An Ontology for Risk Management of Digital Collections

Andrew McHugh

Submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

School of Computing Science

College of Science and Engineering
University of Glasgow

October 2016

Abstract

Maintaining accessibility to and understanding of digital information over time is a complex challenge that often requires contributions and interventions from a variety of individuals and organizations. The processes of preservation planning and evaluation are fundamentally implicit and share similar complexity. Both demand comprehensive knowledge and understanding of every aspect of to-be-preserved content and the contexts within which preservation is undertaken. Consequently, means are required for the identification, documentation and association of those properties of data, representation and management mechanisms that in combination lend value, facilitate interaction and influence the preservation process. These properties may be almost limitless in terms of diversity, but are integral to the establishment of classes of risk exposure, and the planning and deployment of appropriate preservation strategies.

We explore several research objectives within the course of this thesis. Our main objective is the conception of an ontology for risk management of digital collections. Incorporated within this are our aims to survey the contexts within which preservation has been undertaken successfully, the development of an appropriate methodology for risk management, the evaluation of existing preservation evaluation approaches and metrics, the structuring of best practice knowledge and lastly the demonstration of a range of tools that utilise our findings.

We describe a mixed methodology that uses interview and survey, extensive content analysis, practical case study and iterative software and ontology development. We build on a robust foundation, the development of the *Digital Repository Audit Method Based on Risk Assessment*.

We summarise the extent of the challenge facing the digital preservation community (and by extension users and creators of digital materials from many disciplines and operational contexts) and present the case for a comprehensive and extensible knowledge base of best practice. These challenges are manifested in the scale of data growth, the increasing complexity and the increasing onus on communities with no formal training to offer assurances

of data management and sustainability. These collectively imply a challenge that demands an intuitive and adaptable means of evaluating digital preservation efforts. The need for individuals and organisations to validate the legitimacy of their own efforts is particularly prioritised.

We introduce our approach, based on risk management. Risk is an expression of the likelihood of a negative outcome, and an expression of the impact of such an occurrence. We describe how risk management may be considered synonymous with preservation activity, a persistent effort to negate the dangers posed to information availability, usability and sustainability. Risk can be characterised according to associated goals, activities, responsibilities and policies in terms of both their manifestation and mitigation. They have the capacity to be deconstructed into their atomic units and responsibility for their resolution delegated appropriately. We continue to describe how the manifestation of risks typically spans an entire organisational environment, and as the focus of our analysis risk safeguards against omissions that may occur when pursuing functional, departmental or role-based assessment. We discuss the importance of relating risk-factors, through the risks themselves or associated system elements. To do so will yield the preservation best-practice knowledge base that is conspicuously lacking within the international digital preservation community.

We present as research outcomes an encapsulation of preservation practice (and explicitly defined best practice) as a series of case studies, in turn distilled into atomic, related information elements. We conduct our analyses in the formal evaluation of memory institutions in the UK, US and continental Europe. Furthermore we showcase a series of applications that use the fruits of this research as their intellectual foundation. Finally we document our results in a range of technical reports and conference and journal articles.

We present evidence of preservation approaches and infrastructures from a series of case studies conducted in a range of international preservation environments. We then aggregate this into a linked data structure entitled PORRO, an ontology relating preservation repository, object and risk characteristics, intended to support preservation decision making and evaluation. The methodology leading to this ontology is outlined, and lessons are exposed by revisiting legacy studies and exposing the resource and associated applications to evaluation by the digital preservation community.

Acknowledgements

This work was first conceptualized in the Planets (IST-2006-033789) Project, funded by the European Commission's ISandT 6th Framework Programme. Development of the *Digital Repository Audit Method Based on Risk Assessment* (DRAMBORA) was funded in the UK by JISC as part of the work of the Digital Curation Centre and in Europe within the *Digital-PreservationEurope* project, also funded by the European Commissions IS and T 6th Framework Programme. Associated audits were attended by the author and Raivo Ruusalepp. This work was continued in the *3D-Coform* Project (FP7/2007-2013, under grant agreement no. 231809).

Evaluations undertaken at the University of Michigan, the National Libraries of France and Sweden and at CERN were undertaken within the context of the *DELOS: Network of Excellence on Digital Libraries* by the author, Raivo Ruusalepp, Perla Innocenti and Seamus Ross [Ross et al., 2008].

This work was conducted using the *Protégé* resource, which is supported by grant LM007885 from the United States National Library of Medicine.

Eternal thanks to Johanna, Amelia and Louisa for their patience, encouragement and further patience.

Author's Declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Signed:
Printed: ANDREW MCHUGH

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Research Problem and Motivation

Managing and maintaining the accessibility and utility of digital materials is a pressing challenge in today's data driven world. Irrespective of whether one speaks of digital preservation, curation or data management, each task implies a similar set of responsibilities, and a focus on facilitating both contemporaneous information use and safeguarding opportunities for continued consumption many years into the future. Within this thesis we use a number of terms broadly interchangeably, favouring their most demanding definitions, encapsulating sustainability as a critical benchmark.

In 2011 the International Data Corporation presented estimates suggesting that the *Digital Universe*, the full extent of digital content collections throughout the world, would that year exceed 1.8 million petabytes, a reported 62 per cent increase from the previous year [Gantz and Reinsel, 2011]. Further estimates hinted at an expected 44-fold increase to 35 million petabytes by 2020. These difficult-to-comprehend figures can be lent some physicality; the 2010 figure would be roughly equivalent to a stack of DVD's that stretched to the moon and back. The 2020 projections would take that DVD stack half way to Mars. The scale and rapidity of growth associated with our digital information pose several problems, including concerns over information security, power dependency and the costs of data management. However, there is no corresponding growth forecast for custodians of this data; despite the dramatic increase in the scale, the number of IT staff is expected to expand only around 1.5 times by 2020. The purpose of the IDC report was principally to illustrate the challenges of information discovery, but it also throws into sharp relief the challenges associated with maintaining access to such large and rapidly expanding datasets. Dedicated and professional expertise is diminishing, meaning that creators and users of data are increasingly responsible for ensuring their longevity. However, these communities appear to lack curatorial capacity, demonstrated clearly in the results of surveys conducted within this

research and described in sections to follow. Digital preservation and curation are active and not passive processes. Benign neglect of digital materials will seldom result in usable, accessible content in years to come as technology, law and culture change.

Given the scale of data growth we are seeing it is impossible to rely solely upon the proportionately diminishing numbers of trained data managers and digital archivists to safeguard the body of digital heritage. Instead we observe a culture now where the numbers of those with digital custodial responsibility has increased more or less in line with the changing data landscape. Rarely however is such responsibility wilfully embraced - it is more often foisted onto those with no professional interest - or particular competency - in such matters. Increasing numbers of institutions are being required to grasp the challenge of maintaining accessible and available data. In the UK, policies such as those issued in 2012 by the Engineering and Physical Sciences Data Centre and in 2015 by the Economic and Social Research Council require universities and research organisations to establish their own processes and infrastructure to safeguard and ensure the continued availability of data generated as part of funded research. Despite the continued existence of dedicated, expert preservation and curation environments provided by funders such as the Natural Environment Research Council (NERC), science research has been characterised by its increasing data demands. Data curation and management have become necessary and explicit parts of interrogating and manipulating datasets in the new big data context, even in a contemporary sense as the speed of collection threatens to outpace that of data analysis. Research data management is to some extent becoming accepted as a core competency of the scientific process. Even those disciplinary areas that one may intuitively expect to have less onerous data requirements are embracing data as a key ingredient in the pursuit of research. Humanities institutions face complex challenges with issues such as copyright introducing challenges to the digitisation and distribution of textual materials [Stobo et al., 2013]. Custodial organisations such as the *Arts and Humanities Data* are long gone [Open Objects Blog, 2008], with institutions themselves expected to be responsible for maintaining the data their research generates. Social scientists (both qualitative and quantitative) have similarly expanding data and data management requirements. Chapter two details our survey and interview series conducted at the London School of Economics in 2012 which illustrate attitudes, competencies and expectations of a range of social science scholars (comprising several disciplines and levels of seniority) associated with managing and sharing data. Their opinions of existing support provisions and possible policy directions are illustrative of the current demands.

There are few resources available that provide the new data custodian with the definitive and relevant information they require to help them understand these issues. Among the most notable are international standards such as the *Reference Model for an Open Archival Information System* (OAIS) [ISO 14721, 2012] and *Audit and Certification Criteria for Trustworthy Digital Repositories* [ISO 16363, 2012] but their form is such that in many cases under-

standing their practical applicability can be challenging. Likewise, these standards appear to assume homogeneity across preserving organisations that in reality can differ dramatically in terms of legislative context, scale, funding and content coverage.

Other efforts have sought to model digital preservation systems in a variety of ways, including high level reference models [Candela et al., 2008, Antunes et al., 2011], enterprise architecture components [Becker et al., 2011] and data dictionaries [Library of Congress, 2008]. More generic materials are also relevant, including international standards on topics like information security [ISO 27001, 2005] and legal admissibility of evidence [BS 10008, 2008] as well as general organisational framework architectures [Zachman, 1987]. Few effectively bridge the gap between academic theory and practical applicability.

A challenge for those with custodial responsibility has been to distil the tremendous body of literature that exists into a coherent set of requirements that can inform their developments and facilitate their validation [Sinclair et al., 2009, Waller et al., 2006]. That repositories have such potential for variety increases the difficulties associated with presenting a coherent resource with minimal redundancy. Therefore a fit-for-purpose approach is one that is adaptable to the changing circumstances evident across the repository landscape and over time.

A Risk Based Approach

Understanding how to preserve implies an understanding of how to assess our efforts. In the course of this and earlier research, we have developed the concept of risk as a critical component in the determination of preservation capacity and in the validation of preservation solutions [McHugh et al., 2007, Ahmed et al., 2007, Barateiro et al., 2012, Lawrence et al., 2000, Moore et al., 2005]. We define risk as an expression of the likelihood and impact of an event with the potential to influence the achievement of objectives, the success of actions or the sustainability of resources [McHugh et al., 2007]. We favour a view of risk whereby active preservation planning and infrastructural development can be considered synonymous with risk management. The identification and successful management of risk pre-empts the loss of information. In a contemporary sense, definitively demonstrating success or failure is impossible, given that digital preservation is so temporally dependent. For those with the responsibility to preserve, a registry of managed risks offers the next best thing. Risks (a negative element) can be conceptually paired with units of value which are effectively their converse. These units of value amount to individual preservation goals. Our thesis assumes that the mitigation of a specific risk is equivalent to the accomplishment of a corresponding preservation goal. Risk appetite, and risk management capacity are measures of preservation success.

The approach fits well because digital information is at risk, irrespective of whether it is being actively used, modified or manipulated. Technological, organisational, social, legal and financial issues all act individually and conspiratorially to limit access to and interpretability of our collective digital memory. Even where data remain static the contextual factors that facilitate or inhibit their usability and availability continue to change. High dependencies on digitally encoded data in commercial, personal and scientific contexts demand approaches and methods to safeguard their availability over time.

Despite an increasingly intensified research agenda focused on overcoming the issue of digital obsolescence, our understanding of information vulnerabilities and of causal links between environmental or object-specific properties and information loss remains rudimentary. While individual facets, such as format characteristics [Abrams and Seaman, 2003] or legislative responsibilities [Oltmans, 2003] are in isolation well understood, a holistic and continuous understanding is missing. The digital preservation community is large and varied, and a common knowledge base profoundly absent. Even within single environments there is complexity in terms of organization, technology and priorities. A macro understanding of the relationships that exist both within and between preservation environments permits greater overall understanding of risk and of the implications of particular preservation interactions. Given the burden of expectation that now falls on ill-equipped data creators and users as often as on a proportionally diminishing community of data custodians, a linked, coherent picture of digital information custodial best practice becomes critically important; this is the role of this thesis.

Research Questions

Given this context, we define several research questions to inform our efforts. Firstly, can we develop an effective online tool to support the development and evaluation of preservation efforts? Secondly, can we collect data illustrative of real world processes and supporting infrastructures for preservation? Next, can we analyse that data, to make sense of it and draw a correspondence with risk exposure and resolution? Can we take these data and our experience of their collection to reflect on the suitability and value of existing preservation evaluation metrics? Can we develop from the data a structured knowledge base capable of interrogation and integration with a wider range of applications? Finally, can we develop novel tools that effectively validate its usefulness?

1.2 Research Objectives

In pursuit of the resolution of these research questions we present a series of corresponding objectives. Collectively, they may be summarised as the conception of a comprehensive on-

tology for risk management of digital collections. The principal problems associated with the current data landscape are its vast, unconnected knowledge base and a set of core guidelines that remain opaque and difficult to understand in practical terms by even experienced preservation practitioners [Ockerbloom, 2008]. We seek to deliver mechanisms by which this knowledge can be rigorously validated, structured and ultimately applied.

To satisfy our first research question we will develop an associated methodology, characterised as an interactive online tool, for undertaking risk management within such contexts. That is, we aim to streamline and systematise the process of infrastructural assessment and to make the practical more elegantly interlaced with the theoretical. Its accomplishment presupposes and encapsulates a process for organisational analysis and evaluation.

Using this methodology, we will address our second research question, surveying a range of repository services within which data is created, used and preserved. We will do this via primary onsite research and by analysing self-evaluation responses contributed via the online tools we have developed. We will target a selection of organisations that exhibit diversity in terms including but not limited to geographic location, legal jurisdiction, disciplinary association, collection type, budgetary model and scale.

With respect to our third research question we will establish by analysing these surveys a definitive understanding of best practice for preservation, and of issues that limit productivity, introduce additional resource costs or threaten the availability and accessibility of our valuable digital heritage.

Our fourth research question concerns existing approaches; we will evaluate several existing methodologies and criteria for undertaking preservation assessment. Among the most notable are a range of international formalised and *de facto* standards for preservation certification. This thesis will explore their qualities, most significantly their breadth of applicability and perceived utility. We remain conscious throughout that any new developments are most likely to enjoy success if compatible with and complementary to the strengths of existing provisions.

To meet the challenges of our fifth research question we will structure and present best practice knowledge in a suitable taxonomical and ontological format. Again its success will be dependent not only on the extent to which the knowledge contained within is exhaustive and definitive, but in terms of its usability and value. Associated with this will be a relational model designed to express the circumstances within which preservation activities, investments or regulatory requirements intersect with associated risk. Its success implies the ability to traverse bi-directionally - to be able to identify best practice approaches to resolve known risks, and to identify what risks may threaten a preservation context that employs particular tools, strategies or policies.

Our approach with respect to our final research question will be to develop, showcase and

evaluate a selection of practical tools that use this structured knowledge, evaluating their performance in individual production environments and research contexts. These will facilitate preservation management and data curation for both generic and specialist users. The latter group includes 3D model curators and research data management professionals.

We will present a view of preservation that is largely bottom-up, taking its inspiration from findings of a series of investigations undertaken in an international selection of preservation contexts. Acknowledging existing work that has been done (notably in the area of preservation repository certification) we do not present our findings in isolation, but seek to illustrate their relatable qualities. We align conceptual elements to existing criteria within the *Trustworthy Repository Audit and Certification Criteria and Checklist* (TRAC) [CRL/RLG, 2007] in order to illustrate their usefulness in expansion and validation of this and similar top-down resources.

1.3 Methodology

We adopted a mixed method approach in pursuit of our outcomes. Survey and detailed analysis work at participating institutions provided our starting point, involving extensive questionnaire and interview techniques, observed and stated behaviour. Detailed analyses were undertaken of organisational and technical documentation to build institutional perspectives that provided a basis for discussion and additional interrogation during on-site activities. We also took the opportunity to evaluate responses submitted to the online *DRAMBORA Interactive* self-evaluation tool that we developed during the course of this research. *DRAMBORA* is the *Digital Repository Audit Method Based on Risk Assessment*, which with colleagues we conceived to address shortcomings in existing audit methodologies. Its interactive self-evaluation tool was developed to lend usability and connectivity to those organisations using it to direct their self-assessment work. Over one hundred full evaluations have been conducted using that platform by an internationally diverse range of information custodians.

A methodology to perform risk management was continuously shaped in the course of these activities and by reflecting on comparisons with other complementary standards. Literature reviews of existing standards and their evidential foundations were developed. Liaison with the individuals participating in pilot assessments and the wider community provided a means to ensure that the best practice we sought to document remained representative. We employed available audit standards such as *TRAC* [CRL/RLG, 2007] to inform the assessments. We were able to critically appraise not only the institutions being audited, but also the metrics upon which these audits were based. Those who were the focus of our assessments were typically quick to identify any shortcomings of the benchmarks we sought to measure

them against.

Taxonomy and ontology development are inevitably iterative processes. Field and desk research yielded high level classifications of preservation elements, the building blocks of the preservation process. Goals, Activities, Resources, Policies and Rights became increasingly expressive means to understand the connectivity between parts of the preservation system. Our analyses were coded to isolate discrete examples of each and an interactive tool was developed using semantic markup software to formalise relationships between individual abstract instances. This was iteratively developed, with reference to existing practice data and top-down resources such as the various international standards that steer preservation efforts. A suite of tools was built to facilitate the construction of a relational ontology structure and this in turn provided the basis for a selection of applications that use the data.

Evaluation of the resource was undertaken with consideration of a further range of institutional audits, both our own and administered elsewhere, in order to understand the comprehensiveness of the ontology's coverage and its applicability. The ontology and its associated applications were also evaluated against a further program of institutional assessment, with the results revealing their value and perceived value from a range of participating stakeholders.

1.4 Research Outcomes

The principle contributions of this thesis are manifold. The first is an encapsulation of digital preservation practice and best practice presented as a series of case studies of exhaustive analyses of a range of preservation services. These relate to work undertaken in a series of UK, European and US funded projects and correspond to data preserving organisations that vary in terms of jurisdiction, content types, scale and maturity.

Developed iteratively based on our increasing understanding of best practice for planning and evaluating preservation activities, the *DRAMBORA Interactive* online application comprises a major outcome of this research. Developed to reflect the emerging *DRAMBORA* methodology for infrastructural self assessment this resource has achieved impact across a worldwide context. It has been used to support professional activities and also as part of University curricula in well over one hundred international settings. In addition to providing a fertile source of data for the ontology development that followed it continues several years after its release to be relied upon by information professionals in the course of their own efforts.

The outcomes of preservation audits and the user contributions to *DRAMBORA Interactive* provide much of the basis for our final major contribution, a novel ontology for summarising and presenting digital preservation infrastructure information in an integrated and interlinked

form. The *Preserved Object and Repository Risks Ontology* (PORRO) presents a linked model comprising goals, resources, activities, policies and rights and responsibilities within a preservation system, each explicitly mapped to corresponding risks in order to make explicit causality and opportunities for amelioration.

A further major outcome has been the development of a selection of applications that use the ontology and are indicative of anticipated use cases. This includes the *PORRO Browser*, *Collaborative Assessment of Research Data Infrastructure and Objectives* (CARDIO) (a data management capacity management tool) and a specialist preservation component supporting long term availability of three dimensional models and associated metadata developed as part of the *3D-Coform* System.

Our results are documented in a range of technical reports and conference and journal articles [McHugh et al., 2007, McHugh et al., 2008, McHugh, 2011, McHugh, 2012]. These illustrate the development of *DRAMBORA* throughout the pilot assessments, emphasising its shortcomings and document the subsequent development of *PORRO* as a means to inform self assessment, providing a low barrier to entry while facilitating, if not enforcing appropriate rigour.

1.5 Thesis Outline

Chapter 2 explores approaches to preservation infrastructure management, describing current approaches to planning and evaluating digital preservation approaches and infrastructures. It summarises existing tools, metrics and standards for supporting these activities and describes the relationships between approaches aimed specifically at digital preservation management and more generic information management. This chapter also introduces *DRAMBORA* and its sister resource *DRAMBORA Interactive*, respectively developed alongside and as a major part of this thesis. The latter online manifestation of the core methodology extends the functionality of *DRAMBORA* and enables the capture of user contributions as well as facilitating a greater dialogue between a best practice knowledge base and the self assessment process. In collaboration with colleagues at University of Glasgow, the National Archives of the Netherlands and the Estonian Business Archives we conceived *DRAMBORA* as a counterpoint to the top-heaviness and inflexibility of the existing preservation assessment standards. Our efforts to realise it as an online tool have greatly facilitated its usability and visibility.

Chapter 3 comprises our analysis of digital preservation approaches and presents a representative range of preservation experiences from a varied selection of organisations. Its purpose is to illustrate the body of evidence upon which the preservation ontology is based. The accounts within are themselves expected to be of considerable value; seldom do custodial organisations offer comprehensive insights into all aspects of their operations, from policy

and staffing to technological and financial arrangements. This chapter includes a suite of such accounts from a range of preservation organisations that exhibit diverse characteristics.

In Chapter 4 we introduce the *Preserved Object and Repository Risks Ontology*, a classification approach for preservation and an ontology of factors that inform preservation success and risk exposure. This is a core outcome of the research and represents a distillation of the best practice described above to a human and machine readable knowledge base.

Evaluation follows in Chapter 5, including a summary of some traditional difficulties associated with evaluating digital preservation contributions, and a description of the process of engagement with digital preservation expert sources and documented best practice resources in exploring the merits of this research. Our evaluation reveals not only the value of the research outcomes but positions them within an existing context for preservation development and evaluation.

Finally our concluding Chapter describes the broad lessons learned within the research and recommends a new model for evaluation of digital preservation research and outcomes. We consider the extent to which we have been successful in meeting our stated research objectives - we reflect too on the possibilities for further work using the outcomes of this research, and more philosophically on the future opportunities for supporting preservation and data management.

The thesis' Appendices include full assessment reports from each of the described pilot audits, those activities that ultimately informed the ontology's population. They also include a full structure of the ontology introduced in Chapter 4.

Chapter 2

Approaches to Preservation Infrastructure Management

2.1 Introduction

Our research is motivated and propelled by lessons learned in the course of preservation validation research, and by visible opportunities to optimise the supporting methodologies. A common blueprint can support both the establishment and evaluation of a given system. Preservation planning approaches such as *Plato* [Becker et al., 2007, Becker et al., 2008, Becker et al., Becker and Rauber, 2011, Strodl et al., 2007] enable the systematisation of preservation systems development, to which concepts of risk and risk management have become increasingly integral. From risks associated with physical media integrity [Stanescu, 2004] to file format stability [Lawrence et al., 2000], preservation interactions are often considered to be part of an ongoing process of risk management. Our own contribution, the *Digital Repository Audit Method Based on Risk Assessment* (DRAMBORA) [McHugh et al., 2007, Ross et al., 2008, McHugh et al., 2008, Innocenti et al., 2008a, Innocenti et al., 2008b] and *DRAMBORA Interactive* have been similarly pivotal in the introduction of methodological structure to organisational risk awareness.

We developed *DRAMBORA* based on a realisation that risk is a compelling factor in the selection of preservation approaches. Risks are considered in terms of their impact on valued outcomes. The strength of a given plan is determined by the extent to which it avoids or mitigates those pitfalls that prejudice the accomplishment of a desired objective. Preservation is fundamentally about safeguarding against possible negative outcomes: about maintaining things the way they are amid external influences that would seek to disrupt. It is a means by which one can quantify and systematically address whichever threats arise. The most valuable outcome will be the one that best limits risk exposure and vice versa. Risk management requires a consciousness of the value of digital materials, an awareness of the implications

of employing particular preservation strategies, and an understanding of one's own priorities and tolerances (risk appetite). Core tools used in preservation planning such as the *objective tree* focus upon technical and infrastructural factors that influence access to information content (a combination of representation mechanisms and supporting infrastructure), the merits of proposed preservation tools or interventions and the properties of those data that one purports to preserve. Neither preservation planning nor validation has been supported by tools to adequately express or explore sophisticated information interrelationships. Nevertheless, the preservation community has exploited knowledge management approaches elsewhere.

Before considering the landscape of preservation management tools, we turn our attention to the justification for their existence. We seek to demonstrate the problems facing not only large scale organisations with preservation as their core remit, but also those individuals relying on digital content (and by implication its sustainability) but lack dedicated support.

2.2 Challenges of Preservation Management

In order to quantify to some extent the extent of challenge facing the digital preservation community we undertook a large scale survey and series of interviews at the London School of Economics and Political Science in 2012. LSE was a willing host, and one that approached us prompted by increasing stakeholder expectations surrounding data curation and preservation. Theirs was a disciplinary context where attitudes to and reliance upon data varied across its core disciplines. In that sense it was considered to be a useful setting to better understand current and emerging expectations and challenges facing organisations with respect to short and long term information curation. Our goal was to understand attitudes, challenges and responsibilities associated with the management and preservation of digital resources, specifically those used within research. Functions of preservation planning and assessment cannot be assumed a level of significance or importance. The establishment and roll out of certification services for repositories may be prioritised by custodial organisations and their stakeholders. However, we propose that the availability of flexible support and best practice insight is more compelling to data managers and information curators. Our view, informed directly by our experiences evaluating repositories, is that the process of being questioned and asked to justify an approach is of greater value than a certification. Our research indicates too that almost all data users are expected to perform at least part of a custodial role.

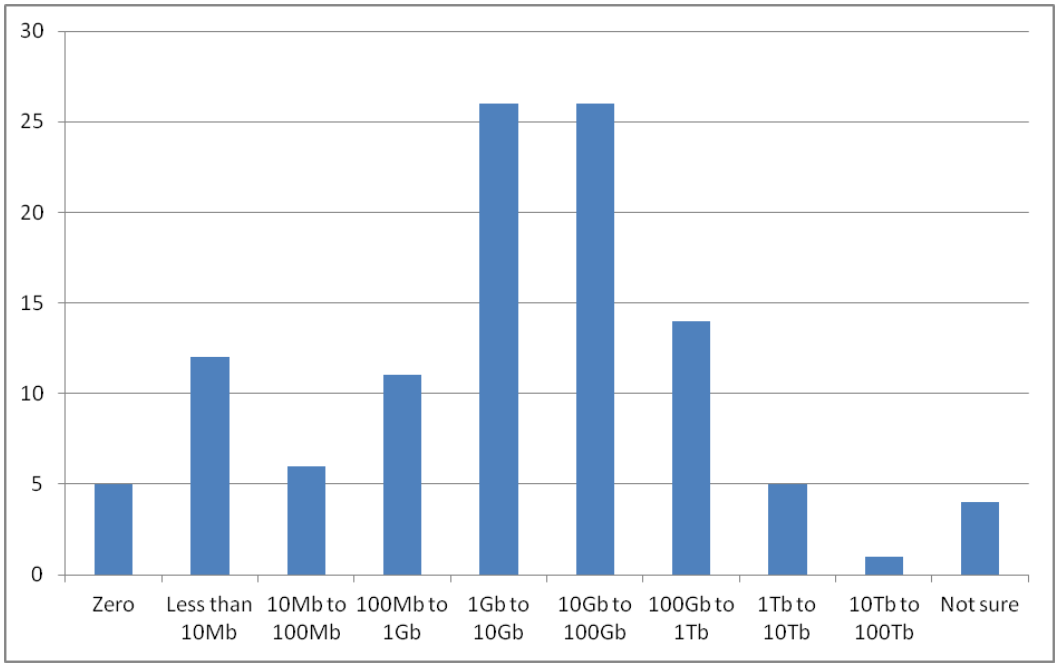
In collaboration with colleagues at LSE's Library, Research Office and IT Service we issued to all LSE academic staff an electronic questionnaire that queried attitudes to research data management within the institution. One hundred and eleven LSE staff responded - they comprised a range of disciplines and roles. Each respondent was asked whether they would be

prepared to be interviewed to discuss these issues further. A cohort of sixteen scholars agreed to do so and met with us to provide their views on research data and on the infrastructures that were available and/or required to support its management. They comprised a range of levels and disciplines, from early career researchers to senior Professors. In disciplinary terms LSE is rather homogeneous in comparison with many Higher Education institutions but our sample included most of the social sciences, including researchers using both qualitative and quantitative methods.

Our results revealed the complexity associated with data management, and the increasing responsibilities of those creating and administering digital information. A consistent message was that associated infrastructures and institutional and personal competencies were rarely commensurate with the scale of and increasing dependence upon digital information.

Responses indicated a general trend in data size per respondent of somewhere between 1Gb and 1Tb of data, tending towards the lower end in most cases. To answer this question respondents had been instructed to calculate their data footprint using operating system level file management software (e.g. *Windows Explorer*) (see Figure 2.1 for details of data size by respondent).

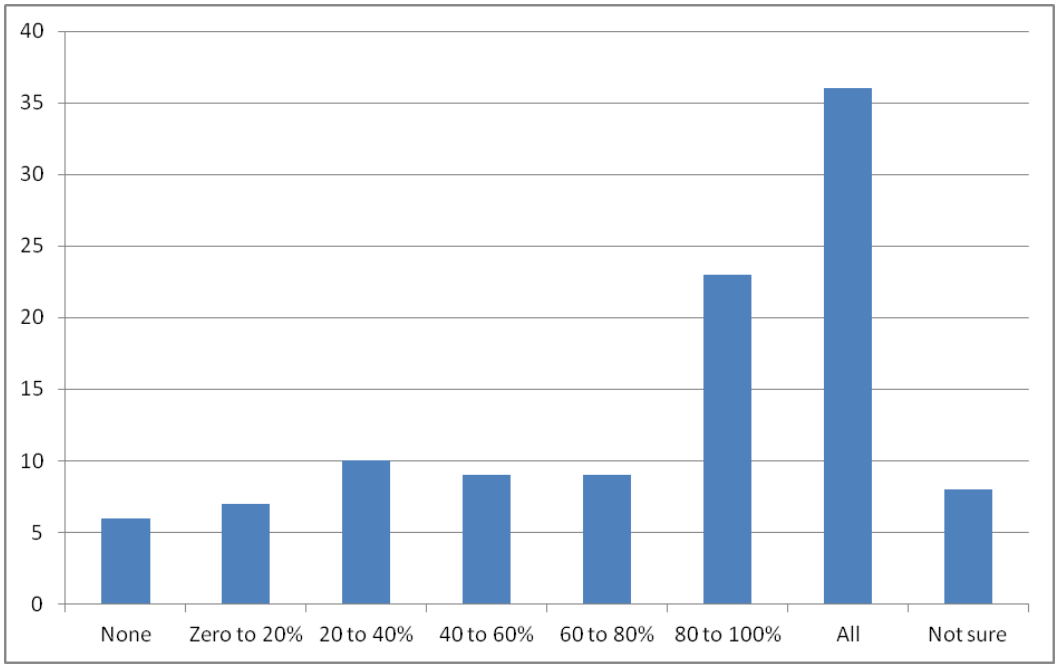
Figure 2.1: LSE Case Study: Data Size By Respondent



The proportion of datasets that respondents considered to require preservation (illustrated in Figure 2.2) offered an indication of the perceived value of their collections. More than half of the respondents suggested that at least 80 per cent of their collection should be preserved.

Those with larger data collections typically demanded their complete preservation (see Figure 2.3 for details of proportions of data requiring preservation by data size), perhaps indicat-

Figure 2.2: LSE Case Study: Proportion of Datasets Requiring Preservation



ing a greater awareness of their data's value, or simply highlighting those disciplines which were more data-driven.

Figure 2.3: LSE Case Study: Proportion of Datasets Requiring Preservation By Data Size

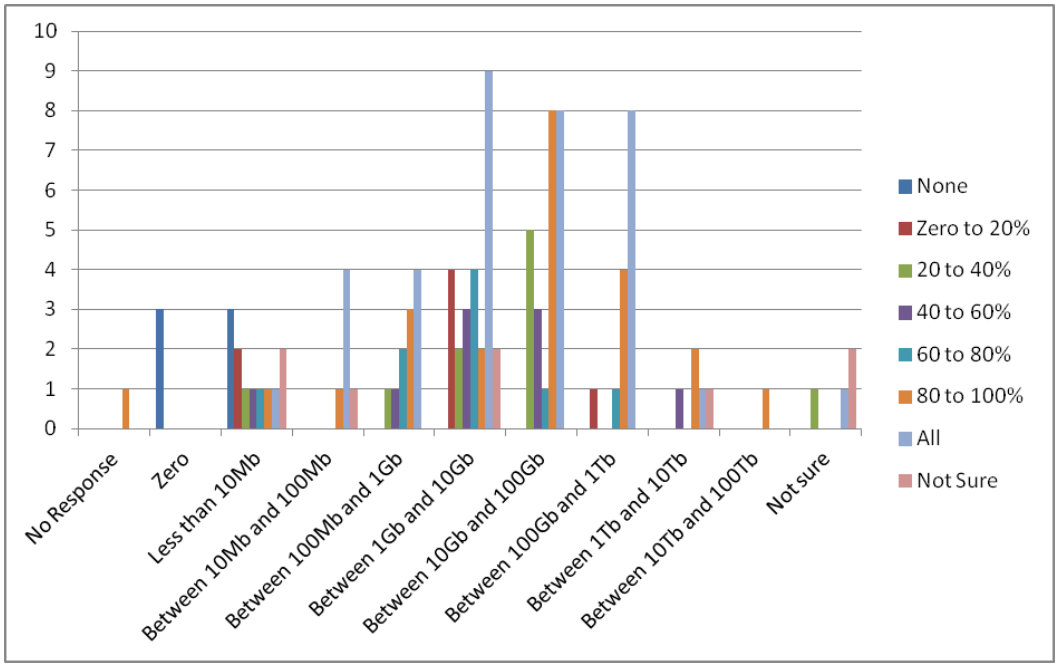
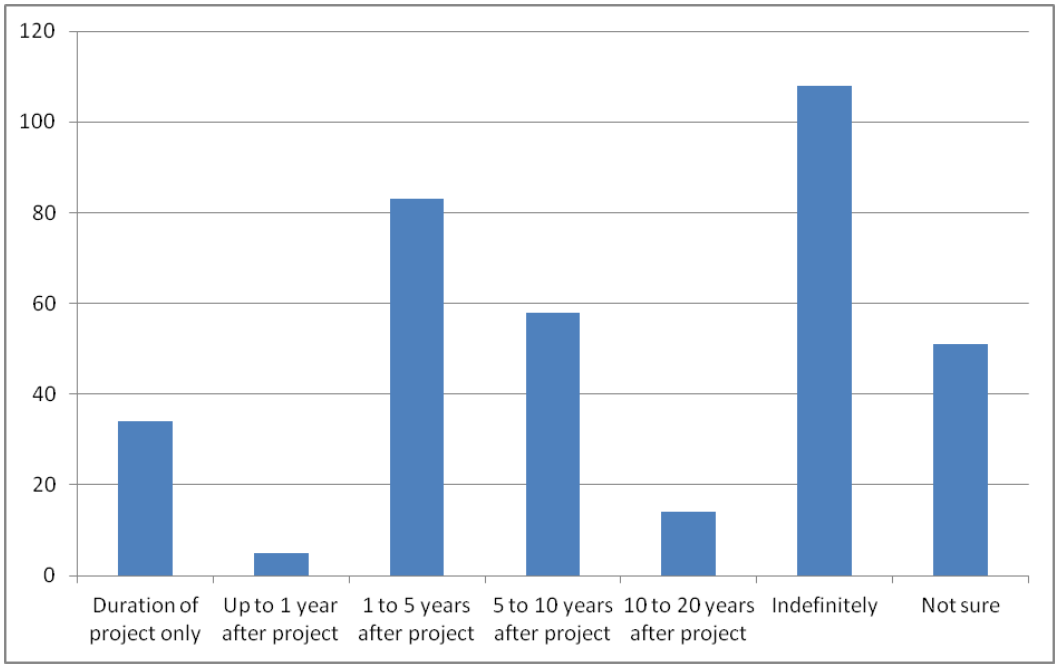


Figure 2.4 illustrates total data retention requirements for each respondent. The fact that the most popular answer was "indefinitely" indicates an emphatic appetite in favour of long term data preservation; when allied with a consensus in favour of keeping most or all data the need for appropriate custodial provisions and expertise are clear.

Figure 2.4: LSE Case Study: Perceived Data Retention Requirements



Illustrated by Figure 2.5, respondents were also invited to consider how long they would wish data to be retained beyond the end of its current period of use, for each of four potential user communities. The groups were a) the respondent him/herself, b) other researchers within the same institution, c) external researchers operating in the same field and d) external researches in other fields. The results indicated that preservation was seen as primarily a personal priority - many respondents argued that a retention period of 1 to 5 years would be sufficient to meet the needs of colleagues or external stakeholders. The number of 'not sure' responses is perhaps indicative of a lack of clarity of the responsibilities funded researchers have to maintain their data over time. Being expected to manage one's data does not equate to an understanding of the associated responsibilities.

Respondents were also asked to consider where responsibility lay for preserving access to their data (see Figure 2.6). The results appear to place responsibility emphatically upon researchers. Continuing a theme, individuals appeared to be embracing responsibility for their data's continued management and availability, but evidence of capacity and required competencies was very limited. We suggest that the response spoke more to the lack of supporting infrastructure than any widespread desire to take on the responsibility for managing data. Given the significant demands for retention scale and term, the extent of the challenge, and the need for supporting infrastructure, is clear.

When asked whether they would be interested in the establishment of a research data catalogue service or infrastructure a majority were either very or somewhat interested (Figure 2.7).

Figure 2.5: LSE Case Study: Perceived Data Retention Requirements By Group

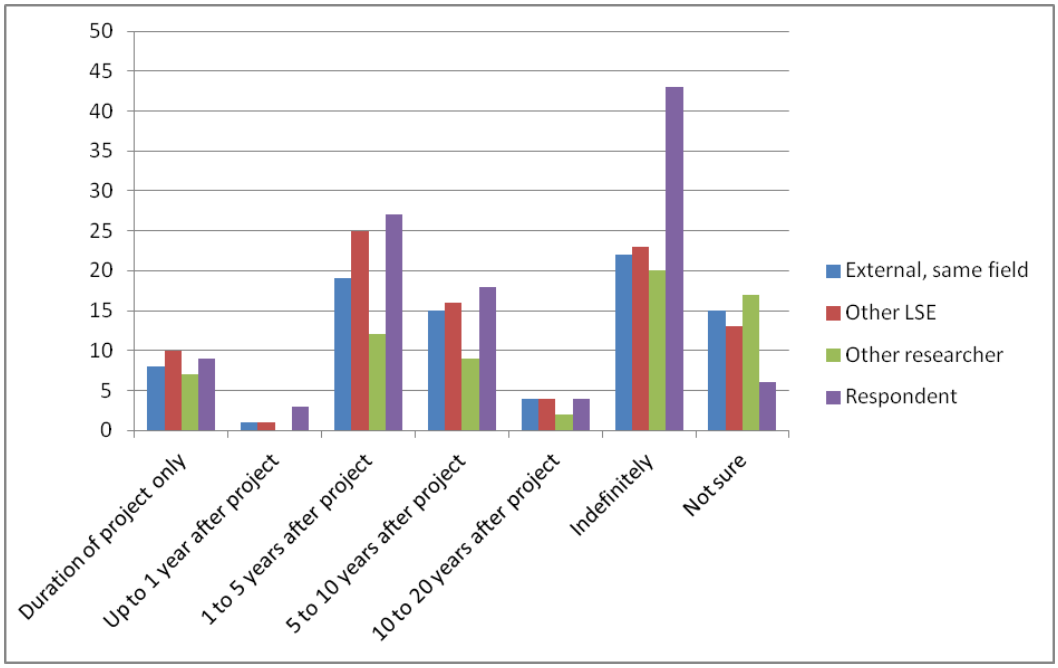
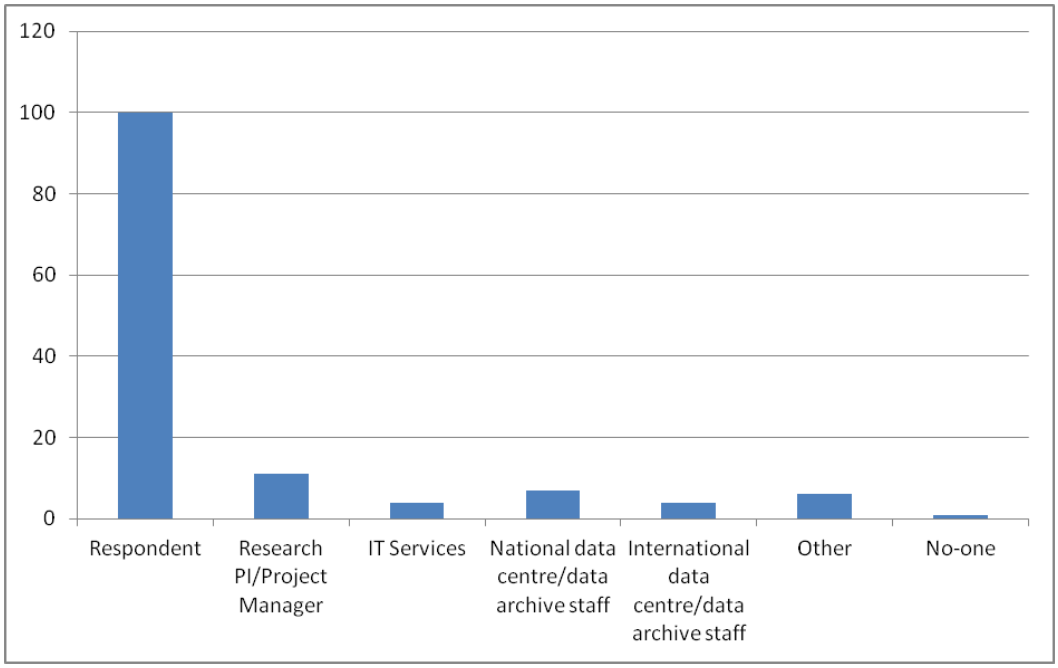


Figure 2.6: LSE Case Study: Perceived Preservation Responsibility

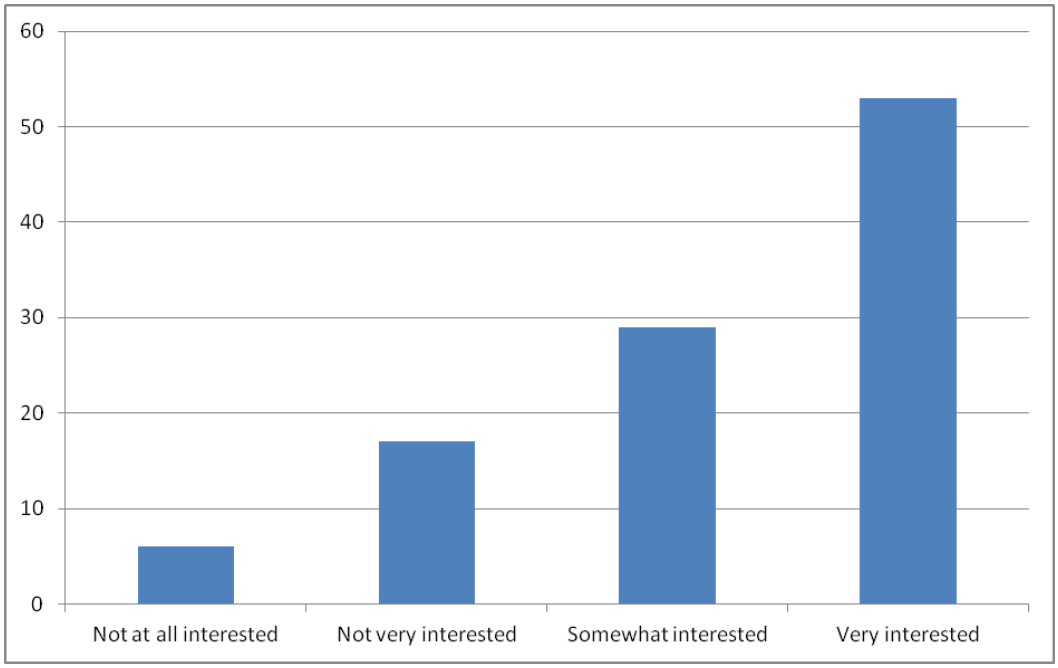


In terms of total data we recorded a potential range across the respondents as follows:

Min Total Data (TB): 16.69
Max Total Data (TB): 166.87

The variation is accounted for by the difference between higher end options in the questionnaire (e.g., the sole respondent claiming to have between 10 and 100TB skews the results and

Figure 2.7: LSE Case Study: Interest in Institutional Data Catalogue



automatically introduces a 90Tb ambiguity in the results). Given the factor of ten that distinguishes the top and bottom end of each selectable range the variation will always maintain that 10x relationship. Extrapolating these figures to the entire School research community further exacerbates this fuzziness, but a crude sense of minimum and maximum contemporary data requirements remains compelling.

Four respondents were unsure of their data usage and depending on their circumstances this could potentially further skew results (they were omitted from this analysis). With 107 respondents left this suggests a personal data footprint at LSE between approximately 150 GB and 1.5TB.

The issue is complicated when one considers the scale of the preservation challenge, achieved by combining those questionnaire responses corresponding to data scale and proportion requiring preservation. This gives a matrix of four values corresponding to the values in low/high parts of each selected range. Again, the most useful statement we can make is of overall minimum and maximum preservation data requirements, which are as follows:

Min Total Preserved Data (TB): 12.37
Max Total Preserved Data (TB): 151.25

We can conclude that although our estimates of data storage requirements are not exact, they are likely to correspond closely to preservation storage requirements. Excluding those unsure about the scale of their data, 6 respondents were unsure of their preservation requirements. One claimed to have between 1 and 10 Tb of data so their data may be significant.

Those unsure were again left out of this analysis. The suggestion is that most data requires preservation, although there is greater variety of requirements from those with less data, a fact largely obscured by the requirements of the large data holders.

Figure 2.8: LSE Case Study: Preservation Requirements Against Data Size

	0	0-	0.2	0.2-	0.4	0.4-	0.6	0.6-	0.8	0.8-	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0.000004	0.000002	0.000004	0.000004	0.000006	0.000006	0.000008	0.000008	0.00001	0.00001
10	0	0	0	0	0	0	0	0	0	0.000008	0.00001	0.00004
100	0	0	0	0	0	0	0	0	0	0.00008	0.0001	0.0004
100	0	0	0	0.00002	0.00004	0.00004	0.00006	0.00012	0.00016	0.00024	0.0003	0.0004
1000	0	0	0	0.0002	0.0004	0.0004	0.0006	0.0012	0.0016	0.0024	0.003	0.004
1000	0	0	0.0008	0.0004	0.0008	0.0012	0.0018	0.0024	0.0032	0.0016	0.002	0.009
10000	0	0	0.008	0.004	0.008	0.012	0.018	0.024	0.032	0.016	0.02	0.09
10000	0	0	0	0.01	0.02	0.012	0.018	0.006	0.008	0.064	0.08	0.08
100000	0	0	0	0.1	0.2	0.12	0.18	0.06	0.08	0.64	0.8	0.8
100000	0	0	0.02	0	0	0	0	0.06	0.08	0.32	0.4	0.8
1000000	0	0	0.2	0	0	0	0	0.6	0.8	3.2	4	8
1000000	0	0	0	0	0	0.4	0.6	[illegible]	0	1.6	2	1
10000000	0	0	0	0	0	4	6	0	0	16	20	10
10000000	0	0	0	0	0	0	0	0	0	8	10	0
100000000	0	0	0	0	0	0	0	0	0	80	100	0

Figure 2.8 illustrates the proportion of data requiring preservation (horizontal axis) and the scale of data. One thing it appears to illustrate (highlighted by the red line) is a relationship between larger data scale and higher retention requirements. Those with smaller datasets vary much more in terms of preservation requirements. Conversely, those with a great deal of data typically appear to favour the preservation of all or at least a greater proportion of that data.

Sixteen interviews with members of the survey cohort offered an opportunity to explore in more detail researchers' attitudes to data and its management and preservation. Participants favoured a broad definition of data, comprising both qualitative and quantitative materials (unsurprising given the hybrid nature of the cohort). Common examples of the former included transcribed, coded interview data, with the latter including spatial and tabulated data sets made available by public agencies. Disciplinary origins of the data were similarly diverse, including medicine, finance, government and law (the institutional setting limited the role of STEM subjects). Data perceived as valuable ranged from open, freely accessible sets to those with significant commercial value or implicit sensitivities. Data were overwhelmingly considered to be digital. Those interviewed with analogue materials of value typically planned to digitise them, again illustrating the increasing demands of even those disciplines traditionally associated with non-digital research methods and resources.

When researchers were asked how they decided which data to retain, the overwhelming majority indicated an intention to keep everything, prompted by difficulties associated with predicting future value, the largely non-reproducible nature of data such as interviews and a general perception that storage costs are comparatively cheap, making viable a keep-all

approach. Less consensus was demonstrated in attitudes to sharing curated data, with disciplinary norms persuasive. Those with quantitative data appeared to embrace sharing most readily. Some feared losing the opportunity to themselves exploit datasets - others had philosophical and scholarly views in favour of sharing. General confusion regarding responsibilities to share (such as conditions of funding), appropriate schedules for doing so and the status of derived outputs (e.g. from commercial datasets) was evident throughout the sample. Communities with no explicit sharing culture (or a determination to avoid sharing) were more likely to see materials isolated and threatened.

Even in comparatively non technical disciplines the systems used to create, process, manage and disseminate data were wide ranging. Applications in use included the mainstream *Microsoft Office* suite as well more specialist applications such as *Stata* and *MATLAB*. Data management planning processes and associated tools were largely unknown. Even those aware of data management planning responsibility typically approached the issue as one of compliance, and devoted little time beyond satisfying bare bones requirements, mainly those originating from funders. Others demonstrated behaviours that were evidently risky. One respondent described transporting sensitive interview data that if disclosed would be "awful, absolutely awful" throughout a politically unstable country on a laptop hard disk. Capacity concerns around core storage and backup infrastructure were described, with several respondents storing data on personal hard drives backed up using commercial synchronisation services such as *Dropbox*. Few were aware of standards for research data management, or over the longer term, preservation. However when introduced to the concept of a data management plan most agreed that it could be useful. It was felt that storage, if not more active data management could reasonably be centralised.

The principal challenge for research data management cited by respondents was limited time. Data documentation was seen as a less worthwhile activity than traditional academic activities such as publishing ("Had I followed [data documentation best practice] instead of having published 100 articles, probably I would have published 30"). This was exacerbated by a perceived lack of recognition or career advancement for publishing data, irrespective of its subsequent usage. Others highlighted the mismatch between short funding cycles and the long term challenge of data preservation. Nevertheless, there was widespread acknowledgement of the importance of data management. Several respondents approached the issue pragmatically, concluding that data takes too long to create to allow it to be lost (in many cases it cannot even be recreated). Others' perspectives were informed by pressure from academic peers to release data to validate conclusions; by funder expectations with respect to data quality and availability and by the role of data as a currency/commodity in enhancing career development. Some more altruistic justifications focused on opportunities to eliminate community redundancy, or on fundamental academic principles such as validation and replication. A small cross section of respondents continued to argue that there was little in-

centive to curate data any more than is strictly necessary to raise one's academic profile and to satisfy funders.

The issue of infrastructure is central to this thesis and among the respondents several perspectives were described regarding those services and resources relied upon for managing research data. Centralised network drives and backups were identified as critical by many, others citing the importance of centrally managed data acquisition and licensing, *ePrints* repositories (for papers and publications) and liaison staff to connect academics with appropriate support. Others remained sceptical about existing central provisions, claiming no interaction with the centre on data. The issue of trust appeared critical, with concerns raised over storage and high performance computing capacity. Respondents described their concerns surrounding instances of illegal or inappropriate uses of licensed data. Notable gaps in the institutional provisions included data inventory management systems; brokering services to facilitate licensing of data from third parties; acquisition and license management resources and levels of central storage space and high performance computing capacity that reflected the institutional need. Data management and liaison staff were also identified as valued potential additions to the provisions. Given the sensitivities associated with accessing many external data sets (e.g. current UK census data), data holders demanded the introduction at local institution level of appropriate secure access environments. Opportunities for training on data management, related central services on offer and archival appraisal were widely welcomed. This reflected a general confusion about where to go to seek help for things like commercialisation of research data, acquisition of new content, the applicability of Freedom of Information legislation, existing datasets and opportunities to preserve data.

Respondents suggested that an appropriate research data management policy should provide base-level standards of what is expected, and offer concrete guidelines with practical applicability. It should be funder-independent but still inform the preparation of research proposals. It should encourage good practice through recommendations, but not constrain (the more it looks like compliance the more it was thought likely to be ignored). It should encourage principles of sharing and appropriate consideration of what should be kept. Finally, it should be succinct and limit unnecessary complexity ("[not] a ten page document that no one reads"). One respondent summarised, suggesting that academics' two main priorities, to have autonomy, and to share, should be prioritised most highly. More than anything else, there was almost universal insistence that any recommendations or requirements should be accompanied by a commensurate commitment of support and an appropriate institutional provision.

This survey provides compelling evidence of the challenges associated with maintaining institutional infrastructures for data management and preservation. Accounts such as these illustrate that those expected to demonstrate conformity to data curation best practice (by, for example, research funders) are often ill equipped to sustain the scale and complexity of data

within their custody. Better means for promoting preservation and research data management approaches and evidence, flexible to meet the expectations of a range of stakeholders are clearly required. End users and infrastructural administrators and service providers have expectations and responsibilities with respect to data. It is critical is that these be aligned to best practice. Currently, this is typically propagated through less accessible resources (such as international standards) and is limited to the institutional competencies of expensive, dedicated curatorial organisations.

In contrast with this work, which was partly focused on the individual responsibilities and expectations of researchers, the full audit assessments that follow in Chapter 3 and inform many of the outcomes of this thesis primarily took place in dedicated, curatorial organisations. These are the very types of organisation whose responsibilities for data preservation we expect to be increasingly inherited by smaller organisational units, individual researchers and non-specialists. Our goal is to establish a record of best practice and therefore it is natural that we look first to those with demonstrable preservation expertise. However, it is critical that the risks we characterise, and the responses we present as optimal are relevant irrespective of where they arise or how they are utilised. In evaluating the outcomes of this research we explore this very issue. The ontology presented in Chapter 4 was deployed within the context of the Digital Curation Centre's *Collaborative Assessment of Research Data Infrastructures and Objectives* tool [DCC, 2011]. We demonstrated the value of ontology elements and properties to a user base approaching questions of curation capacity and risk exposure from specific and individual professional and personal perspectives. A *CARDIO* assessment at LSE revealed concerns associated with unclear data ownership and responsibilities for preservation, structuring of data for long term curation (e.g. metadata and format choices), implications of relevant regulations and legislation (including intellectual property law) for data management, and adequacy of training. These concerns were evident across a selection of users and respondents that included researchers and those providing research, information and IT services in the institution. As illustrated in the accounts presented in Chapter 3 they resonate very closely with the challenges identified, and often successfully resolved, in dedicated preservation environments.

2.3 Preservation Planning

2.3.1 Approaches and Standards

Preservation planning describes the systematic approach to the challenges posed by objects or environments to continued information availability, accessibility and usability. *Plato* [Becker et al., 2007, Becker et al., 2008, Becker et al., Becker and Rauber, 2011, Strodl

et al., 2007] is a *de facto* preservation planning approach with high knowledge demands. It requires consciousness of the value of digital materials, an awareness of the implications of employing particular preservation strategies, and an understanding of one's own priorities and tolerances for achieving success. A favoured means of informing the process is the creation of *objective trees*, used to express strategic and operational priorities in a hierarchy. The goal is the formation of a rationale or justification in favour of a particular approach, wholly based on the environment within which risk arises and has impact.

Preservation planning is particularly constrained by perceptions, expectations and priorities of individuals. The procedure implies the input of diverse constituencies, each of which may have differing preservation priorities. A case study at *ArsElectronica* [Becker et al., 2007] included collaborative workshops with curators, art historians, computer scientists, preservation specialists and management; each role could be reasonably associated with a myriad of policy and procedural responsibilities.

If we wish to understand the wider impact of individual choices, and the nature of relationships between discrete objectives, the existing approaches are limited. For example, within an electronic publishing context we may wish to ensure on one hand the preservation of sufficiently high resolution images to enable legible representation of a smallest meaningful element. On the other we may wish to enforce file size limits to best support scalability. In such circumstances the incompatibility or tension is self-evident, but the example is illustrative of the implications that relationships between system or information facets may have in terms of decision making. In other cases the link between priorities may be less explicit, and as a consequence less well understood. One might for instance prioritise support for embedded metadata in image formats. However, a format satisfying this requirement may have poor tool support and as a consequence staff may become dissatisfied and seek shortcuts in their coding of metadata. Competencies for using related metadata tools such as relational databases may deteriorate. Formalised understanding of such wider potential consequences and associations would not only assist the interpretation of experimental results, but also the conception of more informed *objective trees*.

Risk often emerges at the intersection between information or system facets. For preservation systems, typical origins of risk include combinations of discrete processes; conflicting properties of preserved or operational resources; characteristics and capacity of actors contributing to preservation; and the nature of the context surrounding preservation, including but not limited to legal, geographical, financial, technological, cultural and historical issues. Understanding risk exposure is further complicated by the fact that risks both influence and follow as consequences of other risks. The existence of a risk can generate others, and can influence risk severity in a dynamic fashion. This concept continues to risk management, whereby the implementation of particular approaches may affect risk exposure elsewhere. For example, if software access systems are upgraded to newer versions in response to emerging security

vulnerabilities there may be an exacerbated likelihood of legacy content being no longer fully supported. A common incompatibility is expenditure; if money is invested in the process of risk resolution there is greater likelihood of resource shortfalls in other areas.

It is in the expression of such relationships that existing approaches such as *Plato* appear to fall somewhat short. Risks do not arise in isolation, and likewise cannot be evaluated without consideration of cause and effect. Effective risk management and preservation planning both demand understanding of the implications of decisions. Within an information environment this demand can only be satisfied with an expressive documentation model.

More recent work developing ideas and lessons from the *Planets* preservation planning approach and *Plato* is ongoing within the EU Framework Programme 7 funded *Scalable Preservation Environments* (SCAPE) project [SCAPE, 2014, King et al., 2012]. A taxonomy of representative decision criteria and influence factors has been compiled, based on an extensive evaluation of a number of case studies [Becker and Rauber, 2011]. These are in turn mapped to models to support decision making in software quality, format assessment and object properties. It is an attempt to establish a definitive, common framework for decision factors for preservation planning. Properties are established for a preservation outcome, in terms of an object (in a broadly conceptual, or semantic sense), its corresponding format (its structural representation) and any contextual outcomes, such as costs incurred. Likewise, properties of the preservation action are recorded, encapsulating runtime (e.g., performance, memory use), static (non-runtime considerations such as licensing costs) and judgement (not objectively determinable) aspects. Researchers used the *SQuaRE* quality model [ISO 25010, 2011] as the basis for software quality evaluation. In their adoption of this model acknowledgement is made of "business" factors which are relevant in an organisational decision making context and although not integral to this model have to be considered alongside intrinsic aspects of preservation actions.

The conception of *objective trees* need not be limited to object-centric characteristics; one should also factor into preservation planning those factors considered more passive, such as repository characteristics or other contextual factors. Similarly we may wish to consider in wider terms the meaning of preservation action and intervention. All too often the role of infrastructural or procedural facets can be diminished in favour of evaluating mainly those object-related implications of employing a certain migration or emulation strategies. First and foremost we must develop means of understanding wider implications of particular approaches. This only becomes meaningful when we begin to acknowledge the inevitable relationship between "coal face" preservation interventions and wider business decision making. Every decision, interaction or investment within a preservation context, irrespective of the extent to which it appears detached from the surface of a magnetic tape or hard disk platter may have implications for the preservation of content.

Preservation planning and evaluation both benefit from encapsulating this wider view, and leveraging wider preservation insight. Our ability to relate and understand the implications of those relationships demands more sophisticated information modelling approaches. The expression of an appropriate ontology presupposes an understanding of the fundamental knowledge requirements associated with each process. The outcomes of the *SCAPE* project provide the digital preservation community with a useful resource for cataloguing object-centric properties and influence factors but there remains a disconnect in terms of more infrastructural issues. We present an approach that could be considered a companion piece to *SCAPE*'s decision criteria.

2.4 Preservation Audit and Certification

2.4.1 Approaches and Standards

In 1996 the *Task Force on Archiving of Digital Information* declared in its seminal *Preserving Digital Information* that "a critical component of digital archiving infrastructure is the existence of a sufficient number of trusted organizations capable of storing, migrating, and providing access to digital collections" [Waters and Garrett, 1996]. In 2002, in response to this clarion call, the US Research Libraries Group (RLG) and Online Computer Library Center (OCLC) issued a joint report entitled *Trusted Digital Repositories: Attributes and Responsibilities* [RLG/OCLC, 2002]. This document outlined a set of requirements for the establishment of reliable custodial organisations for digital information. Among a range of attributes including organisational, financial, policy and technology considerations the report's authors recommended the development of framework and process to support the certification of repositories to demonstrate and display their competencies to the many stakeholders reliant on assurances of capacity and capability. This concluded work begun four years earlier at the *Archival Workshop on Ingest, Identification and Certification Standards* (AWICS) [Steinhart et al., 2009] which had proposed a combination of individual, programme, process and data level assessment to satisfy an overall certification requirement.

The workshop explicitly referenced existing resources available to facilitate such assessment, including international standards on quality assurance [ISO 9000, 2005], professional programme accreditation models such as those used by the Society of American Archivists [Society of American Archivists, 2009] and individual competencies examinations. A preliminary checklist was developed at the workshop to serve as a foundation for the development of new resources.

Meanwhile, the Consultative Committee for Space Data Systems' (CCSDS) 2002 *Reference Model for an Open Archival Information System* (OAIS) was given the status of interna-

tional standard [ISO 14721, 2012]. It too called for the definition of accreditation and certification processes to establish a concept of compliance with its recommendations as part of a roadmap for follow-on standards. In 2012 the standard was revised with an explicit reference included to an emergent certification standard published in 2011 as *CCSDS Recommendation for Space Data System Practices: Audit and Certification of Trustworthy Digital Repositories* [ISO 16363, 2012].

This latter standard was several years in development. Shortly after the publication of *Trusted Digital Repositories: Attributes and Responsibilities*, an international working group was established by the Research Libraries Groups (RLG) [OCLC, 2012] and the US National Archives and Records Administration (NARA) [NARA, 2012] with its purpose the definition of criteria for trusted repository audit and certification. The group comprised experts from domains including space data, archive and library science and data curation, including this author. Initially published as a draft for public comment, the criteria were eventually released as the *Trustworthy Repository Audit and Certification (TRAC): Criteria and Checklist*, with the Center for Research Libraries (CRL) [CRL, 2012b] assuming responsibility for its continued development. As part of the *Certifying Digital Archives* project [CRL, 2012a], we partnered with CRL to undertake a series of further pilot audits in a range of US and European settings, with this author a member of the auditing team. Our final release comprised eighty four individual criteria that should be demonstrable by those organisations seeking trustworthy status, divided into three sections covering organisational, object management and technical infrastructure respectively. This would then provide the basis for the development of an ISO standard, with several experts transitioning from the RLG/NARA group to the CCSDS led *Mission Operations and Information Management Area Digital Repository Audit and Certification* working group (MOIMS-RAC) [CCSDS, 2012], chaired by David Giaretta of the UK Space Agency and Science and Technology Facilities Council [STFC, 2012].

Much of the criticism of *TRAC* in its various forms derived from its rather monolithic and prescriptive nature. We conceived the document by conference call. The expertise of the individual authors could not be questioned, and a limited consultation period enabled practitioners to have their say, but the criteria were firmly top-down in their origin, and presented by and large as a one-size-fits-all solution. Having joined the group quite late on in the development of *TRAC*, our first contributions were reactive; we reflected on its content in a response to the *TRAC* team's consultation on behalf of the Digital Curation Centre. Our focus was on the ubiquity and variety of digital content and its resistance to simple characterisation. Digital preservation is associated with a diversity of challenges that vary according to content characteristics (for example, scale; complexity; associated rights issues), organisational qualities (such as budget; legal jurisdiction; and level autonomy) and strategic priorities (perhaps most notably the extent to which particular preservation functions, such

as ingest, access or management are important). Associated demands and responsibilities are similarly variable, and organisations are generally most interested in ensuring that their funder, end user or depositor expectations are given greatest consideration. Conforming to the demands of a one-size-fits-all set of criteria that at times may be considered arbitrary is understandably less important than meeting stakeholder needs.

We also reflected on the lack of clarity associated with identifying what exactly it means to conform with *TRAC*'s criteria. This point was partially responded to in subsequent releases, including the ISO standard, with the inclusion of limited references and example evidence. Nevertheless, a disconnect continues to exist between the criteria themselves and a relatable, practical implementation.

Other similar efforts have followed the release of TRAC, reflecting demands for lower cost or less onerous criteria, and for jurisdiction-specific provisions. The former is best realised by the *Data Seal of Approval* (DSA) [Harmsen and de Leeuw, 2010], developed in the Netherlands by the Dutch Data Archiving and Networked Services (DANS) [DANS, 2012]. It comprises a set of requirements and a quasi-certification that can be issued following application via a self-assessment process. The *DSA* Board assumed management responsibility for the requirements and for issuing the *Seal of Approval* in 2009. Unlike *TRAC* and *ISO 16363* which are detailed in comparison, the *DSA* comprises just sixteen broadly expressed guidelines. They cover similar issues, but describe required commitments of data producers and data consumers, in addition to the archive itself. This is a welcome development that reflects the reality that successful preservation is dependent not just on the efforts of a nominated custodial organisation, but on actors throughout the information lifecycle. Nevertheless, for practical purposes it is the archive that is considered the "primary implementer" of the guidelines, and should assume responsibility for verifying and demonstrating evidence of the other actors' commitment and capacity. A focus on certification makes this position inevitable - someone or something must be the primary subject of scrutiny and the repository is an obvious choice.

We question the extent to which the *DSA* guidelines are self evidently meaningful. Examples such as "the data repository applies documented processes and procedures for managing data storage" are typical in terms of their generality. The process of application requires repositories to complete a self assessment questionnaire which is then peer reviewed by a *DSA* general assembly member, who verifies that conformity with the guidelines has been demonstrated. Detailed information on the processes and requirements for joining the *DSA* community, general assembly and board is available from the *DSA* website. To assist the applicant, limited additional guidance is available, which explicitly aligns the *DANS* criteria with *TRAC* and other guidelines. In addition, *DSA* references the *Foundations of Modern Language Resource Archives* [Wittenburg et al., 2006] and *Stewardship of Digital Research Data: A Framework of Principles and Guidelines* [Research Information Network, 2008].

There remains doubt as to what might be considered to be a practical expression of conformity. The pursuit and award of the *DSA* is mainly prompted by repositories seeking a validation or acknowledgement of their competencies. Although end users and other stakeholders may feel reassured by the *Data Seal of Approval* emblazoned on a conforming repository's website, in reality it is difficult for them to comprehend what this really means. Although self assessment responses are reproduced on the *DSA* site the reviewer responses are a simple binary *Accept* or *Reject* (only awarded certifications are presented on the site, although it is unclear if any have been withheld) and of the eight DSAs already awarded between March 2011 and September 2012 there is very limited use made of the reviewers' additional comments field (which might provide an opportunity for further clarification) [DSA, 2012]. Nonetheless, there does appear to be a robust organisational framework in place for *DSA* which appears to elicit trust; only those who have already been through the *DSA* process may contribute to its continued development.

In terms of jurisdiction-specific resources, softening the problems associated with a single-fit approach, the German standard [DIN 31644, 2012] published by the nestor *Working Group for Trusted Repositories Certification* [Nestor, 2012] (established as part of the Federal Ministry of Education and Research) is a notable example. It was published in its second version in November 2009 and is structurally similar to *TRAC* (comprising sections covering organisational framework, object management and infrastructure and security). The release of this standard was prompted by a wish to reflect specific financial and legal requirements within a German context. As authors of *DRAMBORA* and *TRAC* we collaborated with the authors to define a shared set of ten core requirements for trustworthy repositories, as a means to help ensure the resources' collective coherence [McHugh et al., 2008].

Continuing this spirit of cooperation, in 2010 a memorandum of understanding to create a *European Framework for Audit and Certification of Digital Repositories* was signed between David Giaretta (chair of the CCSDS/ISO Repository Audit and Certification Working Group), Henk Harmsen (chair of the *Data Seal of Approval* Board) and Christian Keitel (chair of the DIN *Trustworthy Archives - Certification* Working Group) [APARSEN, 2012b, Giaretta and Lambert, 2012]. This sought to galvanise collaboration in the establishment of an integrated framework for auditing and certifying digital repositories. It described three levels of certification and, by association, trustworthiness. Basic certification would be conferred upon those repositories awarded a *Data Seal of Approval*. Extended certification would follow for those who in addition to obtaining a *DSA* complete a full, peer reviewed and publicly accessible self assessment based on *ISO 16363* or *DIN 31644*. Finally, formal certification would be available to those who both obtain a *DSA* and complete a full external audit and certification based on either of the other standards. The signatories agreed to ensure overlap between their efforts, to undertake common promotion, to encourage repositories towards higher end certifications and to carry out related test cases in 2010.

The European Commission supported the memorandum of understanding as part of a series of "EC sponsored initiatives on the audit and certification of trusted digital repositories" [Giaretta et al., 2010].

Further European investment in this issue is evident in the work of the *APARSEN* project [APARSEN, 2012a, Giaretta and Lambert, 2012] , which builds on the *Alliance for Permanent Access* membership organisation and is funded by the EU's Seventh Framework Programme [European Commission,]. Its goal is the establishment of a virtual research centre for digital preservation in Europe. This explicitly incorporates work on common terminology and standards. Similarly, its membership and leadership includes pivotal individuals in the development of assessment approaches and resources. At the 2011 *Alliance for Permanent Access Conference* a new organisation was launched called the *Primary Trustworthy Repository Authorisation Body* (PTAB) [Giaretta et al., 2011]. This comprises a number of experts who have contributed to the development of audit and certification resources and standards and positions itself as "the anchor for the provision of ISO audit and certification of digital repositories and plays a major role in training and accrediting auditors". In July 2012 the *Alliance for Permanent Access* reported an ongoing discussion between *PTAB* and the *ISO Committee for Conformity Assessment* (CASCO) and the *International Accreditation Forum* (IAF) over their roles in the processes of audit and certification [APA, 2012]. Until this was concluded and its outcomes reflected in a standard for auditor guidance [ISO 16919, 2011], no formal certification can take place.

That the *PTAB* group appear to have self-appointed themselves as overseers of the international repository audit domain may raise eyebrows, but it has attempted to counter associated criticism by explaining that "to bootstrap the process an initial body of auditors is defined based, as we believe is reasonable, on the membership of the body which wrote the metrics document"[APA, 2011]. Its constitution appears to extend beyond that however, with assumed responsibilities for accrediting training courses, undertaking initial audits and accrediting those national authorization bodies which will in turn accredit auditors within individual countries and allow for the creation of an international network of competent bodies. To date, a series of test audits have been completed - three European repositories (UK Data Archive [UKDA, 2012], French Centre Informatique National de l'Enseignement Supérieur [CINES, 2012] and the Dutch Data Archiving and Networked Services [DANS, 2012]) received EU funding via *APARSEN* to take part, and three US based repositories (the National Space Science Data Center [NSSDC, 2012], the Socioeconomic Data and Applications Center [SEDAC, 2012] and the Kentucky Department for Libraries and Archives [KDLA, 2012]) each contributed their time freely to do so. The German National Library (DNB) [DNB, 2012] also participated in a pilot assessment in this period based on the DIN standard. These were intended to support the establishment of the European framework (as described in the memorandum of understanding described above), identify

metrics which could not be easily or intuitively understood by participating archives and verify that individual auditors shared a common understanding of the evidential and compliance requirements. The exercises also assisted in the definition of auditing processes which are yet to be formally disclosed (an auditors' spreadsheet, based on the ISO standard is available from the *PTAB* website [PTAB, 2012b]).

The range of checklists now is considerable, but the steps taken at the *European Framework* level appear designed to mitigate possible confusion - coherence between them, and also with other relevant standards, is critical. Examples include standards relating to archival information systems [ISO 14721, 2012] and records management [ISO 15489-1, 2001, ISO 15489-2, 2001]. The ISO and DIN standards are expected to provide an intellectual foundation for repository certification, to represent the standard expected of our preserving institutions. However, despite some promises of procedural guidance, seemingly omitted from scope are many of the practical issues associated with performing an evaluation and determining conformity. A companion standard to *ISO 16363* has been released to describe characteristics of certifying organisations [ISO 16919, 2011] and refers to a number of other relevant standards including those covering Quality Management Systems [ISO 9000, 2005] and Conformity Assessment [ISO 17000, 2004, ISO 17021, 2012]. Neither these nor any of the criteria catalogues address the *process* of assessment explicitly. Furthermore, it is clear that the prioritisation of certification (over audit) is a reflection of the fact that funders' interests are usurping those of other relevant stakeholders. End users and depositors have compelling reasons to demand trustworthiness from the custodial organisations they deal with, but are less inclined to pay for such assurances. Funders appear willing to secure their investment by only financing those institutions with demonstrable capacity. However, assessment is not characterised as just pass or fail (although such would appear more aligned to these stakeholders' expectations). The *PTAB* website's *Frequently Asked Questions* page includes the question "is the certification a simple yes/no?", countered with the response "no, the ISO audit and certification process is designed to be one of continuous improvement. Therefore the certification, assuming the repository meets at least minimum levels, identifies areas which need improvement" [PTAB, 2012a]. The document continues to describe possible award levels that while not strictly binary (*exemplary*; *very good*; *good* or *fair*) are more indicative of a less fluid type of evaluation, again in keeping with the expectations of primarily external stakeholders. Nevertheless, it is clear that the group is trying to appeal to repositories by distancing themselves from a clinical yes/no assessment. This reflects a tacit acknowledgement that preservation is too complex and diverse to allow a single set of criteria to be the basis of evaluation in every context or to support distillation to a single pass or fail judgement. There are grey areas in preservation - one must be able to tailor or weight expectations to suit organisational priorities. Skills and infrastructure requirements for the preservation of digitised, out of copyright texts are notably different from those associated

with the preservation of clinical trials datasets for example. Furthermore, *PTAB* relies upon the trust of the whole digital preservation community, a sizeable proportion of which are running their own repository services; their alienation would be a problem for the viability of *PTAB*'s mandate.

Nevertheless, there is notable focus on the views of external stakeholders; emphasis on the role of best practice criteria in informing external evaluation appears to disregard their value in the management or establishment of preservation repository services. To date, few, if any, certifications have been awarded (excluding the small number of *Data Seal of Approvals* issued). Conversely, many accounts (both formally documented and anecdotal) are available to suggest that repositories are using such tools not as evaluation benchmarks, but to inform the establishment and administration of repository services: to support their risk assessment activities [Antunes et al., 2011, Lyrasis, 2011]. There are few practically-oriented resources for informing such activities. The digital preservation community has a tremendous body of knowledge but it is typically dispersed and spans domains such as computing science and archival and information science. This is not to mention the range of individual disciplines with specialist data that poses uniquely challenging characteristics. More so than *OAIS*, *TRAC* and *ISO 16363* provide organisational and structural blueprints that are far more likely to be referenced and checked off by practitioners interested in ensuring the completeness and appropriateness of their activities than an external auditor.

Reflecting these more commonplace use cases, we developed the *Digital Repository Audit Method Based on Risk Assessment* (DRAMBORA) within the UK JISC funded Digital Curation Centre [DCC, 2012] and EU Framework Programme Six funded *DigitalPreservationEurope* [DPE, 2012], initially as a methodological accompaniment to the *TRAC* criteria. Its development was prefaced by a series of pilot repository audits which used the (then draft) checklist and sought to establish complementary information gathering approaches [Ross and McHugh, 2006b]. Our activity was also intended to highlight issues associated with the checklist in terms of coverage, interpretability and applicability in a range of organisational, jurisdictional and technological contexts. These pilot assessments are described in more detail in Chapter 3, in terms of the lessons learned that informed the development of our ontology of preservation practice. They predated the *ISO 16363* audits by several years but share many of the same anticipated outcomes. Our list of participating repositories was diverse, and included national libraries and archives, smaller scale cultural heritage digital collections, eScience data repositories and distributed research collections. The cohort spanned continental boundaries and demonstrated a range of business models [Ross and McHugh, 2006a].

We conceived *DRAMBORA* as a means to support self assessment of repository services. Its core risk-based approach reflected the fundamental doubt that is at the core of preservation; until time has passed there are few guarantees about the appropriateness of any preservation

intervention. Some will argue in favour of benign neglect, with future forensic technologies likely to enable the recovery of valuable content. Others will suggest that explicit and active management and documentation of resources is a requirement in order to ensure authenticity and integrity are adequately maintained over time. In reality, preservation is about managing risk appetite - those with responsibility to do so welcome both guidance and validation to support and underscore their efforts. Given the increasing prioritisation of formal certification approaches and infrastructures *DRAMBORA* was also a response to a growing need for those with custodial responsibilities to obtain reassurances about their ability to meet relevant expectations. It was envisaged as a means for organisations to establish a systematic self-awareness that would be a necessary precursor to inviting external auditors to pass judgement as part of a more formal certification process. *DRAMBORA* uses risk as a metric; its principle conceit is that digital preservation is a risk management activity, and that capacity and capability to effectively manage risk can be considered synonymous with digital preservation success. At its core is a systematic approach. It requires users to document a repository context in terms of responsibilities, objectives, activities and assets, and then to align these with corresponding risks. Appropriate management responses can be defined, and characterised in subsequent iterations of the process. Unlike existing (and subsequently released) evaluation instruments, *DRAMBORA* is intended to be used in a bottom-up fashion, with the specific preservation priorities of the evaluating institution representing the core benchmark for success.

Subsequently, we developed *DRAMBORA Interactive* as a freely accessible online version of the resource to facilitate the evaluation process and to enable the capture of representative responses to the core questions [1]. At the time of writing 834 repositories have been registered within the system by users. Disregarding spam, exploratory and teaching related[2] registrations (and adopting a conservative metric) approximately 123 repositories have been fully and formally evaluated using the tool. This figure was derived by analysing the progress of audits in each case and excluding any that had not completed the classification of risks, objectives, activities and mandate. The sample of 123 repositories offers a compelling and varied selection; it includes repositories from 21 countries (see Figure 2.9 which displays the country association of repositories registered in DRAMBORA) including the United States (79 repositories), the United Kingdom (15 repositories), the Czech Republic (5 repositories), Germany (3 repositories) and Australia, Canada and the Netherlands (2 repositories each). Institution types (displayed in the chart within Figure 2.10) include Universities (51 repositories), Archives (28 repositories), Libraries (14 repositories) and Museums (9 repositories).

DRAMBORA is by no means immune to the criticisms of ambiguity and inapplicability that

[1] *DRAMBORA Interactive* is available from http://www.repositoryaudit.eu

[2] Several Universities including University of Glasgow, Simmons College, the University of North Carolina at Chapel Hill and University of Illinois have used *DRAMBORA* within their teaching activities

Figure 2.9: DRAMBORA Registered Repositories By Country

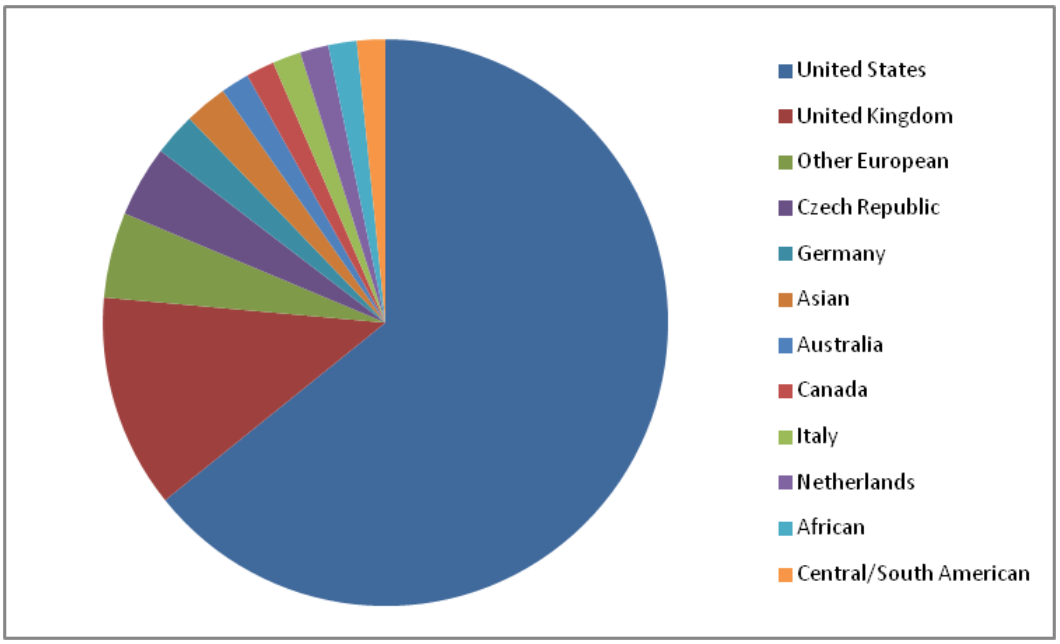
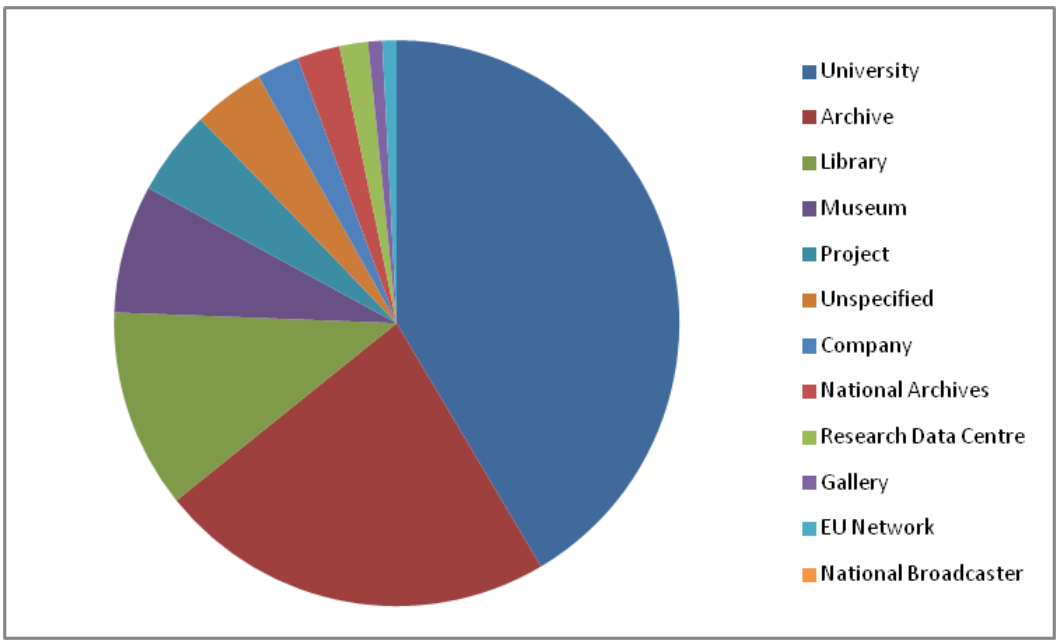


Figure 2.10: DRAMBORA Registered Repositories By Type



TRAC and *ISO 16363* face. During and following its iterative development, we conducted a number of pilot audits at an international range of institutions (see Chapters 3 and 5). In the course of these we gained experience of real world preservation practice that was unique, at the very least in terms of its variety. Despite their largely successful outcomes, a caveat often accompanied the otherwise positive feedback received from participants. Without the expert contributions of external auditors, it was argued, the process would have been less successful. Individuals were constrained by the limits of their own knowledge - systema-

tising the process of assessment was useful, but it was widely felt that the process of risk management demanded the input of those with experience of *comparable* activity. This concern was addressed to a limited extent with the inclusion (in the published and subsequent interactive online versions of *DRAMBORA*) of around one hundred example risks and mitigation strategies. These were developed by reference to the audit experiences, as well as the *TRAC* instrument that we had used as their basis (at least initially). We reflected the latter's structure within the list of risks, incorporating those associated with organisation, technology and preservation workflow respectively. In several cases they were the inverse of good practice promoted within *TRAC*, or the results of extrapolating failure within that framework to a practical outcome. In other cases they were the expression of shortcomings we had identified in our audit work. Risks were defined with a name and description, but in reasonably generic terms to limit the extent to which they were prescriptive. Each risk was accompanied by details of its typical organisational spacing (e.g. physical environment or personnel, management and administration) and role ownership (e.g. management or technical) and prompts to assist in determining its relevance. The latter were intended to encourage self-assessors to reflect on the extent of their risk exposure. For instance, a risk entitled 'staff suffer deterioration of skills' included prompts such as 'are skills refreshment opportunities available to staff?'. Example manifestations were also included to illustrate the types of circumstances within which risks may arise, or their practical effects. Finally, in order to introduce a positive dimension each risk was accompanied by a set of possible mitigation measures, that included both strategies for risk avoidance and treatment in the event of its occurrence.

We acknowledged that this remedy was not wholly satisfying, and even though we aimed to limit the extent to which these were prescriptive and prominently labelled them as being not exhaustive that it implicitly contrasted with the bottom-up philosophy of the *DRAMBORA* approach. There followed further demand for prescriptive guidance from end users struggling to define organisational objectives, activities or responsibilities. Again, in order to assist such definition, but contrary to *DRAMBORA*'s philosophy, we used similar methods to provide a series of examples that end users could refer to in establishing their own organisational picture. These were classified in *DRAMBORA* according to ten "functional classes" which were available to structure responses and facilitate internal communication for self assessment participants, echoing ten common core requirements for repository trustworthiness collectively agreed by the DCC, DPE, nestor and CRL (the corresponding organisations for *DRAMBORA*, and the then-nascent DIN and ISO standards) in 2007 [McHugh et al., 2008]. *DRAMBORA*'s implementation of these requirements (as functional classes) was as follows:

Mandate and Commitment to Digital Object Maintenance Functions and characteristics that correspond to the commitment of the repository or the institution within which it is

based to the maintenance of digital objects, or describe its responsibilities with respect to this.

Organisational Fitness Functions and characteristics corresponding to the repository's organisational viability, sustainability and value, mainly incorporating issues of resource availability, including human resources.

Legal and Regulatory Legitimacy Functions and characteristics corresponding to legislative, regulatory or common law rights and responsibilities of the repository.

Efficient and Effective Policies Functions and characteristics corresponding to the repository's policy infrastructure that facilitates its activities and the completion of its objectives.

Adequate Technical Infrastructure Functions and characteristics corresponding to the technical and security provisions maintained by the repository to facilitate its activities and assist the achievement of its objectives.

Acquisition and Ingest Functions and characteristics corresponding to the repository's negotiation, submission, receipt and ingestion of data from creators and suppliers.

Preservation of Digital Object Integrity, Authenticity and Usability Functions and characteristics corresponding to maintenance of object integrity, authenticity and usability.

Metadata Management and Audit Trails Functions and characteristics corresponding to the documentation recorded by the repository to describe digital objects and processes to which they are subjected.

Dissemination Functions and characteristics corresponding to the repository's distribution of stored content and end user access provisions.

Preservation Planning and Action Functions and characteristics corresponding to the curation and preservation of digital materials within the repository [McHugh et al., 2007].

To be wholly successful, a resource would perform a surrogate facilitator role, with a flexibility that would provide further detail where appropriate or prompt greater examination on issues that might be only partially understood by self assessors. However, each of these efforts (including *DRAMBORA*) is constrained by a reliance on prescriptive, inflexible criteria that may not reflect the objectives or characteristics of a given preservation context. These should be adaptable to permit the evaluation of contexts ranging from, say, a national repository or a single academic maintaining their research data.

In fact, *DRAMBORA*'s reasonably widespread use across a range of disciplines and organisational types has yielded a range of data that provides evidence of priorities and approaches

for preservation, related to associated risks. As described in the Chapter 4, we use this and other assessment data to conceive our own knowledge-base of best practice for administering a preservation system.

2.5 Generic Information Security

A range of generic information standards contribute to best practice awareness in digital preservation. Digital preservation can be considered a superset of a range of organisational and technological end goals (as described in the ontology chapter that follows). Critical standards that are consistently referenced include those on information security [ISO 27001, 2005, ISO 21827, 2008, BS 7799, 2006], legal admissibility of electronic evidence [BS 10008, 2008], quality management systems [ISO 9000, 2005] and risk management [ISO 31000, 2009].

These can be considered alongside more community oriented (although not explicitly digital preservation related) standards such as those associated with archiving and retrieval of digital technical product documentation [BS 9300-003, 2012], performance measures for libraries [ISO 28118, 2009] and records management of information and documentation [ISO 15489-1, 2001, ISO 15489-2, 2001].

With specific applicability to measures for auditing and certifying digital repositories we look to more generic standards on conformity assessment [ISO 17000, 2004, ISO 17021, 2012] which have informed the bespoke digital preservation standards in place [ISO 16363, 2012, ISO 16919, 2011].

Chapter 3

Digital Preservation Approaches Analysis

3.1 Time Proven Perspectives

> "Our yesterdays follow us; they constitute our life, and they give character and force and meaning to our present deeds."
>
> Joseph Parker

As a discipline, digital preservation has wrestled with a number of issues as a consequence of its explicit temporal dimension. Among the most fundamental relate to validation and evaluation of approaches. The standards and criteria referenced in the previous chapter purport to present a definitive account of how preservation should be undertaken, but with the passing of time an essential factor in the determination of the success of any preservation intervention, the pursuit of compelling approaches or standards (and their subsequent expression as 'best') has at times been more art than science.

The studies that comprise this chapter and form the basis for the ontology described in Chapter 4 took place between 2008 and 2014. Each institutional analysis is illustrative of good and bad practice. Each is presented not only in terms of the circumstances during the assessment but also in terms of future improvements and expectations. Little has substantively changed in terms of preservation best practice in this time. New tools and content types have emerged but their impact in terms of fundamental functional approaches has been limited. This analysis, and any optimal approach to digital preservation disregards the transitory and focuses instead on permanence; architecturally, organisationally and in terms of content. We embrace the tools of the day but as vehicles for implementing more fundamental ideas. We have the opportunity to look back at some circumstances that appeared optimal and ques-

tion whether time has disproved or validated such assessments. We also have the chance to re-evaluate identified risks based on the outcomes that have followed.

3.2 Preservation Case Studies

3.2.1 Introduction

In order to better understand the issues associated with preservation best practice we analyse our preservation profiling activities and evaluate the extent to which our findings exhibit continued, contemporary validity. We undertook a series of repository evaluations in 2008 and 2009 (within the parameters of the *Digital Curation Centre* [DCC, 2012] and *DigitalPreservationEurope* [DPE, 2012] projects). As well as providing the participating organisations with an objective and expert insight into the effectiveness of their operations, we sought to explore means for performing evaluation and for verifying the global applicability of metrics, criteria and methodologies already conceived. The cohort of participating institutions represented a diverse subset of curatorial contexts, each contrasting in scale, scope, funding basis, means of deposit, user community and in terms of the nature and origins of their digital holdings. We remain grateful for the welcome and unrestricted accessibility each offered. This was in several cases conditional on us providing assurances of anonymity; this is indicated by the accounts that are presented generically.

The selection of these repositories was based mainly on their availability and willingness to expose their collections and processes to scrutiny. Professional relationships with a number of repository administrators led to discussions about the viability of performing audits. Those that ultimately took part were the ones that agreed and committed staff resource to undertaking the process. There were clear benefits presented to incentivise participation, both in terms of enhancing operational performance and providing an opportunity to engage directly in the development of the evaluation standards.

3.2.2 Approach

Metric and Method

Our initial audits used a draft version of the RLG/NARA check-list (subsequently formalised as the *Trustworthy Repository Audit and Certification Criteria and Check-list* [CRL/RLG, 2007]) as a best practice benchmark. Reflecting their interest in the international audit and certification developments, four of the five organisational assessments repository administrators were already familiar with the document, and with the work that helped shape it, such

as the *Reference Model for an Open Archival Information System* [ISO 14721, 2012]. The final repository was less well versed in the surrounding intellectual framework, and this had undoubted implications, explored in more detail below. An ancillary goal of our audits was to determine the legitimacy of the check-list's metrics, their applicability in a range of circumstances, and their usability both as self-assessment criteria and as a tool to structure and support third part repository audit.

Around a month prior to each visit we issued a local repository contact with instructions to package and distribute a selection of documentation that would support the assessment. This included a range of literature corresponding to policy in several areas, financial information, and content and system documentation.

A two day schedule was established for each onsite visit, comprising interviews and discussions with a range of appropriate nominated individuals. This would typically commence with an opening meeting and tour of the facilities, followed by reviews of organisational characteristics (including staffing, finances, designated community, policies and contracts); ingest procedures and archival storage; preservation planning and strategies; information and access management and finally the technological infrastructure comprising the repository. A worked example would be designed and discussed which was an opportunity to witness or simulate the journey of objects within repository. The visit was concluded with an opportunity for any final questions and a concluding presentation.

Each of the review sessions was linked to corresponding sections of the draft checklist, our adopted means to try to ensure comprehensiveness of coverage and consistency in the series of evaluations. Nevertheless, the conversation was typically wide ranging and not constrained by specific *TRAC* criteria.

Evidence Requirements

Establishing a representative picture of digital preservation best practice demanded methodological coherence and a sound evidential platform. Three principle questions characterised the evaluation across each participating institution:

1. What was documented?
2. What did staff members or other stakeholders believe, think or know happened to facilitate preservation?
3. What actually happened?

The extent to which initial evidence (i.e. that received and analysed prior to the on-site audit activities) was comprehensive, significantly influenced the ease with which subsequent

on-site analysis, comparison and corroboration took place. In most cases evidence was presented as documentation and this was generally useful in establishing an institutional picture. In contrast with personal testimonial it provided an objective foundation to support further inquiry.

In each of the assessments, documents were made freely available from the host institution, ensuring that it was straightforward to identify what was documented. Primary testimony was a useful means of corroborating that written policies, procedures and practices were well understood and representative, providing evidence that staff and stakeholders were aware of the extent of functionality and services offered by the repository, and illustrating a level of intrinsic transparency. The identification of a critical mass of stakeholders or staff whose views, beliefs or understanding differed markedly from that which was documented carried significant evidential impact. Lesser degrees of dissent motivated further interviews, or, where feasible, a conclusive practical demonstration was sought. Testimony unsupported by documentation was persuasive only to the extent to which it was corroborated by other means, whether by interview with alternative staff or stakeholders or through first hand observation. Observation of practice, while likely to carry the greatest evidential impact was suited to only those issues that related to a demonstrable procedure.

Witnessing the repository function in a particular way was representative of its capacity to do so, but not that such behaviour was firmly embedded within the fabric of the repository's infrastructure. Interviews and observation evidence offered a transitory view of the world; the development of a compelling account of preservation practice demanded policies, skills, techniques and functionality that were sustainable, assured and persistent. Documentation was our favoured means of establishing this.

Formal primary documentation included not just paper records, but also online content (such as web pages or wikis) and object or repository metadata. Documentation was typically illustrative of the satisfaction of capacity or commitment-style requirements. Other requirements demanding the existence of particularly policies could be demonstrated with a physical document. Metrics within documentation were illustrative of resource requirements such as appropriate staffing numbers. Service levels and contingency measures were also revealed in documentation.

Unsupported testimony that lacked corroboration, documentation or visible procedure was of limited value, although was used as a basis for further investigation. Subsequent investigation would either elevate this testimony or result in its rejection. The exception was when evaluating organisational commitment. While generally identified at repository level, commitment or appetite could also be determined within the will of individuals, particularly repository management. These softer aspects could rarely be conclusively demonstrated in this way, but such testimony was nevertheless to some degree compelling. In the event of

contradiction by any other evidence type, unsubstantiated stakeholder testimony could in most cases be immediately discarded.

The extent to which testimony was substantiated was important in determining its evidential impact. If several interviewees described a common world view it implied a degree of credibility. Any agreed deviation from documented evidence tended to undermine that documentation and required it to be disregarded. This was also indicative of a poorly integrated documented policy.

Among the most useful secondary documentation was the content of reports from prior audits or certification that had already been awarded (or withheld). In such circumstances that value was largely dependent upon the aspect of the repository that had undergone assessment, and the trustworthiness of the associated procedure or awarding organisation.

Where institutional representatives stage-managed the demonstration of systems or processes the value was to some extent limited. Known issues could be avoided, and non representative strengths (for example, extraordinarily extensive metadata records for particular objects) presented as typical. However, it was considered unlikely that hosts could conceal significant system shortcomings during an in-depth demonstration, and the results were therefore quite compelling. In terms of their evidential weight, these usurp both testimony and documentation as representative of what actually happens.

Our own personal system interactions were preferable. Evidence collected this way was generally conclusive, more so than both testimony and documentation. Documentation would only carry greater weight in the event of temporary system problems error status documentation could limit the inferences that auditors could reasonably make in these circumstances.

These evidential prioritisations were complemented by criteria established by the US *Center for Research Libraries* [CRL, 2007] which detailed informal ratings as follows:

5. Compliant with all metrics fully and consistently, and able to provide complete, up-to-date documentation of all systems and procedures and certifications of system security.

4. Compliant with all metrics, and able to provide complete, up-to-date documentation, but with minor inconsistencies in areas that are not likely to lead to systemic or pervasive defects.

3. Compliant with all critical metrics, and able to provide complete, up-to-date documentation of major systems and procedures, but with minor inconsistencies in areas that are not likely to lead to systemic or pervasive defects.

2. Compliant with all critical metrics, with a minimum of inconsistencies in areas that might lead to minor defects of a systemic or pervasive nature; documentation is complete and updated on a periodic basis.

1. Compliant with all critical metrics, with a minimum of inconsistencies or deficiencies in areas that might lead to minor defects of a systemic or pervasive nature.

Evidence retrieval at times faced difficulties. The most obvious was the reluctance of repository staff to fully cooperate with the process, and their consequent resistance to the release of documentation, disclosure of information in interviews or provision of auditor access to systems, whether due to reasons of sensitivity or otherwise. Those that ultimately participated were required to confer upon auditors discretion to request any documentation relevant to the assessment - for those the incentives of improvement and the opportunity to contribute (and help shape) the development of community standards for evaluation were sufficiently compelling. Non disclosure of the specific practical aspects of each audit was an unfortunate commitment required by several participants (although this did not affect the shaping of the outcomes).

Other barriers were less easily overcome however, particularly when the context for such work was an international one. Documentation was typically written in native languages, which represented a significant barrier. Similarly the role of auditor demands considerably diverse skills and knowledge: in most instances, given the breadth of coverage these audits demand, from the point of view of any single auditor, some aspects of documentation, whether technical, financial, legal or archival could prove difficult to interpret.

3.2.3 The National Library Repository

Background

The *e-Depot* repository at the Dutch Koninklijke Bibliotheek documented in the first case study had a formal relationship with two internationally established Dutch publishers which was integral to the establishment of its e-journal preservation storage resource. Successful negotiations with Elsevier and Kluwer concluded in 1996, and subsequent collaborating publishers include Oxford University Press, Taylor and Francis, Sage, and Springer. At the end of 2005 the repository accommodated around 3,500 e-journal titles, comprising some 5 million articles and totalling around 6.3 Tb.

Prior to the assessment the repository staff were encouraged to complete a self assessment based upon the draft audit criteria and complete separate financial and technologically-focused questionnaires. Further documentation and comments regarding the audit questions were also submitted. The fact that several responses were not in the English language was illustrative of one of the challenge of undertaking international audits.

Methodological Notes

Despite the large volume of documentation that was made available to us during the on-site process, the most influential evidence was solicited via themed interviews, structured to broadly correspond with the sections of the RLG/NARA audit check-list, and involving a selection of relevant staff. The majority of the conclusions from this audit were drawn from a combination of written self-assessment (mainly corresponding to the RLG/NARA check-list and completed and submitted to our on-site activities), and the series of staff interviews. Only the self-assessment documents and a further short document describing some criticism of the RLG/NARA check-list were available to auditors prior to their arrival on-site. Interview questions were primarily designed to address points of uncertainty within the check-list responses, and the specific criticisms of the metric that had been presented. The general process throughout each session was to question only those responses that appeared to demonstrate non-compliance with the check-list's prescribed metric or that questioned the value of those metrics. Significantly less time was spent questioning those responses that suggested best practice had already been implemented. Finally, notwithstanding the insights afforded during a short tour of the archive, the audit offered no opportunities for staff to demonstrate the operation of the system or for auditors to see or question the specifics of actual physical processes. In that respect then, no primary evidence about the hands-on digital object management undertaken within the archive was available.

Documentation was made available on arrival at the repository, and appeared both extensive and persuasive. However, as it was subject to little formal analysis, and was rarely responsible for the provocation of questions during interview sessions one must be cautious of overstating the role it played during the audit. Within the concluding presentation auditors were congratulatory about the level of documentation that the archive had accumulated and made available. However, one might assert that, from a formal, analytical perspective, its legitimacy was granted based on little more than its quantity. A trustworthy repository will have extensive documentation, but it is not necessarily true to say that any repository with extensive documentation is trustworthy.

The role of the check-list itself was limited. A more effective strategy would combine the different kinds of evidence available (interviews, documentation and observation of practice) to evaluate the satisfaction of the requirements of individual sections, with individual metrics providing opportunities for more specific analysis. The audit of the *e-Depot* dwelt mainly on the pursuit of evidence to explore points of ambiguity or organisational failings perceived only from the initial self-assessment. In this sense, the approach was reactive, and even then mainly to points of failure - compliance was frequently determined on the basis of little more self-affirmation, whereas this ought to have been admissible as just one of several kinds of evidence that confirmed the pedigree of the repository. The self-assessment documents un-

doubtedly represent a valuable starting point for the audit process. However, rather than treat them definitively, the on-site process may have benefited more from being an exercise in determining their legitimacy, employing interview, document analysis, observation and experimental techniques to establish a sense of the extent to which each metric's requirements are satisfied.

3.2.4 The National Archive's Data Centre

Background

The National Archive's Data Centre preserved and provided online access to archived digital datasets and documents from UK central government departments. Data stored remained in the legal custody of the National Archives, but were managed by the Data Centre, who provided preservation and dissemination services. The datasets accessioned varied tremendously. Many were decommissioned databases that had been superseded. In other circumstances they were snapshots of running servers still in production use. Similarly, the types of data, their size and their subject matter exhibited considerable diversity.

Limited access was offered to the Data Centre's staff intranet and wiki and a range of additional materials were supplied prior to and during the visit. These included several procedural manuals, organisation charts, staff job descriptions, a selection of data transfer documents, the Data Centre contract's service level agreement and a copy of the service's Business Continuity Plan. In addition, several Data Centre staff contributed to the completion of a self-assessment against the RLG-NARA criteria and these responses provided a useful basis for subsequent enquiry. Additional recourse was made to the Data Centre and service websites, which provided further insights in some areas.

Methodological Notes

The preservation service provided at the Data Centre was undertaken as a contractual obligation under agreement with a national archive, and as a consequence the documentation available within the organisation was perhaps second to none, at least within the context of this pilot process. Perhaps the greatest lesson learned in this exercise focused less on the specifics of the available audit tools or methodology than on the preparatory work that repositories might undertake to facilitate a successful assessment. As a prerequisite and consequence of its contractual relationship, the outcome of a competitive tender process, the Data Centre maintained a tremendous body of documentation relating to almost every part of repository operations. Within an impressive catalogue of policy and procedure manuals, issues such as digital object acquisition, digital preservation, information security, staff

training, legal responsibilities, policy review and access were each explained in considerable detail, enabling an auditor to quickly gain a comprehensive picture of the repository, which could be immediately compared with the criteria implicit within the audit check-list. Similarly, the repository had willingly undertaken a challenging process of certification under the *ISO 9000* series of standards, relating to quality assurance across every aspect of the organisation. It was also already subject to detailed inspection by its primary client to determine the suitability of its physical infrastructures.

The check-list self assessment document returned by the Data Centre was of particular value since unlike every other received during this programme, it was completed by a broad range of staff representing every level and repository function. This provided an opportunity to confirm that repository policies had absorbed into the consciousness of all staff, and not just a single overseer. Discussions with staff indicated a broad and in-depth awareness of policy and procedure in every area and an organisational cohesion consistent with documentation, enabling auditors to more easily take demonstrations of repository functionality at face value with less need to adopt a more adversarial investigative approach. By embedding a culture of assessment, improvement and transparency firmly within the repository, the demands of inviting external auditors to perform further assessment were minimised. In the Data Centre, these characteristics were implicit as a management objective, and to the fore to facilitate the effective running of the repository. An increase in the organisation's 'auditability' appears to be a resultant side effect. The goal of auditors is to identify good management practice; the goal of repository staff is to manage their repository effectively. Both are consistent with a requirement for a formally documented and internally expressed self awareness.

3.2.5 The UK Research Council Data Centre

Background

The UK Research Council Data Centre in the third case study provided electronic archiving facilities for a range of data producers, perhaps most notably the research projects funded by the Natural Environment Research Council (NERC) but also significant international meteorological organisations such as the UK Met Office and the European Centre for Medium-range Weather Forecasts. Since 1985 it had grown to represent the NERC's primary and sole data centre for data originating from atmospheric research, and consisted of over sixty terabytes of data in a variety of formats.

Complete access was afforded to the Data Centre's staff intranet and wiki which incorporated a comprehensive *Operations Manual* and associated documents (including a prototype risk register) and a range of additional materials were supplied prior to the visit. These included the Natural Environment Research Council [NERC, 2012] Data Policy Handbook, with as-

sociated guidance notes and example data policies; an organisational chart with job descriptions; an example data protocol document, the service level agreement that documented the datacentre's mandate, scope, deliverables and funding period; excerpts from presentations depicting lines of management within the Data Centre and associated NERC reports. In addition, the centre's Curation Manager completed a short self-assessment exercise based on an earlier draft copy of the RLG-NARA check-list and this was utilised as a foundation for significant parts of the subsequent investigation. Additional recourse was made to the Data Centre website which provided further insights in a range of areas.

Methodological Notes

The audit of the UK Research Council Data Centre was facilitated with the availability of substantial and varied documentation, interview subjects that were both responsive and forthcoming and an organisationally enthusiastic attitude to the demonstration of practical processes undertaken during the archive's normal operation. Near comprehensive documentation was supplied to auditors prior to the visit (including a self assessment based on the RLG/NARA check-list), offering an opportunity to establish considerable foundational understanding of the organisation, its contextual spacing, the nature of its business and its digital holdings, its technological infrastructure and the services and functionality it is committed to providing. In isolation, this falls far short of representing conclusive proof of the trustworthiness of the repository (although its very existence provides a persuasive indicator of managerial effectiveness). As a starting point however, the documentation, which included extensive details about the archive's systems and procedures, technical architecture, staffing, funding, depositor relationships (including legal relationships) and risks represented an essential starting point. Equipped with an initial world view, we could spend our limited time on-site seeking confirmation; staff interviews would provide compelling insights into whether the documentation was representative of real day-to-day practice and observation of the completion of tasks, interactions with the system and management process would prove even more conclusive. The check-list structure informed the organisation of interview sessions. However, with the extensive evidence already submitted and considered, interviews had some adversarial characteristics. The exercise became akin to cross examination, where truths within the documentation were corroborated, gaps were gradually filled in and concerns confronted. Every interview room had facilities to access the archive's computer system, which facilitated both the demonstration of any concepts or processes that arose as well as the recovery of any additional electronic documentary evidence, which could be checked whenever referenced in conversation.

As the audit continued, the evidence-based focus gradually narrowed; our initial goal to accumulate a broad understanding of the archive evolved into increasingly granular level

of inquiry, culminating in the determination of whether selected individual criteria from within the RLG/NARA check-list had been satisfied. In that respect the check-list provided a pivotal structural support: its broad scope determined the parameters of both initial general investigation and its individual metrics the focus of more specific subsequent assessment.

3.2.6 The US State Digital Archive

Background

The US State Digital Archive was established to provide long term preservation archival services for digital materials originating from any of Florida's state University libraries. In 2002 a three year public grant prompted the archive's development. Preservation functionality was prioritised ahead of access features, and consequently the organisation operated as a principally "dark archive". Its commitment was that all files deposited by agreement with its affiliates remained available, unaltered and readable from media, with preservation achieved using the best format migration tools available. Its technological foundation was a set of scripts and programs, which at the time of the evaluation was due to be released under an Open Source license.

Extensive documentation was provided in advance of the assessment, comprising job descriptions, organisational charts, service level agreements and a completed self assessment check-list. Policy documentation was mostly available from a comprehensive policy guide document. Technical information contained in a corresponding document describing the bespoke archive software system. Financial information was made available to auditors on site.

Methodological Notes

The US State Digital Archive was the final repository to be subject to assessment in the initial *Digital Curation Centre* pilot programme; the adopted methodology was therefore quite mature by this stage. Like many of the audited organisations, the US State Digital Archive submitted a self assessment document based on the RLG/NARA check-list in anticipation of the audit visit, and this proved again to be a useful source of insights. By this stage of the programme it was clear that the self-completed check-list was of greatest value when read after more neutral documentation; responses amounted in some respects to a dialogue between the repository and the auditor. Accompanying the self assessment document was a variety of additional documents, which included organisational information, policy information, software specifications, and example deposit agreements. The audit began not with the arrival on site, but upon receipt of this documentation, with a thorough analysis providing numerous

insights into the repository infrastructure that would be subsequently explored. Two days of on-site activities provided an opportunity for discussions and demonstrations of system functionality and work-flow, and these highlighted a number of implicit concerns. The auditors and audit methodology were by this point sufficiently well established that although interviews were still structured according to the broad categories of criteria within the check-list, it was much less necessary to labour over every specific criterion. A more fluid process evolved; although checks for completeness were made by reference to the check-list at the conclusion of each session, interviews were mainly structured by the evidence provided by the repository. The value of a more bottom-up process of evaluation was increasingly evident.

The team based at the US State Digital Archive appeared in a number of cases to be broadly conscious of their organisational shortcomings but the audit exercise enabled their systematic encapsulation and expression, and allowed them to be more effectively addressed. We acknowledged during this assessment the value of experience accrued during the previous pilots. In isolation, the RLG/NARA check-list criteria offer a useful structure around which to base assessment, and a number of clues about the shape that repository activities might best assume. However, the specific details of how repositories should conform to these criteria were not really expressed within the check-list. Certification is ultimately about comparison, using objective metrics, and with peer organisations. For this to work there is an implicit requirement that tools and methods must support comparability. By exposing ourselves to a range of environments that purport to satisfy the check-list's criteria, we are equipped to determine optimal means of check-list conformity. An additional level of granularity can be expressed, whereby metric conformity or non-conformity is no longer an atomic consideration. Instead, we can determine the extent to which specific practical approaches are capable of satisfying individual criteria, and introduce a notional understanding of what this means in terms of a more universal understanding of conformity. For example, exposed to just a single repository, we may see evidence of provisions whereby staff may request practical training during an annual skills review session, that appear to satisfy metric A2.3 of *TRAC* ("Repository has an active professional development program in place that provides staff with skills and expertise development opportunities."). This may however appear less than satisfactory when we visit a second repository that offers, in addition to an annual skills review session, a system requiring each staff member's line manager to monitor performance levels to suggest appropriate training. The latter approach ensures that any training opportunities that staff members may themselves be unaware of remain available, and is therefore preferable. But without the exposure to a range of implementations that aspire to conformity, it remains difficult for auditors to determine where improvement might plausibly be sought. It might be said that the role of a consultant is to distil broad and varied knowledge, accumulated with considerable experience over a significant period, into advice or services for a client that

lacks the resources to themselves gather that experience. Our role was broadly identical; in order to understand the practical realities of check-list compliance, one must be exposed to a wide variety of implementations. An aspiration to conform is just half of a picture that must also include a practical capacity to conform.

This suggests that the success of the audit is completely dependant upon the availability of sufficiently expert auditors. It is they who must interpret audit criteria and determine what it means in practice to conform. Any opportunities to objectify the process, and convey this knowledge to those within the repository profession should of course be explored. Accomplished auditors are equipped through their experiences to ask telling questions of repositories, which might be understood as *Key Lines of Enquiry*. There is a danger that unless expressed as at least a semi-formal framework within which evidence can be gathered and assessed, the audit process may appear to be unduly based on feel, and dependent on the perceptions of specific auditors, which limits opportunities for comparison. With the prioritisation of self-assessment, each repository manager requires access to a body of knowledge that can be an effective surrogate for such experience.

3.2.7 The Cultural Heritage Archive

Background

The Cultural Heritage Archive based at a UK University was an electronic annex to an older physical archive, which itself consisted of several hundred thousand examples of notes, photographs, negatives, drawings, books, catalogues and gem impressions. Three databases represented the bulk of the electronic content. Much of the the archive's electronic content acquisition was proactive, with staff encouraged to actively pursue newly available catalogue information and photographs for accession.

Little documentary evidence was available to facilitate this final assessment, a point which is returned to a number of times during the case study, but in advance the archive supplied a user manual for the bespoke database system that acts as a technological foundation for the archive, and the *Technical Appendix* from an AHRC project grant application detailing some aspects of project management, a commitment to preservation and the text from the now defunct Arts and Humanities Data Service deposit waiver application, which entitled the archive to maintain and preserve its own collections. Without this waiver agreement AHRC funded projects were required to deposit collections with the central management resource.

Methodological Notes

The audit of the Cultural Heritage Digital Archive was undertaken with little documentation available throughout the process. A small selection was gathered prior to the on-site activities, mainly encompassing descriptions of system procedures and functionality and some documentation describing the archive's practical commitment to preservation. What distinguished this audit most significantly from the others described in this paper was the manner in which the RLG/NARA check-list was employed. In each of the other examples archive staff had familiarised themselves with the document's metrics and provided, in advance, a series of self-penned responses. In the case of the Cultural Heritage Digital Archive, the archive was only comprehensively exposed to the check-list during the on-site activities. In the absence of sufficient alternative documentation, the discussion with staff closely reflected the check-list's structure; in that respect the on-site activities resembled a measured, facilitated self-assessment exercise. A useful consequence was that auditors were afforded an insight into the check-list's applicability, relevance and usefulness within an archive that had little prior knowledge its metrics. Efforts to obtain practical insights into the repository operations were made - each interview was conducted with a computer workstation nearby, ensuring that the physical processes of ingest, archival storage, data management and access could be demonstrated. Typically, the information gathering process began with the check-list requirements. Posed as questions, one or more of the individual metrics would encourage discussion from appropriate individuals, which included management, technical support individuals and object management specialists. In the absence of comprehensive documentation little recourse was available to printed matter, and instead auditors would frequently request that further illustration be provided by way of practical example within the Archive's digital object management system.

The absence of extensive written documentation, particularly prior to the on-site activities, hampered our efforts to obtain a comprehensive and definitive assessment of the Cultural Heritage Digital Archive. One might argue that it was the lack of opportunity for repository staff to familiarise themselves with the check-list's terminology that proved most problematic. But what was clearer was that the check-list could not define, legislate or reflect a new form of best practice; rather its role is limited to reflecting and encapsulating broadly accepted truths. An effective sequence for evidence gathering began to emerge - the check-list provided an initial focus for repositories, but as auditors, our primary starting point had to be documentation. The absence of documentation from the Cultural Heritage Digital Archive was symptomatic of the same organisational shortcomings that meant a large proportion of check-list metrics appeared unfamiliar and onerous.

3.3 Findings

This section presents a distilled perspective of preservation practice within the organisations that we audited, with excerpts from each corresponding case study included to illustrate and provide evidence of practice. Structurally, the overview reflects our audit check-list which provided the intellectual foundation. Each broad section of analysis concludes with a table summarising a relationship between core audit issues (essentially derived from the available check-list literature) and one or more (most commonly several) associated and incorporated goals. This is illustrative of the first phase of transposition of audit findings to a structured information network. From these goals we continue to deconstruct the preservation activity (in the following chapter).

Organisational Infrastructure

This section includes coverage of those organisational aspects necessary to operate the preservation or data management service. Typical considerations include mandate and institutional commitment; organisational viability and sustainability; legal and regulatory legitimacy and policy infrastructure efficiency and effectiveness.

We observed a range of organisational infrastructures throughout the sample set. They included individual services positioned within a single institutional setting, commercial companies operating under public sector contract, centralised, US state funded services with several University clients and repositories supported primarily with short term research funding.

Mission and Mandate

Mission and mandate were typically defined at a very high level within a mission statement or similar succinct message of organisational commitment. Exemplary practice was identified in the **US State Digital Archive** which outlined eight key responsibilities. These described the major constituent parts of activities, and their relationships with their depositors. These were to implement bit-level or full preservation of submitted content (determined according to preservation agreement); to restrict those who were authorised to deposit or sanction the withdrawal or dissemination of content; to provide detailed ingest and error-related feedback for every submitted package; to preserve 'original' files as submitted, maintaining integrity, viability and authenticity; to employ appropriate preservation strategies to persistently maintain a usable version of each file for which full preservation was sought; to provide dissemination information packages (DIPS) on request; to provide appropriate reports to affiliates for management purposes; and to ultimately achieve and maintain certification as a trusted digital repository, when the infrastructure to support this becomes available.

The Archive's mission statement was "to provide a cost-effective, long-term preservation repository for digital materials in support of teaching and learning, scholarship, and research". This was endorsed by a Data Centre board, lending the commitment a weight of legitimacy.

Perhaps as a consequence of its contractual basis, the **National Archive's Data Centre** lacked a true mission statement that was sufficiently succinct and widely distributed. Instead, its contract (and corresponding legislation) made explicit business aims. This remained inaccessible to wider stakeholders and therefore was of limited widespread value. The Data Centre website's 'About' page contained an expression of mandate and objectives, but failed to explicitly define the legislative relationships that justified its existence. Further background was freely accessible from the web pages within data transfer overview documentation, which described in more detail the applicability of legislation, the obligations arising from it and the particular data that the Data Centre was responsible for preserving. It was thought preferable to have this information presented in a more prominent location, encapsulated within a succinct and clearly defined mission statement. The Data Centre's parent service did have its own mission statement ("the service aims to be the preferred provider of information, communication and learning technology service across the public sector") but, while far from incompatible with the Data Centre's commitment to long term data management and access provision, was hardly synonymous.

Organisational, Governance and Policy Best Practice

The **UK Research Council Data Centre** provided useful evidence of organisational and governance best practice. There, a steering committee was responsible for advising on programme development, and ensuring the implementation of associated data management plans. A data management sub group, including representation from the Data Centre (or other appropriate data centre(s)) was convened to support a coordinator in these activities. The steering committee was responsible for ensuring that data management was carried out effectively (by providing adequate support and resources during the programme); an appropriate data management plan was created; a realistic proportion of the overall programme budget was devoted to support data management; and all holders of programme awards complied with the data management policy of the programme, as outlined in the data management plan, (although some scepticism was expressed in terms of enforceability). Circumstances had arisen where funded researchers had failed to deposit content, or had provided incomplete datasets. The Data Centre was not involved in the grant process, and therefore could not compel deposit where it would be worthwhile. It was suggested that the Data Centre would have benefited from access to a more detailed online register of grant awards, to support their pursuit of worthwhile data. In practical terms, discussions suggested that the Data Centre was simply not offered a great deal of data; however, it was also suggested that this

situation was changing, and therefore it was imperative that the Data Centre was equipped to cope with a consequent upsurge of deposits.

Organisationally, the Data Centre existed within a research institution setting under a service level agreement that described core services, infrastructural developments, and support for other data centres operating on behalf of a common research funder.

Less clarity in organisation was evident in the **Cultural Heritage Archive**, where qualities of individuals were often indistinguishable from the service as a whole. Individuals' dedication, self-motivation and wide ranging contacts had offered a degree of operational security to the Archive during its lifetime but there was a notable risk that the departure of key individuals could threaten the organisation's ongoing viability. Partly this was an organisational concern - at the time of the assessment the post of Principal Archivist existed (and was therefore centrally funded by the accommodating University) only for as long as the existing postholder remained in place.

Succession and Service Continuity

The **US State Digital Archive** identified the need to engage with other organisations to meet the range of challenges that prejudiced the integrity of its digital assets. Building relationships was expected to enable the conception of succession or escrow arrangements, further remote storage of backed up materials, and ultimately, assuming the emergence of their adopted technology platform as a widely used tool, collaboration in systems development and format description.

Two 'options' existed in the event of the Archive's cessation of operations, as described in the Archive's *Policy Guide*. The first was to simply return content to the appropriate depositor, in the form of a Dissemination Information Package (DIP). This was practically viable, although one may question whether returning content was a compelling succession arrangement. In addition, the success of this approach presupposed that the Archive's operations would be maintained for a period that was sufficient to permit a comprehensive dissemination. The second option, which was at the time only planned and not practically implementable, was to send content to an alternative preservation repository in a DIP exchange format (the precise format of which was yet to be conceived). While in principle much closer to a true succession plan, the practical barriers of no format and no repository greatly impeded its viability. However, discussions were already ongoing with a partner library on the opposite US coast with whom research monies were being sought to collaboratively define an appropriate exchange format. It was hoped that this joint endeavour might be extended to represent a reciprocity agreement capable of facilitating succession and the remote accommodation of content.

In addition, given the anticipated public release of the platform software, it was quite feasi-

ble to suggest that if widely adopted, the barriers to information exchange across common systems would be reduced, and that many of the practical or technical difficulties associated with succession planning would be similarly mitigated. Nonetheless, it was acknowledged that several of the barriers associated with succession planning and feasibility are not technical, and instead are based in inter-organisational, political and legal concerns. It was suggested that the Archive continues to pursue a formally expressed collaboration to seek formal assurances for succession and service continuity, and to define means for effective inter-organisational digital object exchange. Although it was suggested that funding was reasonably assured for the foreseeable future there was no evidence of a legal or regulatory compulsion upon the state to continue to support the Archive. It was suggested that if such assurances could not be obtained then this should be considered and documented within an overall risk mitigation strategy.

Succession arrangements at the **Cultural Heritage Archive** were identified as being vague if not non-existant. A perception existed that should the Archive fail, the University which provided the operational context for the Archive would assume custodial responsibility for the archival holdings and ensure their continued and ongoing availability, such was the extent to which their value was recognised. However notwithstanding this confident attitude, there were apparently no formal assurances that this would be the case.

At the **e-Depot** the issue of software escrow was subjected to some scrutiny during our audit, and also emerged as an area of some concern. The repository's technical infrastructure was essentially a proprietary system, consisting of both off-the-shelf and bespoke software developed by IBM. No escrow agreements were in place, which might leave the Library in a dangerous position in the event of the withdrawal of the *Tivoli* software suite or the discontinuation of its support. Dismissing such concerns, staff explained that the issue was given significant consideration, but that IBM were deemed the only adequate supplier given their technological requirements. They perceived the likelihood of vendor collapse as extremely minimal, and irrespective of this, since the data and system software were separable, such an eventuality could be survived until appropriate alternative software became available. Additionally, it was argued that since a number of large global banks used and rely upon the same IBM software, there was sufficient international weight to ensure that IBM would continue to maintain the software in its current, or similar form.

Repository Staffing

The **US State Archive's** biggest staffing priority was to make new appointments, with the most high priority being a manager for the Archive, who could engage with affiliates and plan and direct the future administrative and operational direction of the Archive. The Archive's primary goal within the near future was to increase affiliate numbers and the quantity and quality of content within it and to enhance its reputation.

The **UK Research Council Data Centre** exhibited sufficient number of staff, and evidence suggested that the organisation had a broad understanding of the implications in terms of workload and personnel requirements of a move towards a more 'OAIS compliant' infrastructure. However, the organisation would have benefited from committing extra resources towards preservation and archival roles, with its existing staff primarily consisting of subject experts. Archival lifecycle activities such as ingest, data and metadata management could have been assigned explicit ownership by individuals or roles within the Centre. Furthermore, legal advice should have been at least solicited given the doubt that to some extent surrounded agreements with depositors and content creators. These recommendations were primarily focussed on enhancing the repository's scalability.

Training was available for staff, although it was presented in a fairly ad hoc fashion, based mainly on staff demand. At the time of our audit, recent training had been offered in scripting languages such as *Perl* and *Python* whereby several copies of prominent learning texts were purchased and staff were encouraged to learn as a group. There were however few mechanisms in place to identify knowledge or skills gaps and therefore staff were expected to maintain a degree of consciousness of their own shortcomings, which was perhaps unrealistic. Better structured training programmes could have been developed and associated with particular roles within the Centre, in order to ensure the effective development of staff.

Most repositories enjoyed staff stability - the **Cultural Heritage Archive** had been successful for over twenty-five years with little evidence of service disruption or data loss appeared to relate, to at least some extent, to the fact that the Archive had enjoyed tremendous staff stability throughout its period of existence. It was identified in several audits that although low turnover of staff implied that those employed were both experienced and competent, it also meant that staff were more expensive, a consideration in terms of financial sustainability.

The **National Archive's Data Centre** exhibited an exemplary approach to staffing. The heads of Application Services and Digital Archives, in association with senior staff, were responsible for allocating staff resources, and monitoring appeared to be undertaken to detect staffing shortfalls based on the requirements made explicit within the the Data Centre contract. To this end, there appeared to be adequate staffing provisions at the time of our audit.

Professional staff development within the service was covered by both in house training policies and procedures, and by the wider infrastructure provided by the parent university. With respect to the latter, job appraisal schemes provided opportunities for staff and line managers to jointly mould personal and professional development, to identify their own training needs and to arrange for them to be formally addressed. A training budget existed for the Digital Archiving Department and given the involvement of the service in a variety of other training activities there were ample opportunities for specific training. Furthermore, the Data

Centre's Inhouse and Training Procedure Manual described fairly comprehensive processes associated with staff induction, training needs assessment and review and training delivery.

Community Engagement

The **UK Research Council Data Centre's** designated community was defined more narrowly than its overall potential user base, a target community comprising around 45 per cent of total end users. The Centre's commitment (in terms of preservation) was limited to making materials usable for its stated group of specialist scientists. Although an internal understanding of the designated community of the Data Centre was demonstrated during discussions with the Curation Manager, there was no evidence of a centralised, published definition. Similarly, little evidence was available to suggest that formal mechanisms were in place to monitor the evolution of this community, although the relationships that were maintained with depositors, who were in turn likely to be end users, were close, and provided insights into latest developments. However, of the 8000 or so registered users of the Data Centre web access system, a small fraction (less than 5 per cent) were responsible for depositing.

There was little to indicate the existence of formal mechanisms to react to accumulated evidence of community evolution. One anticipated an ad hoc approach to dealing with changes in community expectations.

At the **e-Depot**, two categories of designated community were identified. The first was primarily publishers, what the Library described as their business to business profile. This was expected to be extended in the future with the addition of additional cultural and heritage depositors; the development of formal service level agreements would enable and facilitate these emerging relationships. The second category of relationship was with end users, described as the Library's business to consumer profile. This relationship was less explicitly stated, and few formal guarantees were offered to those seeking content as to what was available and the infrastructure that was available to support its delivery. Nonetheless, Library users could remotely access catalogue information about publications, access resources onsite, or access faxed or printed copies of articles in libraries elsewhere as inter-library loans. In turn the Library was compelled by contract to provide a 'minimal level of functionality' which included bibliographic searches, publisher publication listings at the volume and issue level, listings of issue content, article views, copyright information views article or "smaller than article components" (e.g., metadata) downloads consistent with the terms of each contract.

Suitability of Policies and Procedures

The **National Archive's Data Centre** exhibited great leadership in terms of policy and procedural formality. A range of policy and procedural documentation was available, which included Security Procedures, Transfer Procedures, Site Exchange Procedures, Digital Preser-

vation Procedures, Paper and Paper Preservation Procedures, Helpdesk Procedures, Inhouse and Training Procedures, Closed Data Access Procedures, Data Protection Act Procedures, Finding Aids Procedures, Contingency Planning Procedures and Style Guides. Each set of procedures was realised in one or more associated documents.

The Data Centre also maintained policies that described in detail necessary steps to introduce, review and retire procedures. All the Data Centre procedures were documented in appropriate manuals, supplemented where necessary by detailed working instructions. These could be changed in three ways. The first was where the Data Centre Service Manager had identified the need to revise a procedure manual, and this duty could be delegated to a member of staff. A draft revision was then presented and discussed in a physical meeting, by email or in the Data Centre Usenet discussion forum. Amendments were then actioned, prior to the creation of a final draft, which was then approved by the the Data Centre Service Manager, and linked to from a central HTML index page. Staff were informed that all prior versions should be immediately disregarded. Alternatively, any staff member could suggest changes at any time. Suggested changes were circulated via email or *Usenet*, comments were aggregated and a brief report conceived, for discussion at a subsequent meeting. Final revisions, and the replacement of earlier versions were actioned as above, subject to the Service Manager's approval. Finally, all procedures manuals were subject to ongoing review, on an at least annual basis. The introduction of new procedures was conducted on a similar basis. Once more, these could be prompted by the Service Manager, an independent member of staff or during a regular procedures review meeting.

The quality assurance and consultation procedures associated with the conception of new procedures were broadly equivalent to those associated with amending existing procedures. A new procedure could be justified by the introduction of a new procedure that extended the range of work; a major change in the way a procedure was undertaken; sufficient numbers of small changes to an existing procedure to necessitate a wholesale review or its granularisation into multiple procedures; or, finally, the insistence by a staff member that a new procedure was otherwise necessary. In most cases of procedural change, the existing manuals would simply be updated. The benefits of this approach to procedures management were clear. Each of these three methods for introducing and modifying procedures was intended to ensure that they remained relevant, representative and comprehensive in their coverage. Procedures were allowed to both dictate the work undertaken within the Data Centre, and reflect emerging working practices that may reveal themselves and optimise the repository's efforts.

Further policy described the mechanism for retiring redundant procedures within the Data Centre. Again, this could be prompted by any staff member, or in discussions as part of a procedures review meeting. In the event of a procedure being nominated for retirement, a staff member was delegated the responsibility for collecting comments to support or oppose the motion (or suggest merely revision of the procedure). These were amalgamated

into a report which provided the basis for subsequent discussion and a final decision by the the Data Centre Service Manager. Retired procedures were moved to a special section of the staff intranet, with a note describing the fact that the procedure did once exist and had been superseded. Details of any procedures manuals that did supersede retired procedures were also recorded. Procedures could be retired if they were classified as redundant; this could incorporate situations where working practices had changed to the extent that the procedure was no longer relevant; staff had suggested that the procedure was no longer relevant; procedures corresponded to work areas that were no longer active; sufficient smaller-scale changes necessitated reformulation of policies or the creation of more granular policies; or where procedures were subsumed within an existing procedures manual.

As well as participating in this pilot assessment, the Data Centre illustrated its enthusiasm for policy excellence by obtaining *ISO 9000* series quality assurance certification.

A less successful approach to policy was evident at the **Cultural Heritage Archive**. Undoubtedly the greatest concern with respect to policies and procedures was the lack of transparency, accountability and documentation that surrounded much of the Archive's efforts. Notwithstanding the clear indicators of success in terms of funding consistency, user numbers and community reputation, there was little in place to facilitate understanding or sustainability, or to enable a newcomer to continue to build on the preceding efforts. The Archive adopted a bullish approach to user and legal accountability, where the fact that services were available free to end users appeared to be the basis for complete limitation of liability. The primary role embraced by staff appeared to be to add value to the materials, with preservation of lesser concern. This view did not appear to conform to terms of funding which compelled the Archive to act as custodian and preserver of digital assets arising from funded activities; indeed the Archive's own deposit waiver applications offered a commitment to undertake these activities as an alternative to a funder appointed custodian. Information integrity measures were defined, but not formally documented, although it seemed that these were exclusively related to the creation, acquisition and ingest of new content. Once assets become resident within the database there was little evidence of ongoing integrity checks. No mechanisms existed to provide on demand measurements of information integrity. Clearly, policies ought to have been more rigorously conceived, documented and formalised, and circulated widely among repository staff to ensure widespread understanding.

Contracts, Intellectual Property and Legal

The **Cultural Heritage Archive** faced a number of potential legal concerns, which to date had been managed, but, it was considered, may have threatened the ongoing viability of the Archive. The legal status of much of the material within the database remained unclear. No formal relationship was maintained with information publishers or producers; instead the Archive's principal researchers operated quite independently, acquiring digital materials

from analogue sources based mainly on their availability. Legal guidance had been sought in the past with regard to the dissemination of copyright controlled image materials; the suggestion then was that since the chance of rights holders seeking redress was negligible the Archive needn't deviate from its existing practice. The repository administration offered three main justifications for continuing to distribute copyrighted materials - the lack of charge levied by the Archive for access to materials; the excellent track record that the Archive had established as an authoritative source; and the community interests that were being served - it was argued that in the absence of the Archive there would be no way for these demands to be met.

The lack of appropriate contracts or deposit agreements, and the legal questions surrounding data gathering procedures were a concern, and almost certainly represented a risk that could have been addressed more systematically. The restrictions imposed on usage (content was free for personal and academic usage; copyright notices; digital watermarks and technology used to encode copyright holders name to images) would not necessarily satisfy content creators in the event of legal objections. Reciprocal agreements were sporadically in place, enabling the Archive to digitise content in exchange for appropriate credit on their web site. These could have been better formalised in order to limit the risk of liability. Even in those circumstances where producers or publishers directly interacted with the Archive no formal written agreements existed. Irrespective of the fact that such legal challenges might have been overcome by withdrawing content, the impact in terms of wasted staff time could have been considerable. The Archive described only positive feedback from publishers, who according to anecdotal evidence regarded the Archive's use of their materials as beneficial. Nonetheless, none would agree to waive copyright, and arrangements would have benefited from being more formally expressed.

The lack of legal controls was also problematic due to its impact on funding requirements, notably those imposed by the Archive's primary funder. The Archive was understandably reluctant to deposit content given the rights issues described above; it was thought that such behaviour might imply ownership. Instead the Archive was required to commit to preservation activities which added to their core objectives to present content. However, discussions suggested that preservation remained a very low priority for the Archive, despite the fact that continued funding was contingent upon it.

The **UK Research Council Data Centre** appeared to operate without formal deposit agreements, and where agreements existed there appeared to be a somewhat loose approach to contractual management. At least one relationship with a strategically important data provider offered a good example. Data originating from research council-funded research was subject to more formal terms, and these were outlined within a data policy and associated programme-specific data protocols and policies. These appeared to be negotiated at a level beyond the immediate Data Centre management (although appeared to involve at least

some degree of consultation with the Data Centre).

Further rights-related questions were associated with ambiguities over the Data Centre's right to change data, which might be regarded as 'preservation rights'. The internally held perception was that the Data Centre probably did not have the rights to change, or even reformat data. Instead the rights were to share data. In practical terms though, it was argued that irrespective of the existence of these rights, the Data Centre would not change data because of the question of trust (the internally held perception being that users did not trust the Data Centre to alter content), and the comparatively small number of widely acknowledged data formats: both of the principle formats utilised by the Data Centre, *NASA AMES* and *NetCDF*, appeared stable and widely used. Ownership of data was similarly unclear; NERC was committed mainly to making data public, and therefore questions of ownership were given considerably less priority. Discussions and analysis of example NERC funding agreements suggested that NERC was unlikely to own data generated from funded research since this was not expressly stipulated in the grant award documentation. Notwithstanding such ambiguity, it was claimed that the Data Centre might not be permitted to make data public if there was a notable associated revenue stream that might be exploitable. In the case of NERC data, contracts existed between NERC and the appropriate universities/researchers providing (and in many cases owning) the data. No direct agreements were formed between the Data Centre and depositors, with mutual responsibilities mainly encompassed in a data protocol corresponding to each NERC programme. There were questions about the extent to which the Data Centre was legally entitled to make preservation management decisions based on their perceived value of archived data.

Monitoring of intellectual property rights existed in embryonic form in the *Operations Manual* on the Data Centre wiki, but this fell short of the "comprehensive overview" demanded by our audit check-list. Nevertheless, it was acknowledged that this issue could be taken to extremes, and *ISO 9000* series certification might be regarded as necessary in order to conform to the check list. However, within this context, it was felt that to make such demands of the Centre was not helpful and ultimately unnecessary. Of potentially greater relevance, conversations revealed that no formal policies or procedures existed relating to requirements arising from Data Protection and Freedom of Information legislation. The Curation Manager's view was that neither Act presented any problems or legal incompatibilities with existing Data Centre practice. With respect to Freedom of Information, his rationale appeared to be that existing Environmental Information Regulations superseded FoI, encompassing all of its requirements (and more). However, in order to ensure that everybody at the Data Centre was aware of legal issues and the appropriate approaches to resolve any associated concerns, it was suggested that written guidance should be developed. This was especially true for those members of staff responsible for answering user queries and providing user support. Legal doubts were apparently quite widespread and the then current non-formalised

approach, which seemed to be based on continuing until legal challenges arose, was regarded as unsustainable. Contracting a lawyer to provide advice was recommended. NERC had established some precedent with respect to legal issues, when it provided the Data Centre with legal input for the purposes of drafting a limitation of liability statement, and in addition the Data Centre's parent organisation had in house legal expertise available.

A final concern in this area relates to the current system workflow evident within the Data Centre. Control of access to datasets was built into the access system/interface, with those responsible for information ingest responsible for determining the appropriate access level for particular datasets and content. There was therefore perhaps scope for concern when dealing with atypical access rules. Nevertheless, this was a small concern, as the system demonstrated sufficient fluidity to suggest that it could be altered to reflect emerging requirements.

Given its partnership-based approach, the legal responsibilities of the **US State Digital Archive** itself were considered limited. According to affiliate agreements, liability for intellectual property rights infringements remained with the depositing organisation, and submitted content would only be accepted after standard agreements had been countersigned. The affiliate organisation was required to commit to being "responsible for compliance with all applicable copyright laws and other laws applicable to deposited materials, and that [they have] the authority to grant to Data Centre non-exclusive rights to copy, display and create derivative versions of deposited files." In the event of legal challenge (which had at the time of the evaluation not occurred), the Archive's policy was to disseminate the content to the owning affiliate and withdraw it from the Archive. If a challenge subsequently faltered then they would replace the object without charge. The only concern associated with this approach was that it might be abused with minimal justification. Since the Archive was operating as a dark archive it seemed that either hoax or legitimate challenges were unlikely, given that it remained impossible for non-affiliate parties to determine the nature of or access the archived content. Nevertheless, with the onus of proof of legality resting on the affiliate, this presented a potential risk.

A final point of legal concern related to materials that were associated with the digital objects, and also stored within the Archive. The Archive's documented 'localization' policy described a process that occurred when a submitted file contained links to other files (such as an XML file which references a DTD or Schema). In such cases, the remote, referenced file was retrieved and added to the archival content (AIP). There were obvious legal concerns, given that affiliates were required to only vouch for the legality of submitted content, not that of any referenced material. This had been earlier acknowledged, and the system was modified (albeit without a corresponding change to the Archive policy documentation) to download only a small format-specific subset of all linked files, most notably DTD and XML Schema files. The alternative would be that every online document cited within a PDF dissertation might be harvested and stored with no permission. A remaining doubt was

whether legal permission was required to store these remaining file types. In all likelihood this would vary on a file-by-file basis, but it was recommended that Archive staff explore, with some urgency, the legal implications of storing each of the linked schemas and DTDs within the digital collections. Liability in such cases may not be assumed to fall upon the affiliate within the current wording of the library agreement (given that it explicitly covered just 'deposited files'). Therefore, if potential legal consequences could be identified these should be addressed by either refining the text of the standard agreement or conceiving and documenting an appropriate policy that alleviates the remaining concerns.

Its contractual origins afforded the **National Archive's Data Centre** a degree of legal protection, since the liabilities were expressly indicated within that contract, and many implicit issues were the parent Archive's responsibility. Similarly, questions about the the Data Centre mission being at odds with its parent University were largely moot since the contract had been endorsed, with the Vice Chancellor countersigning and formally expressing his satisfaction of the Data Centre's alignment with the University mission.

At the end of the the Data Centre contract content would be returned to the contracting National Archive and any remaining local copies destroyed. Either contracting party could choose to back out of the contract giving a minimum of six months notice; there were directions to follow in the event of this happening, but specific details would be negotiated at that time. Generally speaking, conflicting contracts with the commercial sector would be avoided by the service.

The Data Centre maintained ongoing relationships with both government departments and data owners (as well as client and contract managers at the parent Archive) to ensure that its procedures were endorsed where appropriate and that it was ultimately able to fulfil its mandate. Dataset transfer forms changed hands during the initial stages of transfer, following the notification by the parent of data that was to be preserved. Signed by data owners and departmental records officers these provided the means to issue formal authorisation to transform source data, detail parts of data sets that must remain closed or be redacted, and describe conditions for managing transport media. Subsequent receipts issued by the Data Centre further formalised the agreement that preservation would take place, and confirmed the instructions issued by owners and departments. An accessioning tracking system monitored and maintained a record of every interaction between parties and interaction with data.

Intellectual property rights were unlikely to concern the Data Centre too much since it was dealing with public records with an explicitly expressed legislative mandate. Nevertheless, evidence highlighted a concerning shortfall in policy in the event of an intellectual property rights challenge. One staff member recounted in a check-list self-response that the Data Centre had been challenged in the past and had redacted data, in the absence of a suitable policy saying otherwise. She expressed some (albeit tentative) concerns that perhaps everything

might be redacted if challenged. This should have been addressed by both the development of formal documentation describing a policy for this situation and internal awareness raising to communicate more clearly the legal status of the Data Centre records. It seemed unlikely that the Data Centre would be conforming to its contractual requirements if it acted unilaterally on this occasion, and therefore an expression of parent Archive policy in this area was probably quite adequate. In fact, in this case the approach went via the parent which oversaw negotiations prior to making a decision. Further ambiguity surrounded challenges to non-availability; for example, some materials were classified as too commercially sensitive to release but may be covered by Freedom of Information legislation. It appeared that there had been no retrospective assessment of previously closed datasets in light of FOI and this was something that could have been considered. Such requests remained primarily the responsibility of the parent National Archive, to whom the request would normally be issued. Following a representation by the relevant government department, the National Archive would make a decision as to whether content should be released.

Further legal complications arose as a result of some records within the Data Centre being exempt from corresponding public record legislation. Furthermore, on some occasions parts of certain accessions would be subject to intellectual property law. For example, software user manuals had been submitted in the past as part of a dataset's accompanying documentation, and this introduced some ambiguities that could have been addressed within a formal policy document.

Budgeting and Finance

In addition to a large research funder grant, the period of which ended prior to our audit, the greatest proportion of the **US State Digital Archive's** funding originated from centralised, state channels. Although the budget came via a host institution, there was little to no direct independent budgetary interaction. Instead, budget plans were subject to review by the board of eleven State-wide Directors (representing individual affiliate Universities), who would offer their approval, assuming sufficient finances were available. Plans would then pass to the centralised, State Board of Governors, whereby each University's Vice-President met and agreed before final ratification by the council of University Presidents. Budgetary flexibility was evident, and had been exploited in the past. For instance, when systems were transferred from a costly mainframe system to cheaper UNIX systems monies were freed up, enabling the Data Centre to acquire additional human resources. The base budget, allocated annually, continued automatically, although it was noted that in past periods of recession the Data Centre budget had been reduced. For instance, in 1991 when the host institution was asked to reduce 3 per cent of its spending the Data Centre was asked to cut that amount of their yearly budget, which was collected from the mid-year free balance. This accompanied a much wider commensurate series of public sector budget cuts within the state. Within the

model, the Data Centre was capable of maintaining a carry-over fund which was useful to meet costs that recur on a less than annual basis, such as replacement of expensive technological infrastructure equipment. These monies resided in a separate budget, which unlike the main Data Centre budget could be accessed by the host institution. Such tampering would be likely to elicit a strong negative reaction from the other state university libraries, and was therefore considered unlikely. One consequence of the protection afforded to the Data Centre budget (which further emphasised the value of the carry-over fund) was that no overspending could take place; budgetary separation meant that the University was unable to cover any deficits. At the time of the evaluation the Data Centre budget had never been in the negative; rather their annual spending had consistently yielded spare cash to carry over into the following financial year. There was also evidence of anticipatory budgeting for subsequent years, if not within the Archive itself, certainly within the wider Data Centre. For instance, when a significant proportion of the Data Centre's library infrastructure moved from dumb terminals to PCs they were able to project across a five year period the anticipated budgetary requirements and spending.

There was no charging model in place for the Archive services, but the library agreement countersigned by each affiliate included a caveat explaining that although no fee was currently payable, this may be introduced in the future. Discussions suggested that a quota-based system of billing might be adopted, with a view to both income generation and provoking a more thoughtful and selective approach to archiving from depositors. The administrative consequences of such a decision were thought likely to be considerable, and this again provided a clear justification for the appointment of a full time Archive Manager. There were many benefits associated with introducing a charging model, not least from a sustainability perspective. Perhaps the most profound was that the introduction of such a system would immediately reduce some of the concerns that surrounded the scalability of the Archive; continuing without charge was quite conceivable if the level of content remained roughly the same. However, an increasingly widespread use of the Archive services would introduce additional costs across the whole Archive budget. If these could be mitigated by a self-sustainable, charging-based system then the Archive could be less worried about attracting additional depositors. Income generation of this type provides a degree of insurance against possible future funding gaps, which as noted, could not be met by the local host University. The Data Centre Director suggested in discussion that the primary operational budget was guaranteed, but given previous funding dips in periods of recession, it was realistic to think that funding may be less than expected.

There was at the time of the evaluation no formal, distinct budget for the Archive, rather its allocations were consumed within an overall budget for the Data Centre. This was a situation that staff were seeking to amend however, and recent efforts had been made to develop a prototype budget for just the Archive, with the intention to make it increasingly indepen-

dent from its organisational context. A spreadsheet detailed individual costs associated with staffing, software and hardware, and the third party hardware hosting services provided by the host institution's Computing and Networking Services and a remote third party Data Centre. These costs amounted to around 550,000 USD of expenditure from a total Data Centre budget of just under 13m USD. The most significant Archive expenditure was staff salaries; given the strong suggestion above, it was likely that this would extend beyond the current 384,774 USD following the appointment of a full time manager.

Greater budgetary independence for the Archive appeared appropriate. In order to not only manage but also actively demonstrate the sustainability of the archival operations it was useful to isolate expenditure, incomes and assets (or proportions of each) that related to the Archive. This would in turn facilitate business planning, and the allocation of monies for contingency. The Archive could also consider maintaining a similar distinction with regard to the carry forward balance in order to ensure that cash saved in archival operations could be subsequently channelled back to cover those less frequent costs associated with archival preservation functionality. Preservation is unpredictable and flexible assets are therefore extremely valuable, particularly in the absence of a parent organisation capable of providing support in times of financial strain.

It was suggested that greater physical separation of the Archive might accompany a move towards greater financial independence. In contrast, this was not recommended, since the rich skills and other resources evident within the Archive team's operational context provided scope for intellectual and resourcing economies of scale that would benefit both the Archive and its associated services.

In terms of transparency, public law ensured that all organisations funded by the state legislature were bound to full disclosure of financial record keeping, ensuring that transparency was maintained, and that shortcomings in accounting practice could be immediately identified and corrected.

On occasions, costs associated with the accessioning of datasets led to the **National Archive's Data Centre's** resistance to archive. Data had been turned down in the past due to unnecessary expense associated with it - a prominent example was a dataset encoded in a proprietary format associated with an unnamed document management system. Mechanisms were offered by the system's developers to export these documents to PDF, but the Data Centre was unwilling to pay the charge. Had the parent contracting Archive insisted, it was acknowledged that the Data Centre would have had to proceed, such were the terms of the contract. It appeared though that the relationship that existed would make it unlikely for the parent to insist that transfer should take place where it might significantly undermine the financial position of the service and the Data Centre.

The service's turnover associated with digital preservation amounted to approximately 1m

GBP, about 20 per cent of its overall turnover. The Data Centre contract was negotiated at a fixed price, which was adequate to meet most costs, although in some circumstances where the service exceeded expectations there were additional costs that had to be met. History suggested that increased costs would be met with favourable terms in subsequent contract renegotiation. The original Data Centre contract was priced too low; the service was losing money and forced to rely on its own additional funding reserves. A subsequent contract compensated this loss and acknowledged the increased cost of providing the service.

Risk Management

At the **Cultural Heritage Archive**, rather than formalising risks in a risk-register or equivalent document, risks were explored and mitigated by planning for broad scenarios. It was assumed that any threatening technological consequences could be overcome by the technological expertise available in-house, and that although depletion or cessation of funding would inhibit the Archive's growth, it would not be terminal to the continuation of delivery services. Sustainability in the event of a combination of both funding lapses and technological barriers were less well addressed. The consequences of key personnel leaving the Archive were likely to be profound. It was suggested by staff that the Technical Director role could be assumed by another, and that the Principal Archivist role could be continued, given the momentum already established. However, there was a serious shortfall in documentation within the Archive, which could exacerbate the implications of staff loss. The technological systems were documented from an end user perspective but little documentation was available for prospective developers to inherit and understand the system to the extent that it could be confidently administered. Similarly, almost every aspect of archival policies and procedures (although seemingly well established among repository staff, and reflected at least partially in the system's imposed workflow) remained undocumented.

Widespread evidence of risk-based strategy was observed, although this seldom manifested itself as explicit, formal risk management. At the **UK Research Council Data Centre** it was revealed that the availability of content originating from one large data producer (which, it was suggested, the Data Centre relied upon for its very survival) was not guaranteed due to the non-renewal of an agreement between the producer and the Data Centre. An original archive and dissemination agreement ended in 1999 and its terms stated then that Data Centre should destroy all data upon cessation of agreement, which never happened. At the time of our audit, a new contract was in the midst of being negotiated.

The **US State Digital Archive** shared a lack of evidence of appropriate risk management activity. The lack of risk-based strategy resonated throughout much of the organisational, technological and digital object management infrastructure at the Archive.

Summary

Table 3.1 summarises the component elements necessary to be successful in these areas. These were derived by reference to both instances of good practice, and the more limited efforts observed elsewhere among the cohort. The more important question of what these goals mean, or rather how they may be practically accomplished, is explored in the following chapter.

Table 3.1: Organisational Infrastructure

Audit Issue	**Incorporated Goals**
Mission and Mandate	- Establish ratification of preservation mission from parent or governing entity
Organisational, Governance and Policy Best Practice	- Maintain business planning autonomy - Establish appropriate business planning - Establish appropriate coordination and steering platform - Evaluate and certify activities - Maintain best practice awareness
Suitability of Policies and Procedures	- Establish policy-review policy - Establish policy transparency
Succession and Service Continuity	- Establish appropriate strategies for facilitating succession of organisation or content - Establish relationships with succession partners
Repository Staffing	- Establish assurances of staff skills and capacity - Establish portfolio of internal or external staff training provisions - Establish appropriate categories of staff (roles and responsibilities) - Establish budget dedicated to training provision
Community Engagement	- Establish designated community - Maintain end user dialogue - Monitor and respond to designated community evolution
Contracts, Intellectual Property and Legal	- Ensure appropriate contractual management - Monitor and fulfil intellectual property responsibilities - Monitor and fulfil freedom of information responsibilities - Monitor and fulfil other legislative and legal responsibilities - Make explicit (and optionally transfer) preservation rights - Establish and maintain terms of deposit - Establish terms of use
Budgeting and Finances	- Establish appropriate financial accounting infrastructure - Establish assurances that all costs are and will continue to be covered - Establish budgetary protection assurances - Maintain budget carry-over facility - Maintain comprehensive costings breakdown - Establish appropriate contingency funding
Risk Management	- Maintain risk awareness

Digital Object Management

Scrutiny of the repository's digital object management provisions are focused on its core service of maintaining accessibility to and utility of its digital collections. Intrinsic functions are largely derived from the broadly accepted functional model presented in the Reference Model for an Open Archival Information System [ISO 14721, 2012]. They include functions associated with information ingest, physical data management and storage, preservation planning and dissemination.

Preservation Responsibility

The **US State Digital Archive** adopted an attitude of shared responsibility for preservation, between both the Archive and its content owners and depositors. It was perhaps for this reason that institutions were described not as passive 'depositors', but instead as 'affiliates', suggesting a degree of mutual cooperation. To this end, as documented in the Archive *Policy Guide*, affiliates were responsible for negotiating an agreement (counter-signed by representatives of both their institution and the Data Centre), incorporating details of authorised individuals for deposit, withdrawal and dissemination and details of projects and sub-accounts; selecting content for archiving and maintaining adequate local descriptive metadata; ensuring legal permissions were obtained and transferred to the Archive (assuming liability for breach of intellectual property rights occasioned by the deposit); submitting content to the Archive in the format specified in its Submission Information Package (SIP) specification; maintaining records of what was archived (including at minimum the *entity ID* of the SIP and links to locally stored metadata); verifying the success of the submission process via the generated error and ingest reports; requesting withdrawals where preservation was no longer required; and requesting dissemination when access to information was necessary.

Acquisition and Ingest

Robust systems for ingest were identified in several of our audited respositories. The **e-Depot's** workflow was quite typical. Prior to ingest, content tended to originate on installable CDs, in PDF format via the *File Transfer Protocol* or on locally received digital tapes. Content was installed along with the necessary helper applications on a reference workstation; the ingested content was a disk image snapshot of the reference machine. PDF documents (which represented the vast majority of received content) were validated via checksums and batched for processing. Both digital content and associated metadata were ingested, with bibliographic information standardised and a unique identifier (based on millisecond-level timestamps) associated with each object. Descriptive and structural metadata were provided by the publishers.

Every stage of dataset acquisition was recorded within the **National Archive's Data Centre's** *Accession Tracking System* (ATS). Its role was to document all events that related to in-

dividual accessions. These included communications that took place surrounding the dataset (whether internal or with data owners, government departmental records officers or client or contract managers at the parent Archive); suspensions on the accessioning process, for instance where the government departments' inactions result in the stalling of the process; and the final public release of the dataset. This provided provenance and traceability up to the point of the dataset's dissemination. The procedures for transfer were well documented and described procedures for initiating dataset transfer, appropriate communications that must be undertaken, physical transfer procedures, checks, documentation and receipts that must be issued in explicit detail with a suite of corresponding forms.

Formats, Naming and Identification

Various formats were supported for deposit by the range of audited organisations but some were more flexible than others. Strict requirements on file formats imposed by the **UK Research Council Data Centre** appeared to have the negative effect of dissuading depositors from submitting content to the Data Centre. A potential solution would have been to be more open minded about acceptable ingest formats, but employ people or acquire software capable of performing appropriate preservation transformations. The Data Centre appeared to be pushing the responsibility for encoding data in long term formats to the data producers, although this is unlikely to have been that group's primary motivation. A lack of funding was offered as an explanation, although on the other hand, opening up the range of ingest formats might have been a means to solicit greater funding. A corresponding funder policy permitted the Data Centre some discretion in deciding whether or not to accept deposited content. "The sole reason for keeping data", described the Data Centre *Operations Manual*, "is to distribute it for use". This did not presuppose contemporary use, instead acknowledging the fact that even long term curation was undertaken with a view to one day using the information that had been preserved. The two most influential considerations were the usability (format, conditions of use) and usefulness (quality, scale, coverage, gaps, uniqueness) of data.

The **National Archive's Data Centre** meanwhile supported a wide range of media and formats. In terms of physical format, the overwhelming majority of data arrived on CD or some kind of magnetic tape, with some low volume and non-confidential material also appearing by email.

Files ingested into the **UK Research Council Data Centre** shared a common naming convention including details of the instrument, location and date. Both capture instruments and locations were required to be registered with the the Data Centre to ensure their validity. Some ambiguity existed over the extent to which an instrument could change and still maintain the same identity. For instance, if components were replaced, was an instrument the same as it was before? Similarly, some confusion appeared to surround the location and time

information of certain data, such as those mounted on aircraft. Generally speaking, such ambiguities, as well as processing that had been undertaken on particular data, was documented within file-specific metadata. At the point of ingest, documents could be associated with datasets in order to provide format descriptions, details of problems or further information about particular instruments. This documentation was associated with data in one of various ways. The first was to include it within a *README* file located within the relevant data directory. An alternative approach was to create a file with the reference explicit within its title, which seemed to be a more robust means of enforcing the association. Finally, staff explained that some adopted file formats supported the addition of in-line comments.

At the **e-Depot** a notable shortcoming related to the adopted identifiers in use. Consisting of a simple UNIX timestamp generated at the moment of ingest this was potentially problematic if multiple ingest machines were commissioned to operate simultaneously or the procedure was streamlined to facilitate the ingest of objects at a rate faster than the timestamp's smallest unit of time. In the former case a solution would be to add a prefix to distinguish objects ingested by alternative machines. This would not address the latter concern however, and the repository's technical staff agreed that some kind of alternative means of conceiving identifiers would be preferable to mitigate potential future problems.

Likewise at the **US State Digital Archive** affiliates were not limited in terms of the filenames that they could allocate, which could result in unpredictable behaviour should multiple packages be submitted by a single affiliate with identical names.

At the **UK Research Council Data Centre** and **Cultural Heritage Archive** a philosophy of providing access dominated. Conversely, at the **US State Digital Archive**, which was principally a dark archive, curation and preservation were identified as primary responsibilities. Each AIP corresponded to a single intellectual entity (some examples might include a volume, dissertation or home movie). Some files (perhaps those originating from digital collections at affiliate institutions) would have a preservation level of none, and in such cases these files would be excluded from the archived package. The next stage corresponded to a development decision to limit the duplication of consistently referenced files; a global directory existed to accommodate any files that may be linked to by several archived objects. There was actually little evidence of savings in terms of bandwidth or processing - the remote, referenced file would still be checked in all cases in order to ensure that it remained unchanged from a previously retrieved globally stored example. The storage savings were likely to be negligible too - for several reasons (including potential legal issues, as discussed above) only linked schema and DTD files were retrieved, and since these were text and comparatively small in terms of file-size the benefits appeared minimal. Conversely, the associated risks were potentially serious. Relying upon a system of shared files meant that no archived package was independently complete, and if one was acquired in isolation from the rest of the Archive this may be problematic. Any value obtained from capturing remotely

referenced content was at best threatened. In order to maintain the link it was necessary to alter references to point to the global directory, not the remote resource, which could be argued to be in contrast to the archival goals. It was suggested that the Archive should retire the global directory approach in favour of independently complete archival packages, despite the additional resulting storage overhead.

The AIP descriptor was created, corresponding to the original SIP descriptor but with additional documentation of all files, relationships and events that the object had been subject to within the Archive. In comparison with the SIP, which was described thoroughly in the public SIP specification and associated METS SIP Profile, minimal documentation described the structure and content of AIPs. A short AIP definition existed within a system overview document, and the same document described the process within which SIPs were converted AIPs, but it was suggested that this should be extended, given that maintaining an understanding of the AIP was, in the longer term, a higher priority.

The **US State Digital Archive** processed incoming content with a 'prep' module, part of their technology platform. This ensured the validity of the submitted package, removing files that were not described within the corresponding manifest. When invalid or non-well formed SIP descriptors were identified packages were rejected and the process logged. Any files that existed within the submission package that were not documented in the associated package descriptor were rejected, although oddly this step was not formally documented.

Understandability Validation

Although not performed periodically, the **US State Digital Archive's** technology platform supported both MD5 and SHA1 message digest algorithms and was capable of recording both in association with a single object. Mechanisms and policies apparently existed for resolving a situation where a single archival package demonstrated corruption.

At the system level it was vital that the repository implemented a means for ongoing fixity checking, either conducted in a random or methodical fashion. Without maintaining assurances about information integrity until the point of dissemination there were implicit risks that even a well implemented backup strategy might fail to solve, if errors, accidents or malfeasance were noticed too late, with even backed-up content potentially affected.

Technical approaches to information integrity maintenance and verification should have been refined. Checksum provisions and other information integrity measurements were insufficient to create an audit trail for the data and processing in the Archive. Monitoring and checking schedules were well described but in practice rarely applied.

A validation mechanism at the **National Archive's Data Centre** sought to ensure that transformations were true and accurate, and sufficiently representative of the original source dataset. The identification of inaccuracies or deficiencies in the original data could also

be found at this stage and brought to the attention of researchers or data users. Two terms, 'transformation validation', and 'content validation' were internally coined by the Data Centre to describe the two types. Software was available to perform content validation on various types of data, validating against metadata descriptors to ensure that content within database fields corresponded to the documented schema, and if so, that metadata was retained with the preserved dataset. It was capable of checking for example, that columns that ought to be dates were dates, and those that ought to be integers were integers. The tool was also used to automate the creation of data description metadata based on the characteristics of the data where none previously existed. Measurement checks ensured that content corresponded to that described in transfer forms, accompanying documentation or in any other referenced publication. This typically compared averages, counts, or other quantitative characteristics of the dataset with this evidential information. Results of each of these checks were recorded within the dataset's processing record. Irrespective of how poor the results of these checks were, the Data Centre had a policy not to change data, even if errors were obvious and straightforwardly correctable. All such problems and inconsistencies were documented. Only one exception to this had been documented, and related to corruption prior to accession by the Data Centre that prevented data processing. The dataset in question had relied upon fixed-width fields to distinguish individual content fields and the corruption had misaligned the data, with significant effects, whereby some closed data (i.e. not for general viewing) had been shifted into open field positions. This was therefore repaired. It would have been useful to have a more formally expressed policy to document the circumstances within which such interventions would be permissible. Documentation described an example occasion where intervention was legitimate, but it was suggested that the Data Centre extrapolate this into a more generally applicable policy statement.

Immediately prior to committing a dataset to permanent preservation storage, responsible staff submitted it for review by a fellow Data Centre staff member. An individual was appointed with responsibility for data checking, although in the event of his/her non availability other staff might have been required to provide this final quality assurance input. Typically, the review comprised checks for consistency between the transformed dataset, the original source and associated documentation, completeness and accuracy. Documented procedures supported random checking by senior Data Centre staff, and mechanisms were in place to involve more than one individual in this final review process (up to the entire Data Centre team) for datasets identified as presenting particularly challenging problems or layers of complexity. The Data Centre *Checking procedure check-list* was completed immediately prior to the dataset being committed.

Media replacement took place in one of four circumstances. The first was the result of ongoing activities, with the latter three reactions to atypical circumstances that might arise.

- A tape has reached its maximum usage count (set at 10,000 mounts) or age (set at 7 years);
- A tape has been damaged due to hardware failure;
- Unanticipated readability problems;
- Other failures or the procurement of information that suggests a tape or batch might be suspect.

Automated checks within the system monitored usage and tape age - the current values were subject to review on an occasional basis, and systems staff were granted suitable discretion to retire tapes prior to them reaching the maximum age or usage if more convenient.

When faced with errors or discrepancies an administrator made a judgement as to whether it was the media upon which data resided that was to blame. Suspect media was disabled from interacting with the wider system until this judgement was made. When the media itself was found to be at fault a procedure existed for media replacement. In the event of above average media failures administrators were expected to pursue with manufacturers the possibility that a batch was affected with a common fault. Where hardware was identified as being at fault, staff would liaise with vendor engineers who would perform appropriate maintenance and corrections. Consultation would follow to determine whether any media on a failing hardware drive might have been affected.

Initially, administrators would determine from the system which files were stored on the tape that was set to be replaced. Each of these files was recalled to online storage, and verified to ensure their integrity had been maintained. These files were then copied to a new tape, and the system was instructed to disregard the previous tape. Media retirement was recorded within the media register, and the media was then erased and destroyed (presumably according to the guidelines expressed within the *Digital Preservation Procedures Manual*, although this was not made explicit).

The **Cultural Heritage Archive** boasted of no content loss throughout the full extent of its twenty-five year lifetime, despite a number of system migrations. This was regarded with some scepticism given a lack of documentation about exactly what was expected to be within the collection.

A notable shortcoming evident in the **US State Digital Archive's** self evaluation was that there was no process (documented or otherwise) for determining the understandability and usability of archived content. It was suggested that this could be implemented in the short term by exploiting the existing communication channels that existed between the Archive and its designated community. Without implementing a means for verifying ongoing understandability the Archive could not confidently claim to be preserving content (other than at

the bit-stream level). Given the finite breadth of its designated community it was thought to be quite feasible for the Archive to establish a straightforward method. It was likely that such increased interaction with affiliates would require an additional administrative commitment; this would represent a further justification for the appointment of a full time Archive Manager.

Preservation Policy and Service Levels

Preservation approaches varied across our audited organisations. Within the **UK Research Council Data Centre**, a distinction was drawn between three classes of data, known as A, B and C data, which corresponded roughly to the extent to which their preservation was prioritised. The characteristics of each was explained in simple terms by the Curation Manager. Class A data was that for which the Data Centre was the primary archive, and this amounted to approximately one third of all data holdings. Class B was that for which although the Data Centre was not the primary or sole custodian, scepticism existed about the ability of the primary archive to provide adequate preservation services. Class C data was that which was adequately preserved elsewhere, but was sufficiently useful to retain. Only class A and some class B data for which the Data Centre considers itself to be the primary steward was really relevant when considering issues of preservation. The fundamental differences between classes were not formally expressed anywhere at the time of this assessment - the internal classification was a realisation of an appraisal - but had tremendous influence over the preservation activities to which particular data sets will be subject.

Following ingest, data that arrived into the Data Centre's archival storage were likely to have already been subject to some processing. Class A, B and C data were distributed across several disks; a single directory contained symbolic links that corresponded to each dataset, and pointed to the physical space where individual data streams were located. All access, including end user access was via this directory. The server that this directory resided upon probably represented the most vulnerable point of the system. If compromised this could limit the extent to which data can be retrieved or its completeness ensured. By maintaining a single system for archive and delivery, the Centre was limited in terms of the extent to which system changes could be implemented while maintaining an optimal level of service.

The transition of a SIP into an AIP and subsequently a DIP was not regulated by a formal policy, nor documented anywhere other than the resulting directory structure on the archival storage media. The Data Centre only accepted data that it had appraised as suitable for deposit. Therefore, SIPs were always transformed into an AIP/DIP. The choice of structure for archival and dissemination packages was made by the ingest staff, who made their decision based on how the dataset was most likely to be used. Therefore, the primary criteria for converting SIPs into AIPs was ease of use. But no strict policy existed, and the decision to convert a particular dataset in one way or another, was not documented in a separate log,

audit trail or documentation. It was only visible from the consequent presentation of the data in the storage system.

No formal criteria were established to determine when preservation responsibility was accepted by the Data Centre, and neither was the transfer of responsibility acknowledged in a deposit agreement exchanged with the depositor. In practice, the Data Centre assumed preservation responsibility from the moment the data had been transferred (uploaded) to the Data Centre and an e-mail had been received from the depositor containing the script of the completed transfer. The Data Centre could have benefited from formalising this process, especially for the class A datasets. In terms of current practice, whereby the Data Centre generally utilised a single file format for SIP, AIP and DIP, the production of detailed depositor agreements was not regarded as being necessary. However, looking into the future when the Data Centre may have had to consider AIP or even SIP migration as part of preservation processing, it would have to be clear about the rights and responsibilities it had with respect to data.

Since a finite number of AIP 'types' were generally accepted (measurement data, model data, satellite data, Met Office data), the *Operations Manual* specified an AIP configuration for each. However, the AIP configurations were not documented in a sufficiently structured way to permit the automatic verification or validation of an archival package. The AIP definitions maintained by the Data Centre were largely sufficient to meet long term preservation requirements, although poorly documented. The choice of file formats for each class of data stored was more based on data usage criteria than on issues of long-term preservation. In fact though, they also happened to be suitable for preservation without requiring extensive processing at short intervals. Although the Data Centre *Operations Manual* described the process of constructing archival packages from submission packages, no documentation was created in practice that enabled one to verify whether the instructions had been followed. The division of a SIP into constituent AIPs (i.e. files in a directory structure different from that of the original SIP) was not tracked - no checks were performed to verify whether all files transferred to the Data Centre were in the stored AIP.

Perhaps more relevant with respect to this point was the appraisal of datasets for classification into class A, B and C datasets. The appraisal criteria were reasonable. However, the class or category assignment did not mean anything for the AIP configuration: irrespective of whether a dataset was classified as A, B or C, it was kept in the same file format and supplemented with the same kind of documentation. The only difference was in storage and back-up practices whereby class C data was supported by fewer safe copies (if any). Assuming that the AIP configuration was sufficient for preserving the class A data it may have been reasonable to apply it to class B and C data. However, in principle the content with highest preservation priority should have been accompanied by richer supplementary information, in greater quantity. Documentation as such was a weak point at the Data Centre, at least

from the archival point of view. All AIPs of class A should have had their entire custodial history logged, all processing decisions documented, all usage occasions tracked, and all changes to documentation audited. This was not done at the time of this assessment. An example was offered of the Hierarchical Data Format raw data that arrived from the HIRDLS instrument aboard the Earth Observing System (EOS) AURA Mission Spacecraft. Given the nature of the data, which was tied very closely to the spacecraft's instruments, it was vital that the semantics of the data were appropriately documented with sufficient representation and provenance information.

The **National Archive's Data Centre** favoured transformation as a primary preservation strategy. The point of transforming data within the Data Centre was to regularise its form to facilitate access and usability. Target formats were chosen based upon their amenability to subsequent conversion as part of ongoing preservation, and their ability to preserve the content and intellectual ordering of the original dataset. Datasets were organised and documented to facilitate the representation of their implicit information, and not necessarily to reflect their form when they arrived. Nevertheless, documentation made it possible to trace back to see the structure of content upon accession. All steps to transform the data were recorded, with the procedures demanding that staff were satisfied that not only could they repeat the process exactly with only the original data, their description of the process and the metadata that accompanies the dataset, but that a different staff member could do the same.

The initial stages of data transformation required staff to document the source dataset's structure, as well as the content and format of each field within. Following their initial assessment, data specialists were required to formally document their anticipated actions within an 'Approach' document, to be evaluated by other members of the team.

Where content arrived in popular formats such as *Microsoft Access* .mdb files documentation was fairly straightforward to automate. In other cases, where more proprietary or bespoke formats or data structures were employed it could be necessary to use more labour intensive techniques, such as text analysis of data documentation or alternatively manual keying. In even more complex cases, it was often necessary to reverse engineer software to retrieve a description of data structures. It was not clear whether the potential legal implications of such techniques had been formally explored, but it was suggested that this should be done and documented with some priority. The end user license agreements of software vary, but a policy statement encouraging staff to investigate their rights with respect to such procedures was considered appropriate for inclusion in the *Digital Preservation Procedures Manual*. If reverse engineering failed to yield a clear definition then raw data analysis was undertaken, using tools such as *od* or the Data Centre's own *flook*. Concerns surrounded such procedures which amounted to little more than (highly educated) guesswork, but given the shortcomings implicit within the received data, and the pressures placed upon the Data Centre to archive whatever they were given, this was probably unavoidable from time to time. It appeared

that communications with departments and data owners were suitably extensive to limit the likelihood of these circumstances in almost every case. Irrespective of which methods were employed for documentation, the required elements remained consistent. Characteristics that were documented for each accessioned dataset included:

- File layouts
- Record Structures
- Field formats (including field widths, repeat counts and relationships between fields (e.g., whether they were keys or indexes)
- Field descriptions - usually one line descriptions, explaining what a field was for, and also documenting any ambiguity that might surround the data specialists's interpretation

Data were also anonymised at this stage if there had been indications from the transferring department that this was necessary. The transferring department would indicate in their initial correspondence how this anonymisation should take place (summarisation of data or suppression of certain fields). An *Anonymisation Procedures Manual* contained detailed descriptions of the process that should be followed.

Following analysis, description and the completion of any required anonymisation, data specialists decided upon the form within which the dataset would be preserved. It was not necessary to maintain the original table structure, and it in some cases could be desirable to normalise if this had not already taken place. Any temporary, or redundant tables could be discarded, but their prior existence was recorded so that it was possible to maintain an understanding of the form of the data at the point of its transfer. Any proposed conversion or disposal was expressed within the 'Approach' document detailed above, a discussion group or in direct conversation with the Service Manager. These modifications were approved by TNA prior to their execution.

Of some concern were situations where coded values could not be transformed, or indeed translated, since information (such as lookup tables, or references linking data to existing lookup tables) had been omitted from the documentation supplied by data owners or departments. In such cases the Data Centre staff were required to simply deduce the meaning of these codes.

Within the **US State Digital Archive** the shape of preservation activities was based on the agreement between the Archive and a given contracting affiliate. This agreement could describe preservation expectations as one of full, bit or none. Preservation levels were determined at the level of individual files at the ingest stage, based on the account (the identity of

the particular affiliate, or the repository itself), the project code (enabling individual accounts to allocate alternative preservation levels for the same file format) and file format. Although sub accounts could also be defined within individual accounts, these were relevant only for billing and reporting purposes and were irrelevant from a strict preservation perspective. Full preservation meant that all applicable and available preservation techniques were employed, including migration, localization and normalisation. Bit preservation meant that files would be ingested and stored and subject to refreshing and integrity checks, but no further preservation methods would be employed. A preservation status of none was to accommodate content that arrived within a larger package of submitted content, which for some reason had not been isolated and removed prior to deposit.

Full preservation services were only practically applicable to a small subsection of all file formats. These had been identified based on a combination of their preservation viability and their popularity. For each format (the full range was listed on the Archive information web page) a background report was prepared detailing a selection of technological characteristics, and documenting additional associated sociological or legal issues (e.g. adoption rate, licensing implications). These were internally ratified by the Archive group as a whole to determine their completeness of coverage. No additional external registries (such as representation information registries) were automatically referenced although it was acknowledged that the research activity undertaken to understand each format could involve consultation with a variety of sources. Following the completion of an initial report a further document was conceived to detail the action plan that would be undertaken with respect to the corresponding format. This document represented the most critical aspect of preservation planning within the Archive. In some respects, the format-centric approach had limitations in terms of the effectiveness with which one can preserve content, or more specifically, the significant properties of individual items. The principal value of an item may relate to any one of its physical or semantic characteristics. There were implicit risks in adopting a preservation approach that dwelt on formats and not objects. It was acknowledged that until relatively recently the Archive had given very little consideration to the subject of significant properties at all. However an even more granular, affiliate-oriented approach should have been pursued; indeed, much of the overhead related to the identification of significant properties might have been allocated to affiliates as an additional responsibility. To date, no affiliate had explicitly notified the Archive of the properties that ought to be preserved within any deposited content but it was suggested that once a suitable infrastructure was conceived to accommodate such varying degrees of preservation, it should be encouraged.

Three main preservation approaches were implemented within the Archive, and reflecting the overall organisational philosophy these were applied exclusively at ingest. In the case of full preservation the archived AIP would contain both an original bit-stream or bit-streams (that is, the originally deposited file or files) as well as the last-best migrated preservable example

of that file or those files. In some cases the original and last-best preservable example would be synonymous. Normalization, Migration and Localization were all identified as means to manage format obsolescence, and based on format transformation. Normalization was intended to ensure that those files that were in formats that were less than optimal for preservation were created in a more preservation worthy format. For instance, PDF files would be normalised into a set of page-image TIFF files. Normalized files were not saved, rather the process itself was recorded as having been successful. Some question marks remained about the value of this process since the ongoing availability of a successful normalisation method relies upon the preservation of the corresponding tool or script. Migration was intended to alleviate the risk of obsolescence by creating a version of at-risk formats that was considered to be a reasonable successor to that format. This could be an equivalent but higher version example of the original format (e.g., PDF 1.4 files might be migrated to PDF 1.6) or a different format altogether. This then represented the 'last-best' preservation version, replacing any that might have existed within the AIP before. Localization, as discussed above, was intended to ensure that remotely referenced files were, wherever possible, harvested and stored locally to ensure the independent completeness of AIPs. For files subject to full preservation (as specified by affiliates in Appendix A of the Data Centre - Library Agreement), the appropriate preservation strategy was documented within each format-specific action plan. For those files formats that had no corresponding action plan the Archive would commit to bit-level preservation until a suitable preservation strategy was identified. At that time the affected files would be disseminated and reingested, and during this process the appropriate preservation steps implemented. Discussions revealed that decisions to research and conceive background and action plans for new formats were prompted by the nature of content that had been received within the Archive.

Since it took around three months to fully document a format and conceive an appropriate action plan it was suggested that the Archive should seek to modularise the system code to encourage the development of format plugins from beyond the Archive's in-house development group. By facilitating and motivating external development the work could be effectively shared and many more than the eighteen supported file formats (at the time of the evaluation) could be preserved. In addition, adoption of the Open Source software would be likely to increase and its status as a stable archiving solution increasingly consolidated. An action plan review schedule existed in order to identify when format information was approaching obsolescence, although as a result of intensive development commitments there had been evidence of failure to undertake some reviews in an appropriately timely fashion. A wider community of format specialists in a range of institutions would provide a considerably more effective, and ongoing means of policing to ensure that preservation planning remains both optimal and viable. Of considerable concern was the lack of regular integrity checking that was undertaken within the Archive, an issue that was described above. It was

hoped that the commitments made during discussions would be fulfilled, and an appropriate automated procedure conceived to execute fixity checking on a regular basis.

Preservation policy was limited within the **Cultural Heritage Archive**. Organisational uncertainty about the role of the Archive with respect to preservation was evident, and this ambiguity manifested itself in an approach to digital object management that fell short of that described in such best-practice benchmarks as OAIS for instance. Preservation planning was undertaken in an ad hoc fashion. Dealing with preservation issues and considering pitfalls and potential solutions was not explicitly part of anyones job, nor were there any (collective) reports written on this to inform decision making. Data management rarely extended beyond the association of simple web-page information with digital datasets. More sophisticated data management provisions were required for the Archive to consider itself to be OAIS compliant.

At the **UK Research Council Data Centre** a policy for ongoing appraisal was in place but rarely utilised. The Centre presented documentation describing a dataset review process but little evidence of implementation. Within the *Operations Manual* the procedure was described as a Retention Process. The process, as documented, was prompted by an automatic notification that data review was due, with a milestone in the project database conceived to correspond to each review. When this happened, a responsible individual was required to evaluate:

- the content of the data's corresponding catalogue entry (checking that links were working for example);
- the extent to which corresponding web pages were current, appropriate, informative and usable;
- the extent to which data was usable, accessible and adequately documented;
- whether any representation information (specifically software) was required to use the data;
- the effectiveness and security of corresponding ingest mechanisms; and
- the extent to which data was adequately documented, creating and aggregating documentation where appropriate.

Applying these criteria would result in evaluation marks between 1 ('Poor') and 5 ('Excellent') corresponding to both data usability and usefulness. Reviewers were then required to propose subsequent action, which could be to leave the dataset as is, keep the dataset but implement some changes or remove the dataset from the Archive. The latter seemed to

imply destruction, with no overt infrastructures in place to support transfer of stewardship to a more appropriate repository elsewhere. Despite the reasonably robust provisions for data review, discussions with the Curation Manager revealed that the process had seldom been undertaken. The number of datasets currently within the Centre, combined with the time consuming nature of the review process was the most critical factor - simply put, review notifications were arising more quickly than staff could undertake reviews. This was undoubtedly a problem, and threatened the viability of long term archiving within the Centre. The Data Centre's emphasis was very much on ingest, with dissemination enjoying a comparable, albeit secondary level of prioritisation.

In addition to continuing to preserve content the **US State Digital Archive** also supported withdrawal functionality to enable content to be removed. This would take place only upon the request of an authorised agent of the corresponding depositing affiliate. Although files belonging to a withdrawn AIP were deleted entirely from storage, the Archive maintained a record of the object's ingest and subsequent withdrawal, with the affiliate notified of the withdrawal via an emailed *Withdrawal Report*. A common use of withdrawal functionality was to correct a previously submitted package. In such cases withdrawal would be followed by a subsequent ingest of the package, with any errors amended. The Archive could unilaterally withdraw content if the preservation of specific material was subject to external legal challenge, in accordance with the policy described above.

The **US State Digital Archive** operated principally as a dark archive with no end user function. Despite this, there were examples of descriptive metadata maintained in association with archived content. The majority of descriptive metadata derived from SIP descriptors provided by affiliates. Information that would be captured when supplied by affiliates included a SIP package identifier (the only mandatory metadata), affiliate-assigned entity identifier, identifiers of external metadata records, title, serial volume and issue number. File names were also maintained for each file within the SIP. The Archive would add further internal identifiers associated with each individual AIP, file and bitstream. All of this metadata was stored within the Archive management database and within the corresponding AIP.

Limited formats were offered by the **National Archive's Data Centre**. Access to the Data Centre's archived content was almost exclusively via its website. Until the introduction of the Freedom of Information Act in 2005 the distinction between closed and open content was quite clear cut. Confusion followed the legislation's introduction within the Data Centre. For instance, the continued applicability of statutory bar (the mechanism that enables government departments to collect sensitive data with the proviso that it may not be used for purposes other than its original stated one) was unclear. The Data Centre's *Closed Data Access Procedures Manual* described the procedures, although it was of some concern that the most current version of this procedures document pre-dated the introduction of FOI legislation.

Content Dissemination

Within the **National Archive's Data Centre** system, DIPs were created dynamically, constructed from the corresponding database or flat CSV file to provide an interrogable, web accessible dataset. Catalogue data was available alongside datasets, encoded within HTML pages. Individuals very rarely requested their own copy of archived datasets, and instead the web interface was overwhelmingly the most popular means of accessing the Data Centre's content. Nevertheless, if tables were required to be delivered in an incomplete form (either due to legal restrictions or to conform with a specific sub-set request) this could be be done. A checksum was created at the point of the DIP's request, which was intended to ensure that it was both complete and correct with respect to the request issued. Similarly, the original SIP bitstream could be requested when it was of value. An example offered during discussions was of certain *Geographical Information Systems* (GIS) datasets that could not really be understood when converted from their source form and separated from the application that created them.

In technical terms, the system that stored the archived materials had no direct contact with the outside world. The Data Centre web server operated as a client to the archival servers, with limited, read only access. Therefore if the web server was compromised the extent to which a malicious individual could damage the archival storage component of the system was limited.

To facilitate the designated community's identification and discovery of content, the **UK Research Council Data Centre's** website offered just a narrative description of each dataset. Search and browse functionality complemented this metadata. There was however no evidence of the use of formal description metadata standards for resource discovery such as *ISADG*, or *EAD*. Nevertheless, the requirements of their designated community were probably met, with descriptive metadata available on the web sufficient to cater for the primary user group. A separate, but related, project called *Claddier* (funded by JISC) was developing further access methods, including better support for data citation. Metadata was both requested from the depositors and also created by Data Centre staff. The metadata to be included in the SIP was stated for depositors, but covered only a description of the data and its implicit variables. Relationships between metadata and archival packages were maintained by storing metadata within a separate directory adjacent to the data directories of the corresponding data sets. No separate techniques, persistent links or identifiers were employed to make this association more explicit. Since staff interactions with data sets might feasibly result in disassociation of this metadata-to-dataset relationship this might be considered as something of an implicit risk.

As submission, archival and dissemination packages within the Data Centre were generally synonymous, the need to demonstrate that the DIP (or AIP) creation process was complete

and correct was perhaps less pressing. Interviews revealed that the last time someone ordered a dataset to be delivered on transfer media was a long time ago. At the time of the assessment, DIPs were delivered exclusively online. Since the data was accessed via an online interface, DIP creation was virtually a one stage process, with archival packages simply delivered via web or FTP protocols. The onus of ensuring that packages corresponded to requests was placed upon the user. Subsets of web accessible data were not really supported as such.

Summary

Once more, the infrastructures observed above are summarised in Table 3.2 with broad check-list areas expanded to encapsulate the range of individual goals that contribute to their accomplishment.

Table 3.2: Digital Object Management

Audit Issue	**Incorporated Goals**
Preservation Responsibility	- Make explicit (and optionally transfer) preservation responsibility - Establish data ownership
Acquisition and Ingest	- Authenticate source of ingested packages - Define ingest package specification - Document software dependencies - Establish and exercise ingest policy - Establish and exercise selection policy - Initiate stakeholder dialogue - Maintain depositor dialogue - Physically acquire content - Process ingested content - Select and appraise ingested content
Formats, Naming and Identification	- Establish list of supported formats - Establish means to track data object through preservation workflow - Establish naming convention - Verify ingest package conformity with specification - Establish means for data identification - Adopt appropriate preservation formats - Monitor file format obsolescence - Maintain archival package referential integrity
Understandability Validation	- Establish criteria for data review - Establish means for data review
Preservation Policy and Service Levels	- Classify archival data - Establish archival package configuration(s) - Establish criteria for disposal - Establish means for data disposal - Establish levels of preservation - Establish relationship between ingest and archival packages - Establish transformation procedure from ingest to archival packages

Continued on next page

Table 3.2 – continued from previous page

Audit Issue	Incorporated Goals
	- Plan for preservation - Exercise preservation plans - Select preservation strategies
Metadata and Documentation	- Record appropriate metadata - Maintain link between data and metadata - Document archival data - Record and maintain descriptive metadata - Record and maintain representation information
Content Dissemination	- Establish conditions for access - Establish physical and logical provisions for providing access - Establish relationship between access and archival packages - Implement access controls - Implement categories of access - Manage formation of dissemination package - Monitor access behaviours - Monitor unauthorised access

Technologies, Technical Infrastructure and Security

Consideration of these issues is, like the evaluation of organisational factors, not necessary exclusive to data management and preservation. Nevertheless, although satisfying more generic information security requirements is indicative of technical capacity, the explicit focus is on the suitability of technical platforms to support preservation and access. This encapsulates technical sustainability, appropriate provisions to detect and mitigate against information change and appropriate tools to facilitate practical preservation interactions.

Software and Hardware Inventory

A technical questionnaire completed by respondents from the **e-Depot** revealed a significant organisational investment in IBM software, with the *Digital Information Archiving System* software at the heart of the repository, providing the breadth of its functionality. Although this relied on additional off-the-shelf IBM products (such as *Tivoli Access Manager* for authentication and authorisation and *Tivoli Storage Manager* for object management and backup) the system was designed and built specifically for the repository application according to the OAIS reference model. IBM was chosen as the supplier following a tender process on the basis of mainly functional requirements. Questionnaire responses suggested good practice with measures in place to optimise performance and capacity, mitigate risks to system security, and deal with any environmental unpredictability (UPS and climate control). Some concerns were raised with regards to several technical responses. No off-site backup

facilities were employed and although the Library had a disaster plan at the institutional level, there was nothing in place at the repository level, nor was there anything specifically addressing ICT concerns. This was thought to be of particular concern given the Library's low-lying nature, and its propensity for flooding. That backups were stored in facilities two floors below ground, in the same building in which the repository operated raised some concerns. Repository technical staff explained that moisture sensors were installed within these backup storage facilities.

The **US State Digital Archive's** technical infrastructure was part of an original software platform developed within the Data Centre. This operated within a Linux environment (the chosen distribution at the Archive was *Red Hat Enterprise* version 4). In addition to the core operating system the software relied upon Sun *Java* and the *MySQL* database server. Archival storage was managed by IBM's *Tivoli* software. Both the primary and a redundant secondary site featured dedicated machines for processing and storage, which were new and subject to appropriate renewal schedules. At the primary site all but some shared networking facilities were exclusively deployed for the Archive. Similarly, a tape robot at the redundant site was leased solely by the Archive.

All the **UK Research Council Data Centre's** systems aimed to use community-supported software and hardware, including open source systems where possible. There was a variety of bespoke code that was used during various data processing and validation stages but this was generally written in mainstream scripting languages. Some data that arrived within the Data Centre was encoded in closed formats and therefore during the ingest and accessioning stages it becomes necessary to rely on both proprietary software and hardware. Generally speaking, the choice of system infrastructure raised no substantial risks of itself being irreplaceable, irreparable or subject to unanticipated and unavoidable licensing changes that would prejudice continuity.

At the **US State Digital Archive** system updates were undertaken based on a needs and risk based assessment. Numerous security mailing lists were subscribed to in order to determine potential problems associated with software that may need to be patched. New and update packages were installed using the *Red Hat Package Manager* (RPM) and updates made available via Red Hat's *Update Agent*. Upgrading was tested within a controlled environment on legacy hardware that corresponded closely with the live configuration. This also demonstrated that the system operated adequately on even old hardware and offered assurances that its performance and functionality would be optimal on the production system. The Archive staff met with system administrators on a biweekly basis providing an opportunity to plan software and hardware maintenance and customisation to suit any emerging user needs.

Backup and Redundancy

The **e-Depot's** most critical shortcoming was the lack of off-site backup facilities, which was of particular concern due to the local topography. Repository staff assured auditors that this was currently being addressed, and had been highlighted in prior external investigations independently commissioned by the Library.

Three principle methods were available for data backup within the **UK Research Council Data Centre**. These were to a local tape archive stored within a Data Centre fire safe and via *rsync* to separate disk storage and to a bespoke petabyte datastore, maintained at the same facilities but around five hundred metres away. The UNIX *df* command revealed around 62 terabytes of content within the archive. Class A data were backed up to local tape on a mainly ad hoc basis, and to the petabyte store as part of a regular backup job. Smaller Class A data were also subject to daily backup via rsync. Class B data was subject to similar backup processes, although it was rarely if ever backed up to local tapes. Little documented backup policy surrounded Class C data; large data within this category were unlikely to be backed up at all - smaller datasets may be backed up to the petabyte store for convenience. The Data Centre had at least one recorded instance of data loss, when a large Class B satellite dataset was lost following a catastrophic filesystem failure, leading to the loss of everything stored on the RAID array. Reacquisition of the data was possible, albeit complicated, and its size was considerable; other data was consequently afforded higher priority, and this particular dataset was not backed up at Data Centre. It was unclear what explanation was provided to the depositor and users about the non-availability of this dataset. There were also daily dumps of *Ingres*, *MySQL* and *PostgreSQL* databases that supported the catalogue, website and various ancillary systems within the Data Centre. These were stored on tape within a firesafe, *rsynced* to another disk and backed up to the petabyte datastore. With respect to storage, at least two copies of class A datasets were maintained, with at least one copy of class B. The Data Centre should have therefore had at least 4 copies of each class A dataset. Documentation was included in the directory structure of an AIP, and was therefore also backed up. There was one issue identified with respect to the backing up of large datasets; the physical capacity of storage systems would sometimes limit the extent to which policies could be adhered to with larger files manually split between storage volumes and partitions. The petabyte system was a *Storatek* tape library (produced by Sun Microsystems). The Data Centre's own storage system relied upon *Cyberview* servers configured for *RAID 5*. Redundancy was therefore maintained on all disks clusters; however, although a single disk's failure could be tolerated, the failure of two or more within a single cluster would result in loss (albeit in many cases recoverable).

All of the copies of data were maintained in a common geographic area (distributed up to 500 metres). None of the redundant storage provisions could really be classified as off-site. This was of particular concern given the safety notification that one received upon entering the site - *a sounding bell denoted a firm alarm, whereas a klaxon would sound in*

the event of a nuclear incident! This did not appear on any of the Data Centre risk assessment documents, and although no doubt capable of destroying much of the local environment, little had apparently been done to assess the threat or to conceive of contingencies (most obviously to store redundant copies of data in a more remote location).

Tapes were checked randomly, but not systematically and there seemed to be little evidence of documentation of test results. Tapes and servers were decommissioned at regular intervals, and it appeared that resource availability was not a premium concern with respect to the replacement of faulty media or infrastructural hardware. However, although procedures for replacing a tape were in place, there were no formal mechanisms for identifying faults. Similarly, although a disaster recovery plan did exist as part of the Data Centre *Operations Manual*, it could have been more detailed and ought to have been tested in fire drill procedures and the test results documented.

System Recovery

Backups were performed regularly at the **US State Digital Archive**, with system software and the *MySQL* management database included within the procedure. Three archival copies of AIPs were maintained at all times. Although some Data Centre software was also accommodated at the San Diego Supercomputer Center the Archive's content was not currently included. It was agreed that relationships should be continuously pursued with more geographically diverse organisations in order to conceive and build reciprocal agreements to mutually accommodate content. Although no complete system recovery tests had been undertaken, Data Centre staff described a number of occasions where individual items had been recovered. It was noted that system administrator staff claimed to have undertaken simulations of data destruction and recovery from a redundant site, which were apparently successful, but this was not documented in a prominent place, and appeared to be an ad hoc test. Some aspects of this were covered in the Archive's (at the time unfinished) *Continuity of Operations Plan* (COOP), conceived to meet state legislature requirements. This described the steps to overcome problems associated with disaster, although omitted to describe the specific steps required to re-establish the service or the location of key documentation.

Fixity and Content Integrity

Fixity information that was collected and stored within the **e-Depot**, (specifically CRC32 checksums) raised some concerns. Since it was stored within the archival repository alongside additional technical metadata any system compromises that threatened the integrity of metadata or objects could in theory also prejudice the integrity of these checksums (which were principally deployed to determine when and where unauthorised changes have taken place).

At the point of ingest the **US State Digital Archive** system performed fixity checks to ensure that each master copy was identical. Since any AIP interactions were actioned by re-ingesting content these synchronous fixity checks would be undertaken at the point of dissemination or the execution of new preservation strategies. Staff offered a limited description of mechanisms and policies embedded in the system software code to resolve fixity inconsistencies, but as noted this was undocumented. There did not appear to be any explicit procedures or mechanisms to report bit loss or corruption to repository administration. The fact that no bit loss had yet been incurred was a weak justification for the absence of such mechanisms.

Storage media were refreshed annually, with a scheduled job within *Tivoli* to transfer all stored content to new tapes. In addition, *Tivoli* supported a range of functionality to determine tape deterioration or increased error probability. A healthily paranoid level of administration was consistently maintained, and any tapes that prompted concerns would no longer be written to, and content immediately transferred to an alternative fresh tape.

Information Security Best Practice

The **UK Research Council Data Centre** maintained fairly stringent physical access requirements around its main petabyte data store, where backup copies of the highest priority datasets were maintained. Similarly, archival/access storage facilities were subject to physical security systems. Pass card authentication was enforced within the store facility, and although visitors could be signed in, they were required to be accompanied by authorised individuals. More generally, the facility demonstrated security best practice. All employees were required to display identification badges at all times and visitors were required to liaise on arrival and departure with gatehouse security staff in order to be issued with a visitor's pass.

Exemplary physical security was observed at the **US State Digital Archive**. Electronic locks protected the central machine room and each of the core network fibre huts. All doors opening to public spaces were configured to fail to a secure state. Alarms were immediately investigated by local staff or referred to campus police. Key fobs and proximity cards were required for access, with rights granted based on work requirements and staff integration needs. This meant that most technical staff would have access to the areas in which the Archive machines were based. PIN codes were required in addition to physical fobs during non-working hours. A variety of environmental security measures were also implemented, with heat and water detection facilities subject to continuous monitoring. Uninterrupted power supply facilities provided power for all computer equipment in the event of grid failure, and a diesel generator offered a day's power for all systems before it needs to be refuelled. The only perceivable shortcoming from a physical perspective were the lack of hurricane proof windows within the central server room which were in any case due for

imminent installation.

The Archive's redundant site was a bespoke secure Data Centre, and therefore physically optimised to ensure security, accessibility and connectivity. Non-stop security monitoring, video surveillance, air temperature and humidity control and monitoring, and redundant cooling were all available. Lightning protection, smoke detection and fire suppression and emergency power were also provided.

A notable concern was the lack of geographical diversity between the two sites (both were within the same US state), and one might conceive of a disaster (natural or otherwise) that might render both sites non-operational. The Data Centre Director described an informal hurricane threat assessment exercise that suggested that the chances of a single hurricane affecting both sites was very low; however, two hurricanes might occur simultaneously or in quick succession. The biggest continuity issues were largely organisational and the Archive had already demonstrated a willingness to collaborate (with a much more physically remote institution) to address these.

At the **US State Digital Archive** passwords were rotated every one hundred and eighty days, and were strictly enforced to include numbers, letters, punctuation, upper and lower case characters. A single database user permitted insert and delete rights, although these were applied to all tables (although according to established work flows only the affiliate user information should have ever been changed by direct human interaction). These rights could therefore be restricted to limit insert and delete privileges more strictly. Processing scripts were executable by the five IT staff within the group, and configuration files were editable only by these individuals. Of more concern was the potential for human error during the processing of scripts. It was suggested that applications and user interfaces should be refined to maximise automation, limit manual interactions and render the system less vulnerable to accidental or malicious misuse.

System security at the **UK Research Council Data Centre** was less well enforced. Many of the software scripts utilised by the Data Centre staff were run as a shared user, which immediately let any staff member access the full range of functionality that all of the repository's collective scripts offer. If compromised and exploited a responsible individual would be very difficult, if not impossible, to accurately identify.

The **e-Depot** was praised for its commitment to evaluating the extent of its achievements, as well as the areas in which it might improve. The most recent completed assessment was undertaken by KPMG, which highlighted the lack of off-site backup facilities among its chief concerns.

More formal risk management was universally poorly undertaken. The **US State Digital Archive** lacked a formal risk register although some aspects of risk were covered in the Archive's *Incident Event Threat Matrix*. The **Cultural Heritage Archive** undertook no for-

mal, documented disaster planning and although the Archive's Technical Director suggested that the availability of backups guaranteed that the worst outcome could be the loss of a single day's work this was not systematically demonstrable.

Summary

Table 3.3 illustrates the incorporated goals within each high level sub-category of Technologies, Technical Infrastructure and Security.

Table 3.3: Technologies, Technical Infrastructure and Security

Audit Issue	Incorporated Goals
Software and Hardware Inventory	- Establish appropriate hardware infrastructure - Establish appropriate software infrastructure - Establish software upgrade policy - Establish hardware upgrade policy
Backup and Redundancy	- Backup documentation - Define policy and procedures for undertaking backups - Ensure synchronisation of data separated by time or space - Establish appropriate backup redundancy provisions - Establish appropriate backup remoteness provisions - Establish appropriate database (i.e system) backup infrastructure - Establish appropriate provisions for backup - Establish suitability of backup infrastructure through testing
Fixity and Content Integrity	- Maintain data integrity - Validate data integrity - Continuously validate data integrity - Establish media refreshment policy - Validate integrity of backups
System Recovery	- Define disaster recovery policy - Establish appropriate technical documentation base - Establish assurances of recoverability of any lost data - Limit data loss incidence
Information Security Best Practice	- Establish appropriate logical security provisions - Establish appropriate physical security provisions - Establish assurances of site stability - Establish assurances of availability of appropriate technical skills - Establish information security policy

Audit Conclusions By Organisation

Cultural Heritage Archive Assessed according to the strict terms outlined within our audit check-list the approach adopted by the Archive raised questions in a number of areas.

However, given the historical success of the Archive, and the esteem that it clearly enjoyed among its target communities it was difficult for us to dismiss its efforts according to just these metrics. The most notable shortcoming was the lack of formal documentation that characterised much of the business activities of the Archive. There can be little doubt that its current staff were competent in their positions, and that there was a shared sense of duty, responsibility and role. Similarly, the life cycle of content that was accessioned, archived and disseminated seemed well understood. Surrounding procedures were, although not formally communicated anywhere, well known. Without the discussions undertaken during this assessment there would have been little scope for forming any kind of organisational assessment. There was concern that a new staff member would face a similar struggle to understand the organisation's mechanisms, policies and scope without recourse to a resource within which they are formally, objectively and unambiguously expressed.

Associated with documentary shortcomings were issues concerning the Archive's policy in a number of key areas. Legal questions abounded, and there seemed to be few formal assurances that the Archive had legal authority to maintain much of its digital collections. Where agreements were in place they were generally informal or bore more similarity to 'understandings'. Relationships with organisations providing content should have been more formally established to provide the Archive with the necessary protection to enable it to continue its business.

A similar problem followed with respect to the user communities - closely related to internal documentation and transparency was the external issue of community trust. Based on its track record the Archive had established a dedicated user base, and although one could not dismiss the success with which this had been preserved for several years, there was a danger that without better external expression of their policies and procedures this might be threatened.

Preservation policy was perhaps even more of a widespread problem. Frequently described in this case study was an organisational uncertainty about the role of the Archive with respect to preservation, and this ambiguity manifested itself in an approach to digital object management that fell short of that described in best-practice benchmarks such as *OAIS*.

A further issue associated with sustainability that was of concern was the extent to which the staff (most notably the Principal Archivist) were inextricably associated with the Archive. Given the extent to which the latter's dedication, self-motivation and wide range of contacts had offered security and continuity to the Archive, there was a notable risk that her departure would be difficult to overcome. Partly this was an organisational concern - her position existed (and was therefore centrally funded) only for as long as she continued to occupy it. The other factor was less quantifiable, but nonetheless persuasive, and was based on her unique personality and knowledge. That the Archive had existed so successfully for over

twenty-five years with little evidence of service disruption or data loss appeared to relate, to at least some extent, to the fact that the Archive had enjoyed tremendous staff stability throughout its period of existence. Perhaps true sustainability could only be demonstrated, and this concern addressed, following a rotation of staff, whereby new individuals were expected to take over in key roles.

Overall our audit exercise identified a series of shortcomings that for the most part would probably manifest themselves only during a period of organisational disruption or change, or in the event of one or more unforeseen contingencies. However, it seemed that little was in place to mitigate such problems should they arise, and within the organisational model limited resource seemed available to do so. Many of the problems could be traced back to a lack of documentation and discussions highlighted concerns about the limited extent to which policies, procedures and legal relationships were formalised. Policies and work flows were clearly well-ingrained into the management and archival activities needed to be more communicable to stakeholders in order to elicit trust.

National Library e-Depot Our overall conclusion from this process was that the Library Data Centre was operating an efficient and considered repository service that emerged with credit from its exposure to the RLG-NARA audit check-list. The organisational, financial, technological and preservation infrastructures that the National Library Data Centre had established corresponded closely to those outlined within our audit check-list. In addition, a number of more broad characteristics of the Data Centre were applauded, and highlighted as being particularly representative of this overall success. Perhaps the most appealing aspect of the Data Centre was the proactive attitude that seemed to characterise the whole organisation. Eschewing the notion of continuing to plan until completely certain of success, the Library had achieved a great deal by deploying experimental solutions that had through practical experience been shaped into robust and stable systems providing a realistic and achievable example to other aspiring repositories. This was facilitated with a strong commitment of both effort and resources to an active research and development culture. Related to this, and also noteworthy in terms of the Library's successes was the fluid approach to financial management that supported all Library activities and which enabled the flexible allocation of funding where it was most needed, even between financial years and across organisational units.

As well as adopting an organisational structure that facilitated success, the National Library Data Centre was also praised for its commitment to evaluating the extent of its achievements, as well as the areas in which it might improve. A number of examples of external evaluation were identified as having taken place within the audit. The most recent completed assessment was undertaken by KPMG, which highlighted the lack of off-site backup facilities among its chief concerns. KPMG staff responsible for performing the audit were not particularly

expert in the area of digital preservation, and repository staff were themselves responsible for identifying and documenting a significant proportion of the points detailed in the final report. This appeared to confirm that demands for repository audit services were not being fully met by traditional providers. Plans for a comprehensive risk analysis investigation to be undertaken by Zurich Insurance (encompassing every aspect of the Library's operation, both technical and otherwise) were also discussed during our audit, again underlining an overall commitment to excellence.

Although broadly successful in its activities, a number of areas were identified throughout the course of our audit process that raised some further questions. The most critical shortcoming we identified within the repository was the lack of off-site backup facilities, although we were reassured that this was being addressed, and had been highlighted in prior external investigations.

Our second criticism was associated with the identifiers allocated to objects within the repository. Consisting of a simple UNIX timestamp generated at the moment of ingest this may have posed problems if multiple ingest machines were commissioned to operate simultaneously or the procedure was streamlined to facilitate the ingest of more than one object per second. In the former case a solution would have been to add a prefix to distinguish objects ingested by alternative machines. This would not have addressed the latter concern however, and the repository's technical staff agreed that some kind of alternative means of conceiving identifiers would be preferable to mitigate potential future problems.

A further concern was associated with the fixity information that was collected and stored within the repository, specifically CRC32 checksums. These were stored within the archival repository alongside additional technical metadata. System compromises that threatened the integrity of metadata or objects could in theory have also prejudiced the integrity of these checksums, which were principally deployed to determine when and where unauthorised changes had taken place.

The issue of software escrow was subjected to similar scrutiny during our audit, and also emerged as an area of some concern. The Data Centre's technical infrastructure was essentially a proprietary system, consisting of both off-the-shelf and bespoke software developed by IBM. No software escrow agreements were in place The likelihood of vendor collapse was perceived as being extremely minimal, and irrespective of this, since the data and system software were separable, such an eventuality could be survived until appropriate alternative software became available. Additionally, it was argued that since a number of large global banks used and relied upon the same IBM software, there was sufficient motivation for the vendor to continue to support it. This reveals the importance of adaptable or flexible criteria. Even where a course of action has identifiable risks, the most important factor in determining a successful repository is its willingness and capability to undertake the appro-

priate risk/benefit assessment exercises. Self evaluation has an important role even in formal, externally orchestrated audit exercises.

Related to software escrow concerns were fears that without formal succession plans the Library was exposing its content to risk. Once again these were to an extent mitigated; the Library staff argued that its legally defined mandate and obligation rendered such plans unnecessary, ensuring the Library's permanent existence. Notwithstanding this, some doubts continued to persist, particularly associated with those collections not subject to these legal considerations, such as international, non Dutch materials.

A final issue of concern identified was an example of system bottlenecking that was being experienced within the system during the visit, preventing the ingest of objects. This issue concerned a small script responsible for the allocation of identifiers. Since a system restart, this script was no longer operational and consequently no objects could be added to the system until the problem was resolved.

National Archive's Data Centre Disregarding its considerable quantity, the quality of documentation that was available to describe and inform almost every process and contingency associated with the National Archive's Data Centre was impressive. Discussions with staff members within the organisation revealed that most of the shortcomings had already been identified and earmarked for corrective action; a culture of ongoing improvement and a commitment to excellence appeared to exist at the Data Centre and this was reflected in many areas.

One area of concern was the lack of a true mission statement for the Data Centre. Similarly, a more cohesive and well expressed definition of its designated community would have been useful, if only to further legitimise the numerous policies and procedures that had been formally documented. Also in organisational terms some ambiguities existed within the legal context that surrounded the Data Centre. The closed data procedures had not been updated to reflect Freedom of Information legislation and greater consideration should have been given to the data that accompanied (and was necessary for the interpretation of) core materials, but was licensed under different terms.

With respect to digital object management, more explicit and granular planning for specific formats would also have been welcomed. It was anticipated that binary data formats would become more commonplace within government data sets and the approach to dealing with these appeared ad hoc in places. Related to the shortcomings defining a designated community, it was recommended that the Data Centre explore in more detail the understandability requirements of its users, and their anticipated requirements over time, in more expansive terms than just accessibility. Some questions also surrounded those materials that were maintained in their original bit-stream format to support contemporary usability (such as the GIS

datasets described during discussions). Whether these were really being preserved more than simply retained is questionable.

Technologically, and in terms of security the Archive seemed to be adequately supported, although it was argued that ongoing revisions of security documents should have been maintained, and controls to maintain logical security made more explicit within documentation.

Overall though, the Data Centre represented a useful benchmark for contemporary repositories, at least in terms of the criteria expressed within the RLG-NARA check list.

UK Research Council Data Centre The pilot assessment of the Data Centre revealed that policies ought to have been more rigorously conceived, documented and formalised, and circulated widely among repository staff to create a widespread understanding of exactly what the repository was engaged in, and how it was ultimately operating. In terms of staffing, greater resources should have been invested in staff skills development, most notably archival skills which were lacking in comparison with scientific expertise.

It was concluded that technical approaches to information integrity maintenance and verification could be considerably refined. Then-current checksum provisions and other information integrity measurements were largely insufficient to create an audit trail for the data and processing in the Archive. Monitoring and checking schedules were well described but in practice rarely applied. Likewise, data management rarely extended beyond the association of simple web-page type information with digital datasets. More sophisticated data management provisions were required for the Archive to consider itself to be OAIS compliant.

Preservation planning at the Data Centre was undertaken in an extremely ad hoc fashion. Dealing with preservation issues and considering pitfalls and potential solutions were not explicitly parts of anyone's job, nor was there any reporting or internal information sharing to influence or guide decision-making. Related to this, stringent requirements on file formats dissuaded depositors from submitting content. A potential solution would have been to be more open minded about acceptable ingest formats, but employ people or acquire software capable of performing appropriate SIP to AIP to DIP conversion. A lack of funding was an obvious a barrier here, although opening up the range of ingest formats might have been a means to solicit greater funding. Fundamentally, curation activities seemed to take something of a backseat to facilitating access. To be trustworthy in terms of this assessment the organisation would have to embrace its preservation responsibilities, and to identify issues associated with mandate, legal status, services and functions that needed to adapt in service of that.

US State Digital Archive The Archive provided an invaluable service to their statewide affiliates and their efforts were broadly successful. The infrastructure that had been

established, the financial support that had been secured and the firm mandate upon which the Archive was founded were all robust. As well as developing an infrastructure that corresponded favourably with much of the central work in this area, the Archive staff demonstrated a keen willingness to determine the success of their efforts, and had already been quick to identify their weaknesses.

Many suggestions were offered in the course of this audit, but the most important were typically those that the Archive had already identified themselves. The first was the appointment of additional staff, with the most high priority being a manager for the Archive, who could engage with affiliates and plan and direct future administrative and operational direction. The Archive's primary goal within the near future was to increase affiliate numbers and the quantity and quality of content within it and to enhance its reputation.

There was a clearly identifiable need to engage with other organisations to arrange secure storage that was sufficiently robust. Building relationships would have enabled succession or escrow arrangements to be established, further remote storage of backed up materials, and ultimately, assuming the emergence of DAITSS as a widely adopted tool, collaboration in systems development and format description.

At the system level it was vital that the repository implemented a means for ongoing fixity checking, either conducted in a random or methodical fashion. Without maintaining assurances about information integrity until the point of dissemination there were implicit risks that even a well implemented backup strategy might fail to solve, if errors, accidents or malfeasance were noticed too late, and even backed up content demonstrates the emergent problem.

Another key point that emerged was that many of the problems being addressed in the Archive's operations were dealt with on a somewhat ad hoc basis. There was little central coordination of risk or challenges, or of the operational means to overcome them. By composing its own catalogue of risks the Archive could better equip itself to manage resource effectively to meet all of the challenges, at the points where the greatest threats were being faced.

On the whole though, the Archive demonstrated its status as an effective and well managed organisation. Its efforts stood up well to even considerable scrutiny according to the criteria within the RLG-NARA check-list, and also to those within comparable efforts such as the German *nestor* project's criteria catalogue.

3.4 Gaps and Desiderata

The range of materials collected during each of our audits provides invaluable perspectives for those wishing to implement or expand their own digital preservation environments, and for those who would seek to evaluate their performance. Accounts such as these represent an evidence base of practice aligned to the various audit standards - an insight into what it means to satisfy - and to fall short of meeting - various criteria that are collectively considered vital for repositories that would seek to be trustworthy.

We could not find comprehensive examples of best practice in any single institution - differing approaches had equivalent validity and value depending on the operational context. We did find considerable evidence that understanding was lacking in a wide range of areas, perhaps most pertinently in terms of actively preserving data. Each of the audited institutions had custodial responsibility for their respective data collections. This was mandated via a variety of means, including contracts, legislation and terms of funding agreements but was generally explicit. Nevertheless, not all of the organisations appeared to prioritise the long term preservation aspects of their remit. Instead, collection, or more commonly, the provision of access, were cited as primary motivations and the target of most time and resource investment. Given the high profile nature of several of these organisations the fact that none got every aspect of their operations right is revealing. Only the US State Digital Archive, which operated as a dark archive (i.e. with no significant access function) was most focused on preserving its collections. This is understandable; it is easier to justify investment for contemporary than future uses. That is not to say that access and preservation are competing or non-complementary goals but they are not equivalent. Assurances of sustainability are not essential for the provision of access services, but are integral to a preservation effort. Our studies revealed that many organisations are unclear about how such assurances can be characterised. More widespread understanding of what it practically means to preserve will encourage more organisations to ensure their infrastructures measure up to the demands. It is reasonable to expect that many end users will tend to assume that if an organisation is capable of serving a particular dataset today, that it will be able to do so in ten years time. Functionality is a more marketable concept than longevity and end users are more likely to assume the latter exists where there are few metrics on which to base their assessment.

Our sample of audited institutions comprised mostly an elite range of robustly funded organisations. National and state wide libraries and archives and disciplinary data centres may reasonably be expected to be expert in long term preservation. Their success was evident throughout our assessments; the process showed the diverse range of ways that success could be realised, but also that not everyone gets everything right. The account from the Cultural Heritage Archive is illustrative of something else; the challenges faced by those who aspire to create, collect or aggregate data in a local, comparatively non-expert setting. As we de-

scribed in earlier sections, preservation is no longer an issue faced solely by large, traditional, memory organisations and responsibility is being increasingly delegated to the institution, research project or individual. Funded by the Arts and Humanities Research Council (which at the time operated a companion data preservation service called the *Arts and Humanities Data Service*), the Cultural Heritage Archive opted to exempt itself from depositing data in the AHDS because of ambiguities surrounding the legal status of their collection and their approach to collecting. To do so required them to commit to preserving their own data, and to maintaining infrastructures and processes capable of supporting preservation. At the time a single sheet form was sufficient to present a satisfactory justification and evidence of their capacity to do so. In reality, the system in place was revealed by our research to be little more than an effective system for content collection and dissemination. Organisational, technological and information-level sustainability were not demonstrated - there was substantial evidence to suggest that none had really been systematically accounted for. If illustrative of the type of experience that led the AHRC to discontinue the AHDS funding then it was evidence of a gap in overall data management and preservation provision.

Today, data producers or collectors have less discretion whether to deposit or not - often there is no appropriate custodial organisation that will accept and preserve content. If anything, the onus is increasingly on data producing institutions to offer their own assurances of data curation and of widespread, continued availability and accessibility. The best practice lessons of expert organisations - the evidence of what they do can inform the novice research units and individuals who have inherited custodial responsibility. To ensure a fit to these often dramatically different contexts (a national library shares little in common with a University research group) the lessons must be flexible to transposition. A new resource is required: something less than a rigid, absolute set of rules or guidance, but more than the higher level expressions of best practice that can be found in international standards and other such literature that remains inaccessible and incomprehensible to most.

These are this thesis' desiderata - an accessible evidence base for preservation best practice that can be interrogated and used as a development tool for any organisation that has custodial responsibility for digital information, and a means of classifying preservation in terms of its fundamental components. At the very heart of this is the idea of *evidence*. How can organisations be *demonstrably* successful at undertaking preservation? The problems with existing approaches for preservation planning and validation lie not with methodological shortcomings, but with difficulties in identifying and understanding vulnerabilities and the associated consequences. Several knowledge facets that are not currently addressable contribute to risk and inform or justify preservation decisions.

3.4.1 The Relationship with Risk

The repositories we surveyed often struggled to relate their collections and preservation environment to associated preservation risks and opportunities. On occasion, practitioners were oblivious to potential risks or resolution strategies even though they may be considered commonplace elsewhere, cases that Donald Rumsfeldt may describe as "unknown unknowns". Instead of planning preservation based on specific information or representational properties, we saw repositories electing to base their strategies on self evident characteristics, such as file format. Preservation classification is a means of uniting those with things in common, to enable information sharing, and the perception of common risks. Preservation environments that appeared to vary markedly in terms of infrastructure, domain or scale actually demonstrated a great deal in common, but making explicit such relationships was suffocated by a lack of suitably expressive tools. It is possible to build a picture of preservation function and risk that is broadly applicable, while still being practically meaningful.

Other problems relate to the breadth of the preservation challenge. Technical, legal, financial, management and content skills are all required to successfully ensure information availability. It is rare for any individual to have comprehensive oversight across a preservation environment, and therefore understanding the whole can be complicated. Preservation planning is particularly constrained by perceptions, expectations and priorities of individuals. It requires the input of diverse constituencies, each of which may have a differing priorities. Case studies reveal that in any single preservation context stakeholders might vary between curators, art historians, computer scientists, preservation specialists and management [Becker et al., 2007]; each role can be reasonably associated with a myriad of policy and procedural responsibilities. Given often opaque and sometimes conflicting priorities, we observe situations where approaches are selected to suit the interests of an individual, an organizational unit or a single functional component. The consequences of such choices are rarely so isolated as their justification: understanding wider impacts and risk outcomes ensures more informed preservation planning and more easily interpreted validation results.

Risks occur as a manifestation of those factors that threaten the accomplishment of goals, prejudice the availability or quality of required resources and expose those performing preservation functions to liabilities or negative perceptions. Better means for risk definition and understanding will prompt greater awareness of influential factors, their consequences and appropriate responses. Work in the related domain of internet security has revealed ontologies' value in the identification and classification of attacks and threats to networked systems, in terms of their relationships with technology, policy and use [Ahmed et al., 2007, Ekelhart et al., 2007, Fenz and Neubauer, 2009, Raskin et al., 2001, Tsoumas and Gritzalis, 2006]. A more holistic view of risk and its cause and effects seems well suited to the complex environments within which information is preserved, and its availability threatened.

Work has also been done to explore conceptual modeling of preservation goals with risk in mind [Dappert and Farquhar, 2009, Dappert, 2011], others have considered the role of risk management in designing preservation solutions [Barateiro et al., 2010, Barateiro et al., 2012]. Other efforts have sought to form relationships between heterogeneous metadata. The *P2 registry* [Tarrant et al., 2009] uses the semantic web to link data from Pronom3 to support rudimentary risk assessment based on file format characteristics. Ontologies were used in the *PANIC* project's prototypical preservation alert and response system [Hunter and Choudhury, 2004, Hunter and Choudhury, 2005]. More general documentation projects such as *PREMIS OWL* [Coppens et al., 2010], and the *CIDOC CRM* [Doerr, 2003] enable documentation to capture underlying semantics concealed beneath domain-dependent documentation structures. Digital library models [Candela et al., 2008, Kovács and Micsik, 2005] describe the digital library environment in terms of classes, subclasses and implicit relationships. As part of the work of the *CASPAR* project a *Core Ontology for Dependencies* facilitates documentation of information dependencies, both semantic and structural, and *PreScan* [Marketakis et al., 2009] supports automatic extraction of metadata and its encoding in RDF.

Our assumption, which informed our development of *DRAMBORA*, is that preservation can be understood as a complex interrelationship of several factors [McHugh et al., 2007]. Preservation goals are motivated or legitimized by rights and responsibilities, qualified by parameters which in turn direct activities that both rely upon and enhance or develop resources. Risk has a fluid relationship with each of these elements. The development of coherent relationships between real world practical examples of elements and specific, calculable risks enables a rich evaluation of risk causation and recovery - we can begin to see things that increase or decrease a risks likelihood or impact, and trace these to more distant dependencies. A core associated use case for this information is a risk identification tool - preservation practitioners are expected to identify familiar practical circumstances in the network of elements and by traversing the relationships identify unknown risk exposures, or see where risks that they are already aware of can be managed through the introduction of appropriate policy, process or resource.

The success of preservation planning and validation depends on robustly defined object properties and dependencies, explicitly stated environmental and contextual characteristics, and a systematic appreciation of associated risk exposure. Implicit within conforming systems is the capacity to trace these factors as a related matrix. A further application is the traversal of a network of related risks, in order to determine the factors exacerbating each, and represent more clearly the wider implications of a particular circumstance. Linear means for recording such information, as currently exists (like a conventional organizational risk register) have limited expressiveness.

Our preferred solution, presented in the following chapter, is to present risks and preservation efforts alongside one another, and make explicit links of causality and mitigation. The first

requirement is to represent preservation activities and infrastructures according to core goals and associated efforts.

Chapter 4

The Preserved Object and Repository Risks Ontology

4.1 Theory and Components

We approach the concept of risk as an intuitive expression of the probability and potential impact of negative outcomes. In many organisational contexts these may be characterised according to financial loss but in those organisations performing information preservation functions may be more widespread. They may include outcomes such as information loss, loss of organisational sustainability, reputational harm or legal liabilities. Risks are considered in terms of their impact on valued outcomes and conceptualised as the inverse to organisational objectives. The effectiveness of any approach can be understood in terms of the extent to which it enables the avoidance or mitigation of barriers to the accomplishment of given stated objectives. In that sense, risk management can be considered synonymous with digital preservation - itself an active process of identifying, measuring and responding to threats that may manifest on technical, cultural or social layers of activity. Risk management is a tool to quantify and systematically address whichever threats arise. Conversely, the most effective strategies, most prized resources and most enabling policies will be those that most effectively limit risk exposure. As stated earlier within this thesis, risk management requires a consciousness of the value of digital materials, an awareness of the implications of employing particular preservation strategies, and an understanding of one's own priorities and tolerances (risk appetite). We pursue a means of systematically identifying tenets of excellent practice within digital preservation, characterise these as risk management techniques, and align with corresponding risks, which can then be isolated and quantified.

A pivotal question in the conception of this thesis is why risk is a desirable basis for evaluating preservation readiness when there are other approaches that might be considered compelling alternatives. Risk offers a number of appealing advantages. We have covered at

some length the limitations of a compliance, or standards-based approach to preservation evaluation. Standards like *ISO 16363* are valuable contributions to a common best practice knowledge base but are limited. These limits are principally manifested in terms of the extent to which their criteria are intuitively relatable and applicable when referenced in isolation as part of a process of institutional evaluation. In contrast a risk based approach is widely applicable but can reflect specific circumstances and priorities within a given institutional setting. There are advantages in referencing available compliance standards, but as a sole measure of success, these can be uncomfortably ill-fitting.

A risk based approach can be aligned with an established set of community norms just as to very specific institutional objectives. In that spirit, our approach is firmly focused on the process of evaluation, whatever that means to a given organisation. We make no assumptions about whether any given metric is more or less objectively stated.

But why not adopt a market-led approach whereby those preservation environments that prosper within a competitive marketplace will be determined as most successful? This strategy suffers as a consequence of the very nature of digital preservation. Often shortcomings in preservation approach are evident only after time as passed - cause and negative effects are not always synchronous or even directly contiguous in time. A risk based strategy offers the opportunity to distil the challenges of preservation into more granular (and related) terms. That means that one can identify and remedy shortcomings without having to rely on a market reaction that might be months or even years away. A commercial preservation market appears to demand audit and certification systems as a precursor to its existence, so it is difficult to envisage a scenario within which it can rely on traditional market indicators to infer success. Given that many of the institutions within which preservation takes place do not operate on the basis of commercial models the approach appears even more limited in its usefulness.

Other alternatives to a risk-managed approach might include analytics-driven methods, stakeholder satisfaction surveys and institutional peer review. In fact, each has a legitimate part to play in our risk-based approach but in isolation are insufficient. Data analytics findings are limited by the quality of one's queries and these are driven by those operational goals and interactions that comprise our risk-based ontology. The views of stakeholders are, like market forces, of great potential interest but likely to be at best incomplete and at worst myopic and detached from reality. Peer review is unquestionably valuable, but is expensive and depends on a network of willing peers that (even in non-commercial contexts) may be difficult to arrange. With the presentation of a shared and extensible knowledge base we hope to offer some of the benefits of peer review.

Our primary goal is the distillation of the many facets of preservation into a flexible model that can be comprehended and referenced by both humans and machines. The *Preserved*

Object and Repository Risks Ontology (PORRO), our novel representation of institutional and object-centric characteristics defines related elements as risk cause or effect factors. An ontology represents knowledge within a domain in a hierarchical form. We choose it as our adopted structure because it offers sufficient expressiveness to show properties and connectivity between concepts that collectively comprise a given domain. Unlike for example a simple criteria list or hierarchy the ontology format offers flexibility of relationships, including the possibility of recursiveness to characterise the causality of activity and approach within an operational context. Unlike a bespoke relational model it provides a form that can be reused and shared. Combined with instances, the ontology comprises an interrogable and extensible knowledge base for how to do digital preservation. Its value as a form of information representation lies in its wide readability, its modularity and its potential connectivity to representations of other domains. *PORRO* presents a multidimensional picture of preservation best practice that can scale to accommodate emergent wisdom, and can be pitched at a macro or micro level to ensure its accessibility to a range of users and applications. It comprises both a fixed vocabularly and a scalable collection of practical manifestations of its concepts. It is the outcome of efforts to highlight meaningful information interrelationships within the preservation context. In subsequent sections we describe the ontology and its development, presenting insights into its initial population and drawing conclusions about its current and future applications. We prove its expressiveness to illustrate and convey diverse preservation infrastructures and its associated capacity to support wider aspects of preservation planning and evaluation. Both functions are enhanced by an integral alignment between discrete elements of preservation practice and an extensible knowledge base of examples from real world environments. These illustrate what it practically means to pursue goals, perform interactions, develop policies, respond to rights or obligations and establish appropriate resources. The model can be approached in one way as a glossary of terms defined by examples. Risk management capacity is our metric for preservation readiness and we have developed a set of fundamental preservation risks and linked them with those facets of preservation infrastructure that cause, exacerbate or mitigate them.

Digital preservation goals have been distilled into a number of criteria catalogues and international standards. The principle digital preservation standard throughout much of the community's development has been the *Reference Model for an Open Archival Information System* [ISO 14721, 2012] (OAIS). This lends two key models which have characterised much of the research activity in digital preservation. The most frequently cited (albeit argued by Giaretta to be the least interesting [APA, 2011] owing mainly to the simplicity with which many seek to interpret it and demonstrate their conformity) is its functional model. A high level overview of this breaks the challenge of digital preservation into a number of fundamental functional chunks. The standard's glossary and functional overview describes each functional entity as follows, each reflecting part of the repository-centric information

lifecycle.

An *Ingest* function contains the services that accept *Submission Information Packages* from *Producers*, prepares *Archival Information Packages* for *Archival Storage*, and ensures that Archival Information Packages and their supporting *Descriptive Information* become established within the system. A *Data Management* function contains the services for populating, maintaining, and accessing a wide variety of information. Some examples of this information are catalogs and inventories on what may be retrieved from Archival Storage, processing algorithms that may be run on retrieved data, consumer access statistics and billing, event based orders, security controls, and OAIS schedules, policies, and procedures. Archival Storage contains the services and functions used for the storage and retrieval of Archival Information Packages. The *Preservation Planning* function provides the services for monitoring the environment of the OAIS and which provides recommendations and preservation plans to ensure that the information stored in the OAIS remains accessible to, and understandable by, and sufficiently usable by, the *Designated Community* over the long term, even if the original computing environment becomes obsolete. *Administration* describes the services and functions needed to control the operation of the other OAIS functional entities on a day-to-day basis. Finally, the *Access* function contains the services which make the archival information holdings and related services visible to *Consumers*.

OAIS has informed and continues to inform the breadth of goals associated with information preservation. Criticisms of OAIS as a closed model (with no explicit entry or exist points) are countered with a subsequent standard describing producer-archive relationships [ISO 20652, 2006]. We reflected and re-presented the OAIS model in a more practical setting within the *Trustworthy Repository Audit and Certification Criteria and Checklist* [CRL/RLG, 2007], now an ISO standard [ISO 16363, 2012]. Both resources provided an intellectual basis for example preservation goals that we characterised within *DRAMBORA*. These were classified in *DRAMBORA* according to ten "functional classes" which reflected earlier work [McHugh et al., 2008].

4.2 Development Methodology

In developing an ontology representative of preservation best practice we favoured a bottom-up approach, focused on the identification of real world goals, in order to complement the more prescriptive selection that comprise the above-noted standards. Information from our audits described in the previous chapter provided a broad but very detailed perspective of objectives, strengths and weaknesses, in very practical terms. We sought to reflect real world priorities, anticipating that this would lead to the definition of a more intuitive information model. We believe a top-down model in isolation to be uncomfortably limiting, although we

acknowledge that a solely bottom-up model is also problematic.

We determined that an effective compromise was to focus primarily on the outcomes of our audit exercises, but rather than constrain our conclusions according to existing criteria, shape a new model. The subjects' priorities and emphases informed our development of the resource. We undertook a series of content analyses of audit reports and the full body of additional evidence collected throughout each assessment scrutinising, parsing and coding the evidence. Individual information elements were isolated and recorded in an Excel spreadsheet initially. Initially utilising the high level conceptual areas outlined in *TRAC* (see the tables in the previous chapter) and *OAIS* we recorded instances of activity that would contribute, or be detrimental to each. Gradually we exploded individual concepts to establish a set of goals, supported by evidence of real world attempts to pursue their completion.

The structure of the emerging resource evolved as information was collated and iteratively classified, conceived as a taxonomy of repository properties corresponding broadly to a combination of established community concepts. Given the role of *TRAC* in gathering much of this evidence, we naturally encountered residual information alignments between our new record and the *TRAC* structure. This was a quite deliberate methodological outcome. Our intention was not to dismiss or disregard existing certification guidelines. Not least due to their acknowledgement by ISO and establishment as international standards, we consider that any resource that purports to support improved repository effectiveness can do so more effectively by reflecting an existing regulatory ecosystem. Our intention was to reconfigure and enrich such guidance into a form that offered greater accessibility and utility.

Information excerpts were structured according to a corresponding repository property. These were then subdivided into issues associated with data ingest, data management and preservation, data access, organisational issues, policy issues and infrastructural issues. Higher level, macro-classifications reflect the *TRAC* and *OAIS* origins in terms of terminology, but do not share the same granularity of these sources. The simplification reflected the structure of the audits which in turn were designed to reflect the most common distribution of responsibilities evident within the preservation contexts.

We present an example of encoded information fragments in Table 4.1. One hundred and twenty seven categories of information (preservation 'goals') were recorded, with corresponding information facets aligned from each of the assessments that we conducted.

Once an initial set of categories was established we referred to a selection of other repository assessments to align their findings with this taxonomy. Notable examples include assessments of the LOCKSS archiving system [Dale et al., 2007], Portico [Waltz et al., 2010] and the Inter-University Consortium for Political and Social Research [Dale et al., 2006] undertaken by the US Center for Research Libraries [CRL, 2012a]. Accounts from these were deconstructed and encoded to also reflect the adopted classification. Further refer-

Table 4.1: Information Fragment Development Example

Property	TRAC	Archive A	Archive B	Archive C	Archive D	Archive E
Data Access - Access Conditions	B6.4 Repository has documented and implemented access policies (authorization rules, authentication requirements) consistent with deposit agreements for stored objects.	All users can access most content as required by legislation	No formal deposit agreements but access based on pre-ingest negotiations with depositors, and on the basis of policies relating to specific NERC programmes.	* Content free for personal and academic usage * Terms of access published on site	Onsite unless one of few publishers with external access accounts	Policies well established despite dark archive status * Content disseminated by FTP when requested by authorised agent of depositing affiliate
Data Access - Access Control	B6.3 Repository ensures that agreements applicable to access conditions are adhered to. B6.5 Repository access management system fully implements access policy. The repository must demonstrate that all access policies are implemented. B6.9 Repository demonstrates that all access requests result in a response of acceptance or rejection.	* User registration service, initially a precursor to accessing content but since FOI only required to access additional website features * Cron task opens closed datasets when their closure period expires (or when FOI exemptions no longer apply)	* UNIX user:group based security at the directory level * Privileges manually changed when requests for new access granted	* Digital watermarks and SPIFF technology to encode copyright holders names to images * User management system in proprietary XDB system * Logging of anonymous access	* IP authentication * Tivoli Access Manager	* FTP security * Manual script execution and authorisation checks introduce scope for human error * Maintenance of authorised affiliate details is crude, databases updated manually and scope for error is evident

ence was made to a selection of ISO standards on topics including information security [ISO 27001, 2005] and quality assurance [ISO 9000, 2005], introducing a more generic set of applicable information mappings.

Lastly, we mapped to *TRAC*'s criteria. This provided a broad functional classification that would provide the basis for the ontology, which takes a firmly goal-oriented view. Emerging mappings between ontology goals and TRAC criteria were in many cases not 1:1. This is illustrative of the differing priorities between the ontology and the criteria lists, and in some cases represents evident omissions from more formal criteria. Finally, the categories were refined further, combined in some cases and rewritten as explicit goals. This process reduced their number to 104 ontology goals compared to 84 *TRAC* criteria (several of these are mapped to multiple ontology goals due to their compound nature).

The following tables illustrate the range of adopted information classifications and the corresponding *TRAC* criteria. From these, a set of preservation goals was derived, again classified according to the same high level schema. Parameterisations, implementations, dependencies and drivers for each goal (extracted from the accompanying information excerpts) comprise the main body of the organisational aspects of the ontology. They are discussed in more detail in a subsequent section.

Table 4.2: Data Ingest Goals Mapped to TRAC

Goals	**Corresponding TRAC Criteria**
Authenticate source of ingested packages	B1.3, B6.10
Define ingest package specification	A5.3, B1.1, B1.2, B5.1
Document software dependencies	C1.1
Establish and exercise ingest policy	A5.3, B2.4
Establish and exercise selection policy	B2.4, A5.3
Establish and maintain terms of deposit	B1.7, A5.3
Establish list of supported formats	A5.3
Establish means to track data object through preservation workflow/lifecycle	B6.8, B5.4, B6.9, B4.5, C1.9, B6.10, B3.4, B4.2, B4.3
Establish naming convention	B2.5
Initiate stakeholder dialogue	B1.6, A5.3, A3.5
Maintain data integrity	A3.8, B4.4, B2.11
Maintain depositor dialogue	B1.6, B6.10, A5.3, A3.5, B6.3
Maintain link between data and metadata	B5.4, B5.2, B5.3
Physically acquire content	B1.5
Process ingested content	B1.8
Record appropriate metadata	B2.9, B5.1, B2.13
Select and appraise ingested content	B2.4
Validate data integrity	A3.8, B4.4, B2.11
Verify ingest package conformity with specification	B1.4, B6.10

Data ingest goals are those concerned principally with the selection, acquisition, negotiation and initial processing of data.

Data preservation goals are those concerned with physically accommodating the data as

Table 4.3: Data Preservation Goals Mapped to TRAC

Goals	**Corresponding TRAC Criteria**
Adopt appropriate preservation formats	TRAC B2.9, TRAC B3.1, TRAC B4.2, TRAC B3.3, TRAC B4.1
Classify archival data	TRAC B2.1, TRAC B1.1
Continuously validate data integrity	TRAC B2.11, TRAC A3.8, TRAC B4.4
Document archival data	TRAC B4.5, TRAC B5.1, TRAC B2.13, TRAC B2.1
Establish archival packages configuration(s)	TRAC B5.1, TRAC B2.2, TRAC B2.1, TRAC A2.1
Establish criteria for data identification	TRAC B6.10, TRAC B5.1, TRAC B2.6, TRAC B5.4
Establish criteria for data review	TRAC B3.4, TRAC B2.10
Establish criteria for disposal	TRAC B2.4
Establish data ownership	TRAC A5.4
Establish designated community	TRAC A3.1, TRAC B2.10
Establish levels of preservation	TRAC B1.1, TRAC B4.1, TRAC B3.1, TRAC B3.3
Establish logical storage provisions	TRAC B4.2
Establish means for data disposal	TRAC B2.4
Establish means for data identification	TRAC B6.10, TRAC B2.6, TRAC B5.4, TRAC B5.1
Establish means for data review	TRAC B2.10, TRAC B3.4
Establish relationship between ingest and archival packages	TRAC B2.1, TRAC B2.11, TRAC B6.10, TRAC B6.8, TRAC B4.3, TRAC B2.6, TRAC B2.3, TRAC B5.1
Establish transformation procedure from ingest to archival packages	TRAC B2.1
Evaluate and certify activities	TRAC C3.1, TRAC C1.9, TRAC B6.8, TRAC B3.4, TRAC A3.9
Exercise preservation plans	TRAC B4.1, TRAC B3.4, TRAC B3.1
Maintain archival package referential integrity	TRAC B5.2, TRAC B2.11, TRAC B5.3, TRAC B4.3, TRAC B2.6, TRAC A3.8, TRAC B5.4, TRAC B4.4
Maintain best practice awareness	TRAC B4.2, TRAC C3.1, TRAC B3.2, TRAC C2.2
Maintain end user dialogue	TRAC B6.10, TRAC A5.3, TRAC A3.5, TRAC C2.2, TRAC B6.2, TRAC B6.3, TRAC B2.10, TRAC B6.1
Make explicit (and optionally transfer) preservation responsibility	TRAC B1.7, TRAC C1.8, TRAC B1.1
Make explicit (and optionally transfer) preservation rights	TRAC A5.2, TRAC A3.3, TRAC C1.8, TRAC A5.4
Monitor and fulfil freedom of information responsibilities	TRAC C1.8
Monitor and fulfil IPR responsibilities	TRAC B6.3, TRAC C1.8, TRAC B6.4, TRAC A5.4, TRAC A5.5
Monitor and fulfil other legislative and legal responsibilities	TRAC C1.8, TRAC B6.4, TRAC B6.3
Monitor and respond to designated community evolution	TRAC B3.2, TRAC A3.4, TRAC C2.2, TRAC B2.10
Monitor file format obsolescence	TRAC B3.2
Plan for preservation	TRAC B3.1, TRAC B1.1, TRAC B2.9, TRAC B4.1, TRAC B3.4, TRAC B4.2, TRAC B3.3
Record and maintain descriptive metadata	TRAC B5.2, TRAC B5.1
Record and maintain representation information	TRAC B2.8, TRAC B2.9, TRAC B2.7
Select preservation strategies	TRAC B3.3, TRAC B3.1, TRAC B1.1, TRAC B4.3, TRAC B4.1, TRAC B4.2

well as the development and implementation of active preservation strategies and information integrity validation.

Table 4.4: Data Access Goals Mapped to TRAC

Goals	Corresponding TRAC Criteria
Establish conditions for access	TRAC B6.4, TRAC B6.3
Establish physical and logical provisions for access	TRAC B6.5, TRAC B6.1, TRAC C3.2, TRAC B6.4
Establish relationship between access and archival packages	TRAC B2.1, TRAC B6.10, TRAC B4.3, TRAC B5.1, TRAC B6.8
Establish terms of use	TRAC A5.4, TRAC B6.3, TRAC A5.3
Implement access controls	TRAC B6.9, TRAC B6.3, TRAC B6.5, TRAC B6.4
Implement categories of access	TRAC B6.3, TRAC B6.1, TRAC B6.4
Manage formation of dissemination package	TRAC B6.3, TRAC B6.7, TRAC B6.8, TRAC B5.1
Monitor access behaviours	TRAC B6.9, TRAC C2.2, TRAC B6.6, TRAC B2.10, TRAC B6.2, TRAC B5.1, TRAC A3.5
Monitor unauthorised access	TRAC B6.6, TRAC B6.2

Data access goals correspond to the provision of discoverable, usable content and also encompass issues of authentication and authorisation.

Organisational issues cover a varied selection of topics including staffing, legal issues, mandate and finance.

Technology issues are those concerned with the software and hardware platform within the repository, but extend beyond simply digital (computer) technology to include physical plant and infrastructure as well as the range of logical and security provisions available to support the service.

In some respects policy aspects are prevalent across each of the previous categories but we elected to include a dedicated categorisation to reflect their importance. This is to some extent illustrative of the difference between good practice being done and being seen to be done. Throughout the repository assessments we identified that policy was typically written in formal documentation but can also be evident in software code, systems or organisational structures.

Following the definition of a core set of preservation goals we began to extrapolate from these information elements a relational structure - we modelled entities within the *Semantic Mediawiki* online software and linked with information elements that had been isolated in the previous work. *Semantic Mediawiki* is an extended version of the popular *Mediawiki* software that lends additional semantic richness, *SPARQL*-like query support and form-building functionality to articles and content. It is a very accessible and intuitive tool for developing semantic links between concepts that have their origins in full text content. Our earlier work developing *DRAMBORA* provided a vocabulary for defining preservation components as one of “mandate”, “constraint”, “objective”, and “activity/asset” but comparison with the range of information elements originating from our audit exercises exposed these as inadequate.

Table 4.5: Organisational Infrastructure Goals Mapped to TRAC

Goals	**Corresponding TRAC Criteria**
Ensure appropriate contractual management	TRAC A5.1, TRAC B6.3, TRAC B6.4
Establish appropriate business planning	TRAC A4.1, TRAC A4.2
Establish appropriate categories of staff (roles and responsibilities)	TRAC A2.1, TRAC C3.3
Establish appropriate contingency funding	TRAC A4.5
Establish appropriate coordination and steering platform	TRAC C3.1
Establish appropriate financial accounting infrastructure	TRAC A4.3
Establish appropriate strategies for facilitating succession of organisation or content	TRAC A1.2
Establish assurances of sufficiency of staff skills and capacity	TRAC A2.1, TRAC A2.2
Establish assurances that all costs are and will continue to be covered	TRAC A4.5
Establish budget dedicated to training provision	TRAC A4.5, TRAC A2.3
Establish budgetary protection assurances	TRAC A4.5
Establish portfolio of internal or external staff training provisions	TRAC A2.3
Establish ratification of preservation mission from parent or governing entity	TRAC A1.1
Establish relationships with succession partners	TRAC A1.2
Maintain budget carry-over facility	TRAC A4.5
Maintain business planning autonomy	TRAC A4.2, TRAC A4.1
Maintain comprehensive costings breakdown	TRAC A4.5
Maintain risk awareness	TRAC C1.10, TRAC C3.1, TRAC A4.4

Table 4.6: Physical and Technological Infrastructure Goals Mapped to TRAC

Goals	**Corresponding TRAC Criteria**
Backup documentation	TRAC C1.2
Define disaster recovery policy	TRAC C3.4
Define policy and procedures for undertaking backups	TRAC C3.4
Ensure synchronisation of data separated by time or space	TRAC C1.4
Establish appropriate backup redundancy provisions	TRAC C1.3
Establish appropriate backup remoteness provisions	TRAC C3.4
Establish appropriate database backup infrastructure	TRAC C1.2
Establish appropriate hardware infrastructure	TRAC C1.7
Establish appropriate logical security provisions	TRAC B6.4, TRAC B6.5, TRAC C3.2
Establish appropriate physical security provisions	TRAC B6.4, TRAC B6.5, TRAC C3.2
Establish appropriate provisions for backup	TRAC C1.2
Establish appropriate software infrastructure	TRAC C2.2, TRAC C1.1, TRAC B2.7, TRAC C1.10
Establish appropriate technical documentation base	TRAC C1.9, TRAC A3.2, TRAC C3.3, TRAC C3.1, TRAC C1.7, TRAC C1.8
Establish assurances of availability of appropriate technical skills	TRAC C3.3, TRAC B2.7
Establish assurances of recoverability of any lost data	TRAC C3.4
Establish assurances of site stability	TRAC C3.4, TRAC C3.1
Establish hardware upgrade policy	TRAC A3.6, TRAC C1.9
Establish information security policy	TRAC C3.3, TRAC C3.1, TRAC C3.4
Establish media refreshment policy	TRAC C1.7
Establish software upgrade policy	TRAC A3.6, TRAC C1.10, TRAC C1.9
Establish suitability of backup infrastructure through testing	TRAC C1.2, TRAC C3.4
Limit data loss incidence	TRAC C1.6, TRAC C3.4
Validate integrity of backups	TRAC A3.8, TRAC C1.5

Table 4.7: Policy Framework Goals Mapped to TRAC

Goals	**Corresponding TRAC Criteria**
Define disaster recovery policy	TRAC C3.4
Define ingest package specification	TRAC B5.1, TRAC B1.1, TRAC B1.2, TRAC A5.3
Define policy and procedures for undertaking back-ups	TRAC C3.4
Establish and exercise ingest policy	TRAC B2.4, TRAC A5.3
Establish and exercise selection policy	TRAC B2.4, TRAC A5.3
Establish and maintain terms of deposit	TRAC A5.3, TRAC B1.7
Establish archival packages configuration(s)	TRAC B2.1, TRAC B5.1, TRAC B2.2, TRAC A2.1
Establish conditions for access	TRAC B6.3, TRAC B6.4
Establish criteria for data identification	TRAC B6.10, TRAC B5.1, TRAC B2.6, TRAC B5.4
Establish criteria for data review	TRAC B3.4, TRAC B2.10
Establish criteria for disposal	TRAC B2.4
Establish hardware upgrade policy	TRAC A3.6, TRAC C1.9
Establish information security policy	TRAC C3.1, TRAC C3.3, TRAC C3.4
Establish levels of preservation	TRAC B3.3, TRAC B1.1, TRAC B4.1, TRAC B3.1
Establish list of supported formats	TRAC A5.3
Establish logical storage provisions	TRAC B4.2
Establish media refreshment policy	TRAC C1.7
Establish physical and logical provisions for access	TRAC C3.2, TRAC B6.4, TRAC B6.1, TRAC B6.5
Establish policy review policy	TRAC B3.3, TRAC A4.2, TRAC C1.9, TRAC A3.2, TRAC A3.4, TRAC C1.8, TRAC A3.6
Establish policy transparency	TRAC A3.2, TRAC B2.12, TRAC A5.3, TRAC A3.7
Establish relationship between access and archival packages	TRAC B6.10, TRAC B5.1, TRAC B6.8, TRAC B4.3, TRAC B2.1
Establish relationship between ingest and archival packages	TRAC B2.3, TRAC B6.8, TRAC B6.10, TRAC B2.1, TRAC B5.1, TRAC B2.11, TRAC B2.6, TRAC B4.3
Establish software upgrade policy	TRAC C1.10, TRAC A3.6, TRAC C1.9
Establish terms of use	TRAC A5.4, TRAC A5.3, TRAC B6.3
Establish transformation procedure from ingest to archival packages	TRAC B2.1

Feedback from *DRAMBORA* training activities [Mchugh, 2009] suggested that it was not intuitive for end users to consider their systems in these terms. Grouping activities and assets maintained inflexible associations which limited opportunities for reuse.

In response to such concerns we elected to deconstruct these classifications, establishing groupings of preservation "actions", "resources", "policies" and "mandates", related in the first instance by common associated preservation objectives. Our earlier content analysis had yielded structured evidence examples that conformed with one or more of these categories. We could take the practical responses to preservation problems that were evident in real world environments and characterise them in terms of their status. Goals (often prompted by specific mandate or other such compulsion) were pursued with the enactment of policies, in turn implemented by actions which were supported by specific resources. These were analogous to the example evidence entry accompanying each *TRAC* criterion. The collective set of related elements is indicative of much more than just associated documentation or resources though. Instead its purpose is to illustrate why particular objectives are necessary or worthwhile, the wide variety of ways in which they can be approached, how in practical terms they can be accomplished, and any dependencies for doing so.

This informed the object properties which were initially established within the *Semantic MediaWiki*. Sharing the common "objective" domain these were "Is achieved by" (action), "Is supported by" (resource), "Is defined by" (policy) and "Is legitimised by (mandate). We conceived a further object property "Is validated by", coupling the preservation goal with a corresponding *TRAC* criterion (intended to offer some formal, regulatory legitimacy to the corresponding objective). This is an example of how we envisage the interface between top down and bottom up approaches.

Using *PHP/MySQL* we developed a bespoke ontology manager web application which facilitated the ontology's definition and iteration. Firstly we generalised those information elements considered too specific for widespread applicability, resulting in a two tier hierarchy of resources, activities, rights and responsibilities (formerly mandates) and parameters (formerly policy). These would be reflected in a distinction between leaf node elements and corresponding individuals. In ontology terms they were characterised as classes (the generic fundamental parts of the preservation process) and instances, or individuals which are more specific representations and more subjective, based on context or time. The individuals comprise the knowledge base but the ontology is structured principally at the level of entity. This ensures the discoverability of more specific example practice via terminology with more general meaning. It also introduces an abstraction between general best practice (intended to be timeless and non context-sensitive) and specific implementations. Relationships between classes were developed based on evidence from the audits and mappings to example supporting evidence from *TRAC*. Our online application yielded some 8899 individual relationships between the individual ontology classes. Our mapping to TRAC, and to the *Data Seal of*

Approval (see Chapter 5) is available from:

`mchughontology.hatii.arts.gla.ac.uk/ontologybrowser/viewTrac.php`

We defined, modeled and iteratively evaluated relationships, reflecting and illustrating systematic, functional relationships within the given example preservation contexts, as well as risk causality relationships, highlighting where elements were threatened by particular risks, and where they were likely to influence risk probability and/or impact. We based a great deal of this on evidence available from the *DRAMBORA*'s online tool (which we developed and offer freely to information professionals wishing to perform systematic preservation risk assessment) and the tracing of corresponding relationships between identified risks and risk causation / recovery factors and the ontology elements.

A simple *AJAX* web application for traversing the ontology is available from:

`mchughontology.hatii.arts.gla.ac.uk/ontologybrowser/`

A further application that exposes the rdf elements and relationships is at:

`mchughontology.hatii.arts.gla.ac.uk/ontologybrowser/viewRelatable.php`

Our ontology editor, which was used to populate the ontology and make explicit the relationships between individual elements, is available at:

`mchughontology.hatii.arts.gla.ac.uk/ontologybrowser/structureBuild.php`

and

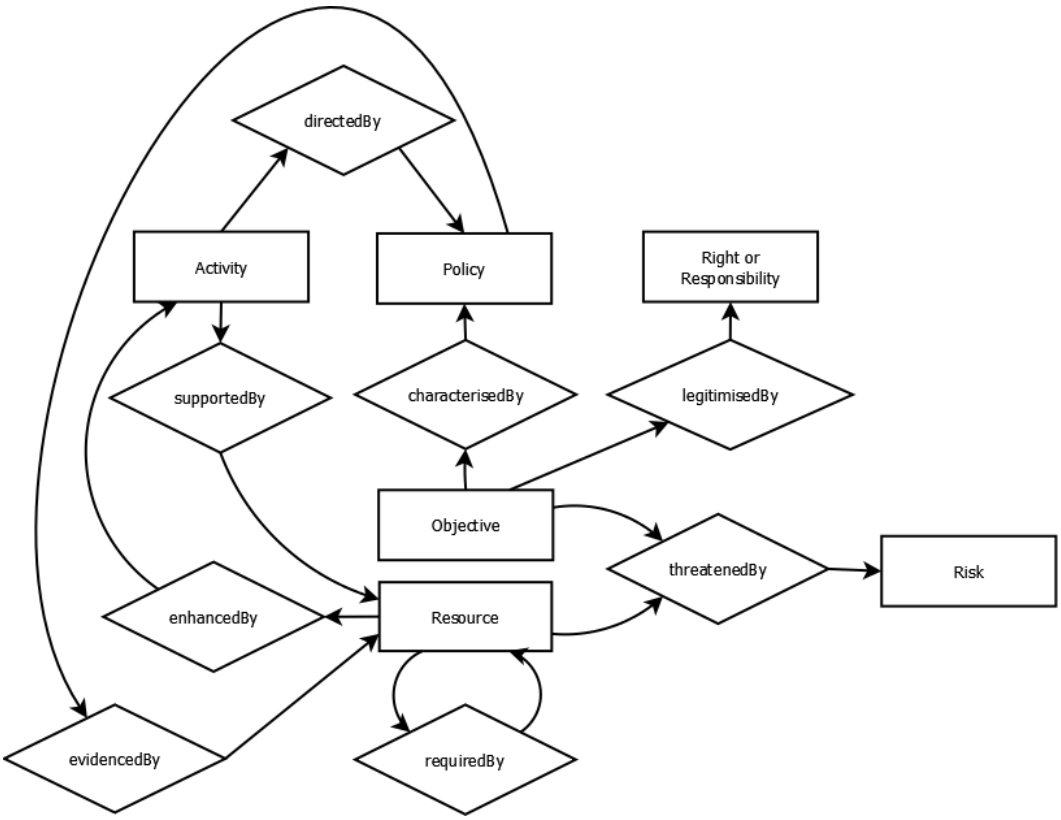
`mchughontology.hatii.arts.gla.ac.uk/ontologybrowser/customBuild.php`

This is the tool that was used to conceive the many relationships that comprise *PORRO*. It enables the individual building blocks to be related individually. The network effect that the ontology offers is the ability to relate concepts that may be non-contiguous in terms of causality but are nevertheless influential.

Initially relationships were developed between those concepts occupying a common limited area (e.g. Organisational $\rightarrow$ Legal issues) [structureBuild.php] before graduating to the development of relationships that spanned the entire conceptual space [customBuild.php].

We related preservation goals derived from reference literature such as *OAIS* and *TRAC* (and the case study details) to rights and responsibilities which motivate or legitimize them. Goals find their first practical expression through their relationship with parameters, which characterize them, illustrating what is required for their accomplishment (those characteristics that transform broadly defined goals into specific, measurable objectives). Parameters also direct the activities which are undertaken to satisfy them and are evidenced by specific resources (most often documentation such as policy, but sometimes implicitly, for example in software algorithms). Activities are supported by, and may also enhance resources. Resources may be dependent upon other resources. Based upon this latter relationship we represent semantic or structural dependencies between content information and infrastructure in *PORRO*. Figure 4.1 presents an entity relationship model that describes the classes and relationships encoded within the ontology.

Figure 4.1: PORRO Relation Diagram



4.2.1 Ontology Classes and Object Properties

The complete set of Classes within *PORRO* is available as Appendix A. The modelled object Properties are illustrated in table 4.8.

We consider generic statements of preservation objectives to be vulnerable to criticism because of the variety of preservation efforts, ranging in terms of physical scale, available

Table 4.8: PORRO Object Properties

Property	**Domain**	**Range**	**Inverse**
P01_has_goal	E021_Custodial_Entity	E029_Preservation_Goal	P01inv_is_goal_of
P02_legitimises	E274_Preservation_Right_Or_Responsibility	E029_Preservation_Goal	P02inv_is_legitimised_by
P03_characterises	E134_Preservation_Parameter	E029_Preservation_Goal	P03inv_is_characterised_by
P04_directs	E134_Preservation_Parameter	E303_Preservation_Activity	P04inv_is_directed_by
P05_supports	E445_Preservation_Support_Resource	E303_Preservation_Activity	P05inv_is_supported_by
P06_evidences	E445_Preservation_Support_Resource	E134_Preservation_Parameter	P06inv_is_evidenced_by
P07_enhances	E303_Preservation_Activity	E433_Preservation_Resource	P07inv_is_enhanced_by
P08_requires	E433_Preservation_Resource	E433_Preservation_Resource	P08inv_is_required_by
P09_is_threatened_by	E028_Functional_Entity	E591_Preservation_Risk	P09inv_threatens
P10_makes_more_likely	E673_Preservation_Risk_Influence	E591_Preservation_Risk	P10inv_is_made_more_likely_by
P11_makes_more_impactful	E673_Preservation_Risk_Influence	E591_Preservation_Risk	P11inv_is_made_more_impactful_by
P12_makes_less_likely	E673_Preservation_Risk_Influence	E591_Preservation_Risk	P12inv_is_made_less_likely_by
P13_makes_less_impactful	E673_Preservation_Risk_Influence	E591_Preservation_Risk	P13inv_is_made_less_impactful_by
P15_satisfies	E029_Preservation_Goal	E002_Preservation_Criterion	P15inv_is_satisfied_by
P16_has_source	E002_Preservation_Criterion	E004_Preservation_Criteria_Source	P16inv_is_source_of
P17_has_equivalence_with	E002_Preservation_Criterion	E002_Preservation_Criterion	P17inv_has_equivalence_with
P18_references	E002_Preservation_Criterion	E002_Preservation_Criterion	P18inv_is_referenced_by
P19_defines_as_evidence	E002_Preservation_Criterion	E003_Preservation_Criterion_Evidence	P19inv_is_defined_as_evidence_by
P20_is_comparable_with	E003_Preservation_Criterion_Evidence	E445_Preservation_Support_Resource	P20inv_is_comparable_with
P21_receives_funding_from	E021_Custodial_Entity	E023_Funding_Source	P21inv_funds
P22_has_staff_role	E021_Custodial_Entity	E026_Staff_Role	P22inv_is_staff_role_of
P23_preserves	E021_Custodial_Entity	E434_Preserved_Resource	P23inv_is_preserved_by
P24_contains	E443_Preserved_Performance	E434_Preserved_Resource	P24inv_is_contained_by

resource, the backdrop of legislation and compliance, and the broad range of data types and formats being preserved. We therefore define goals as equivalent to checklist criteria within repository certification standards (such as forthcoming standard *ISO 16363* which largely builds on community criteria originally released as *TRAC*). These are preservation cornerstones intended to represent a full range of the ambitions of the preservation practitioner, although specific implementation will vary. While our expression of goals is generic, each becomes specific and measurable by relation to one or more parameters which are typically qualifiers expressed as policy. There are 104 broadly stated preservation goals in total, intended to provide a comprehensive account of preservation aims, in order to reflect the diversity of goals evident throughout the digital library and broader preservation landscape.

We define resources as tangible or non tangible *stuff* within the preservation context that influence the existence or severity of risk. Resources include both those things fundamental to the preservation process (and normally intended to assist in the management of risk e.g. software access systems), and those that are valued in and of themselves, as part of a core business objective (e.g. preserved digital objects, financial profit). Resources are exposed to threats of loss or failure with consequences in terms of their contribution to risk management activities, and wider implications in terms of success of procedures and associated wider objectives. Resources' contributions to risk causation may be in terms of their insufficiency, associated conflicts, arising liabilities or a lack of their appropriate deployment. Repository roles, including 'typical' preservation roles such as data archivists and librarians, information architects, system administrators and developers and external roles such as depositors, consumers and information owners are characterised as Resources.

Where applicable, existing ontologies may have a role - we reference examples such as the *eXtensible Characterization Language* Ontology which provides robust property constraints for a range of formats and may be substituted where applicable [Puhl, 2009, Thaller et al., 2008, Becker and Rauber, 2011].

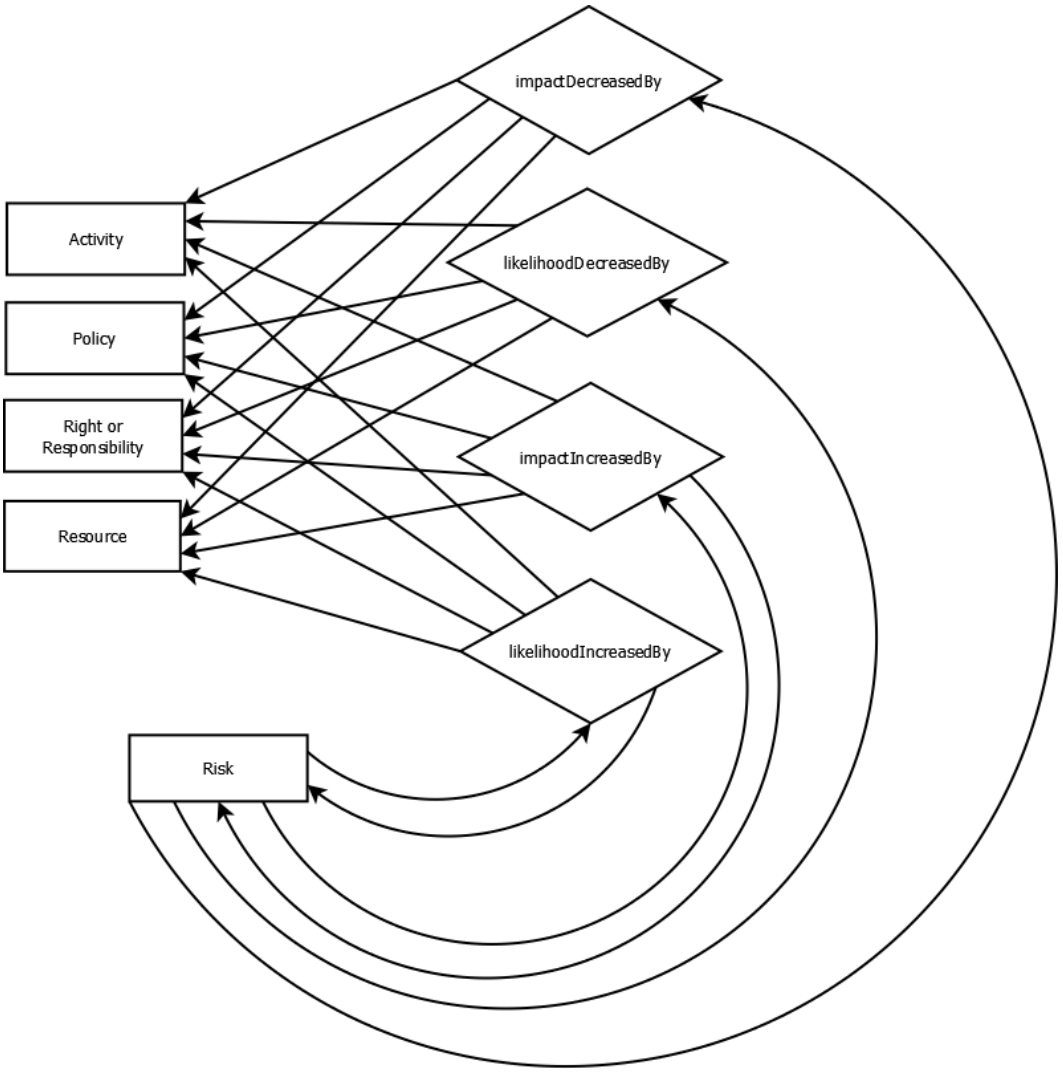
Rights or Responsibilities arise from the context within which preservation is undertaken; we define them to include any kinds of debts, obligations, liabilities or enablers. Contracts or legislative mandates are obvious examples. The conferment of mandate may be a risk limiter, or expose greater risks. A common risk that may arise is associated with incompatibilities between a particular liability and a business objective. A digitization project may face a conflict between its objective to make available digitized copies of its whole collection with intellectual property law liabilities which restrict the dissemination of copyright content. A consequential risk is that either the objective will fail or the liabilities escalate.

We define Parameters as those characteristics enacted in some aspect of operational policy (explicitly or otherwise), that lend a specificity and measurability to goals. Their value defines in a more tangible way the meaning of satisfying individual preservation objectives,

and reflects the diversity of the preservation landscape. For example, the establishment of a designated community (a generic goal) depends upon designated community composition and understandability definitions (each enacted in corresponding policy documentation or other resources, such as a database of users).

We present Activities as those processes intrinsic to the preservation context, associated with the execution of specified policy and/or the resolution of identified risk (including for example the enacting of a preservation plan). Activities may influence risk if they introduce conflicts, are incapable of satisfying their purpose, or are absent where otherwise required.

Figure 4.2: Illustration of Risk Cause and Effect



Consistent with our previous work we define Risks as the expression of the likelihood and impact of an event with the potential to influence the achievement of an organisation's objectives [McHugh et al., 2007]. Risks do not arise or exist in isolation and can both influence, and have their severity or realization determined by other risks, in various ways. Risk types may be broadly subdivided into risks of failure (directly threatening objectives), loss (threat-

ening resources or activities, and indirectly affecting objectives), or liability (imposing liabilities which again threatens objectives). Risks can be influenced by the existence, absence or specific characteristics of individual activities, resources, rights and responsibilities, and by the cumulative effects of multiple concurrent factors. More importantly, risks can both follow from or be rendered more severe as a consequence of other risks. The range of relationships between Risks and other *PORRO* categories of content is illustrated in Figure 4.2. This presents expressions of risk causality and mitigation, including the at times compounding quality of certain risk exposure. The emergence of an individual risk can increase the likelihood and impact of other risks. For example, the property "*Staff Suffer Deterioration Of Skills*" has the relationship "*makes more likely*" with several other risks. Evidence of particular Activities, Policies/Parameters, Rights and Responsibilities and Resources can each similarly affect the probability or potential impact of any given risk, but we model not only a relationship of increased risk exposure, but also mitigation. These relationships are made explicit with one of four properties, corresponding to both risk likelihood and impact either increasing or decreasing.

We assume no priority in terms of causal and consequential factors; *PORRO* is sufficiently expressive to encapsulate not only overtly risk-related factors, but also descriptive, technical and administrative information that may be subsequently relatable to risk. Unlike more static existing resources *PORRO* supports not only the repository evaluation or risk assessment exercise, but also documentation in a more general sense. The ontology approach offers extensibility to enable the adoption of existing domain specific descriptions where necessary.

In the context of our goal to provide semantic structure to our understanding of digital preservation systems we draw a distinction between the development of a taxonomy of terms (preservation system facets) and a structure that reflects properties or linkages between them, manifested as our overall ontology. Both are of intrinsic value and reflect the various use cases for PORRO more generally. The iterative process of developing PORRO took as its starting point accounts from real world audits, increasingly granularised to an atomic level, whereby system characteristics and qualities identified within the course of institutional audits were isolated and characterised as individual ontology elements. A compelling question was of the appropriate level of granularity. Starting with broad functional distinctions (drawing mainly from the widely acknowledged OAIS functional model [ISO 14721, 2012] we divided further, initially based on those topical issues that had emerged as areas of enquiry during the course of our assessments. These ultimately would be manifested as organisational objectives; our set of objectives is intended to provide a comprehensive selection of goals that preservation organisations are established to pursue. Risks too were defined by reference to the specific points of concern highlighted in our audits and the self-assessment responses recorded within *DRAMBORA Interactive*. The existing risk catalogue published within the original *DRAMBORA* release also provided a compelling resource in the definition

of risk elements. Those other elements that we record in the ontology (and specifically the level of granularity with which they were described) were determined by the relationships that they supported. In order to characterise associations between goals and risks we were required to define elements that were suitably independently meaningful and distinguishable. We remained resistant to incorporating elements that were implementations of more general concepts. The process was partly artisan (as manual ontology development typically is), but in order to ensure it remained systematic, representative and functionally appropriate we sought constant feedback from project colleagues (for instance in the Digital Curation Centre and 3D Coform) as well as end users and practitioners with whom we engaged in the course of resource development. While we sought to ensure that we were not constrained by existing top down expressions of best practice (we preferred to view with scepticism their legitimacy and claims of authoritativeness) we nonetheless referenced them regularly to ensure compatibility and that no obvious omissions were evident in our own taxonomy and ontology.

The world view we express with the development of *PORRO* is broadly compatible with the *DELOS Digital Library Reference Model*. We extend this approach however, adding with *PORRO* the dimension of Risk to enable the formation and expression of meaningful interrelationships between discrete system facets. The *DELOS* Model's Quality Node is relevant for preservation planning and validation, but appears insufficiently robust to enable its simultaneous exposure to a wide number of system and object properties. Quality measures in the DL Reference Model are explicitly set, and classified as one of 'Generic', 'Content', 'Functionality', 'User', 'Policy', and 'Architecture'. Sub-classifications including 'Interoperability Support', 'Trustworthiness', 'Fidelity', 'Authenticity', 'Fault Management Performance' and 'Compliance with Standards' are clearly relevant to a preservation agenda. However, there is little opportunity to enforce information property-driven quality control over the preservation process, at least not at a suitably granular level.

PORRO is a new model that offers means for recording diverse information facets to support fully warranted preservation management decisions and conclusions. It makes accessible properties and considerations that while relevant may otherwise be overlooked. Likewise, it presents a holistic view of risk and risk implications, limiting the likelihood of risk impact silos and making explicit organizational and informational relationships, de-emphasizing boundaries. This is true both internally within organizations, and in a more global sense, whereby common characteristics can be established across preservation environments, in order to establish shared knowledge pools that may be broadly exploited.

4.3 Applying PORRO to Real World Circumstances

4.4 Overview of Use Cases

PORRO has four primary use cases. The first is to facilitate the identification of risk, whereby users or agents can traverse the knowledge base to identify linked concepts based on common contextual characteristics. Secondly we seek with *PORRO* to facilitate the resolution of identified risks, mapping risks to appropriate mitigation, whether that entails particular interactions or the creation or acquisition of particular resources, policies or mandates. Thirdly we seek to enable gap analyses to be conducted more straightforwardly, whereby traversal of the ontology can reveal appropriate policy definitions, interactions and resources to enable the accomplishment of stated goals. Finally we seek to support the validation of approaches, as seen in preservation planning. Traversing the ontology yields insights into whether the prioritisation of particular policy, procurement or activities is likely to be beneficial in terms of overall goals. We illustrate PORRO's capacity to satisfy these use cases within the context of two novel applications (in addition to the ontology browser/manager tools referenced above) that use the ontology that we have developed in the course of this research.

4.5 3D Coform Long Term Preservation Component

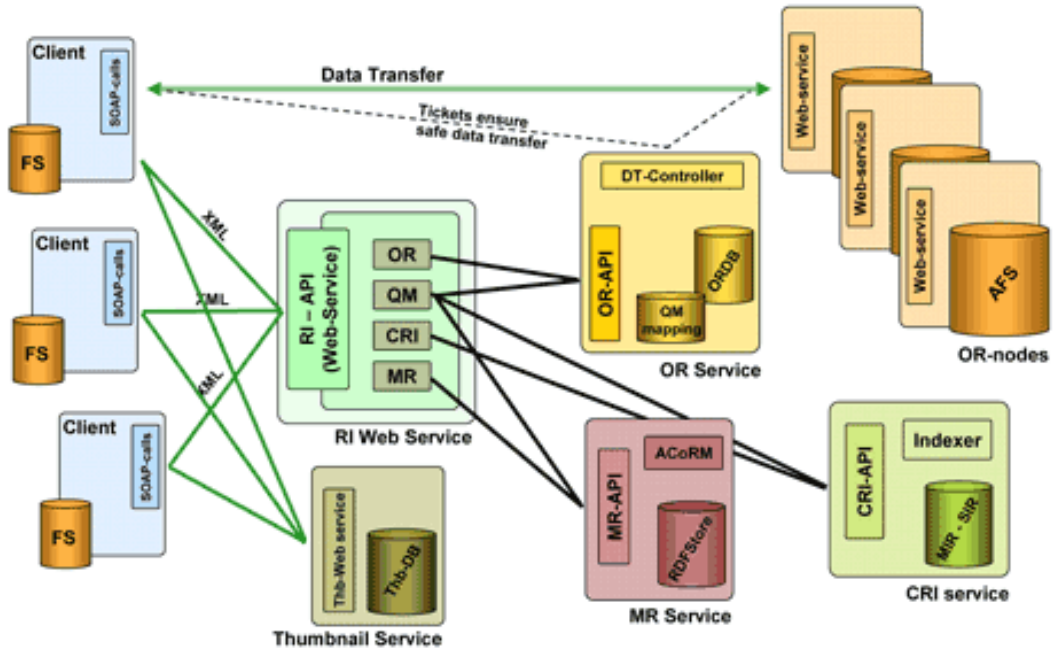
4.5.1 About 3D Coform

3D Coform [Tzompanaki et al., 2011] was a four year project funded under the EU Seventh Framework Programme focused on the sustainable documentation of tangible cultural heritage using 3D technology. This incorporated research and development of 3D capture, processing and repository software, and brought together a range of commercial, University and cultural heritage partners and collaborators including the Victoria and Albert museum, the Louvre and World Heritage sites in Cyprus.

We partnered on this project to contribute to the development of a metadata and object repository for 3D content collections and associated metadata. A critical part of the overall 3D Coform architecture, this relied upon the *CIDOC Conceptual Reference Model* (CRM) [Doerr, 2003] plus some digital extensions to record event based metadata relating to physical and digital materials (see Figure 4.3 which describes in high level conceptual terms the overall 3D Coform Repository Architecture - the preservation modules are part of the RI Web Service).

As part of this repository we developed a long term digital preservation component with the explicit aim to provide the repository with a means for the storage, description, distribution

Figure 4.3: 3D Coform Repository Architecture



and management of three-dimensional data and metadata. It consists of five individual elements. These are a *Preservation Level Manager*, which enables the conceptual formation of preservation packages (including data and metadata - Figure 4.4 presents a screenshot from the Java application); a *METS Export Manager* which enables the physical formation (encapsulation) of packaged data and export to METS format; a *Dependency Manager* which enables the recording of semantic and structural (technical) dependencies associated with rendering and manipulation of packaged data and or metadata; a *Preservation Risk Manager*, established to record and iterrelate preservation risks associable with dependencies and other contextual and infrastructural factors (see Figure 4.5, also presenting a screen from the Java application); and finally an *Obsolescence Manager*, established to prompt preservation interactions on the basis of evident risk exposure.

Our *Preservation Risk Manager* illustrates where and how generic risk factors and risks are manifested within the *3D-Coform* information space. Instead of manually encoding risk relationships between *3D-Coform* content we took PORRO's more generic examples and mapped these to *3D-Coform* information elements to understand where risk exposure may reside. A variably granular level of mappings are permitted between *PORRO* elements and either *3D-Coform* information types (based on a type-taxonomy developed within the project) or specific instance values. One can map a particular generic resource (e.g. Ingest Platform), with the broadly encompassing *DeviceType* Laser Scanners, or, if it is more appropriate a specific individual model or example of laser scanner.

Similarly, we model information dependencies as specialist *PORRO* resource entities, based

Figure 4.4: 3D Coform AIP Manager

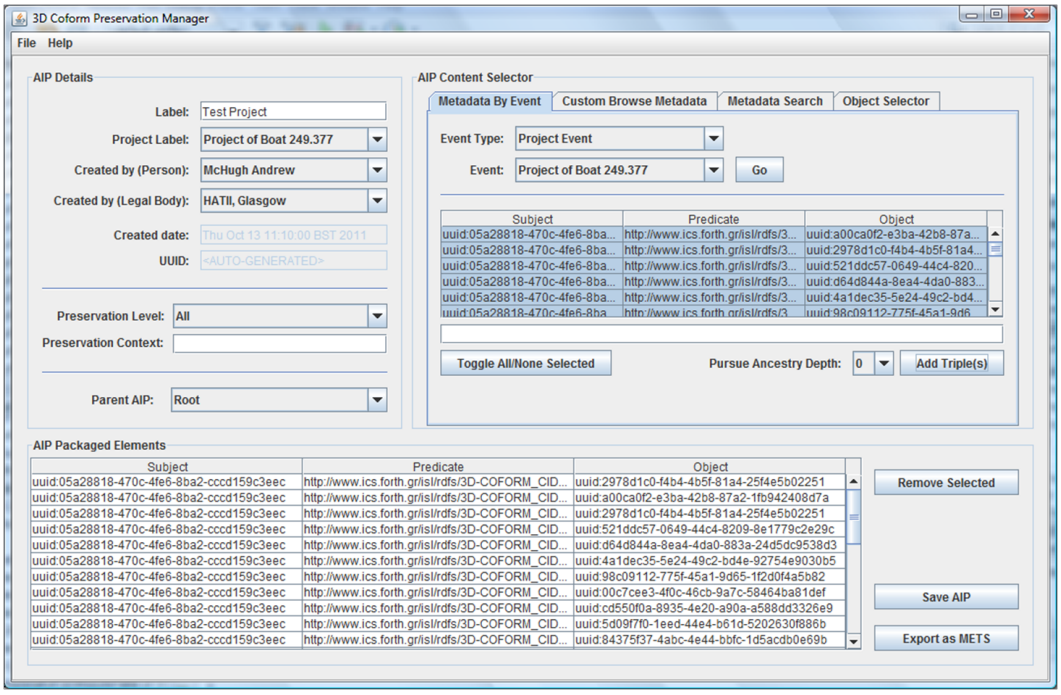
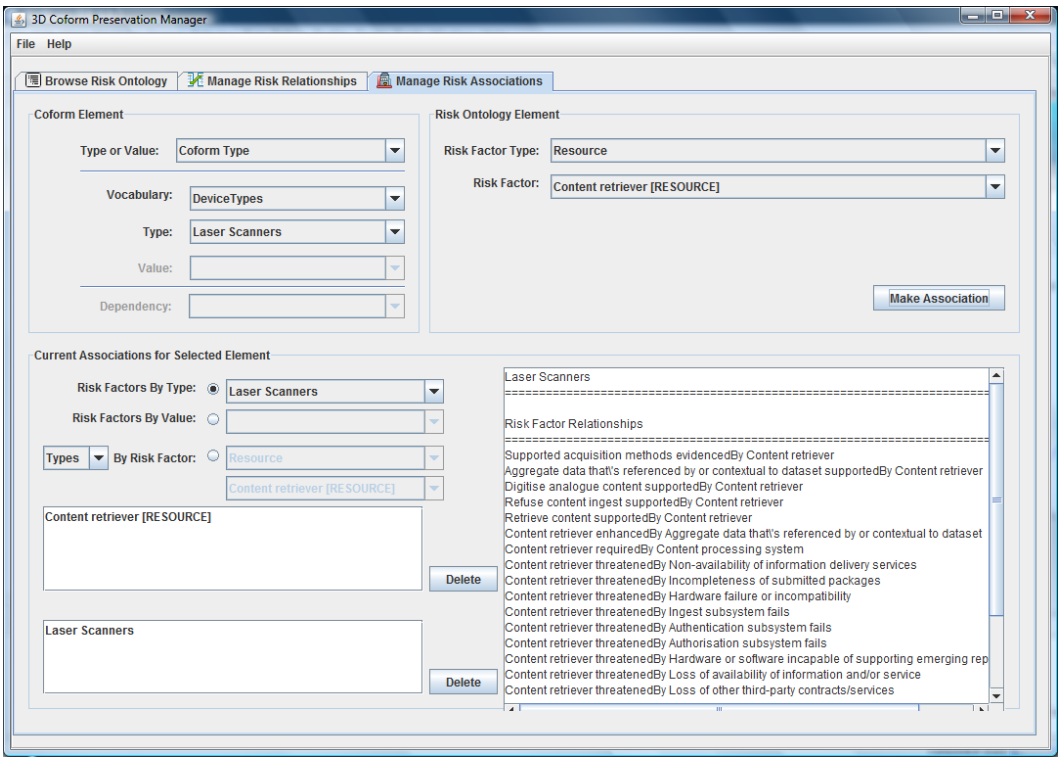


Figure 4.5: 3D Coform Risk Association Manager



primarily on *OAIS*' concept of *Representation Information* [ISO 14721, 2012]. These are implicit within archival packages, which are therefore self-evident. Whether functionality (i.e. tools) or just documentation are actually encoded/exported depends on a defined preservation level. Properties are measurable facets of function. Functional components can

exist hierarchically, and single functional behaviours' may be grouped into wider functions. Multiple versions of any individual dependency may exist; this may mean variability of rendering, processing, and of preserved outputs, which may differ from an "original". Different versions share function, but may exhibit material differences. Within the context of each version there must be an explicit mapping between content and dependency elements.

Traversing *PORRO* reveals relationships with other mapped content, or challenges the user to determine whether generic activities, policies or resources which appear to be required have been adequately implemented. In tandem with a preservation package manifest, which is also created within this long term management tool, this enables a clear risk profile to be presented, with closely associated risks and potential additional risk mitigation approaches clearly identifiable, albeit generically expressed.

4.6 Collaborative Assessment of Research Data Infrastructures and Objectives

Overview of CARDIO

The *Collaborative Assessment of Research Data Infrastructures and Objectives* (CARDIO) tool was developed with colleagues from the *Digital Curation Centre* (DCC) [DCC, 2012], a JISC [JISC, 2012] funded service that provides leadership and expertise in the curation and management of digital resources. Most recently the DCC has focused on issues surrounding the management of research data. Like digital preservation more generally, research data management's success is reliant upon a range of elements. The distribution of influences is not limited to infrastructure, and also implies a number of internal stakeholders with a range of responsibilities.

CARDIO is a tool and associated workflow for performing data management maturity and capability assessment across a data context (typically an institution, project, data centre or department), which supports and demands a collaborative approach. Its origins were in discussions within the DCC that revealed an appreciation of the importance of incorporating a range of perspectives into the evaluation of data management infrastructure. Its central conceit is that only by adopting a holistic view can such activities succeed. Individual functions, roles and systems are considered not in isolation but in terms of their relationships with and influence from others elsewhere. It was conceived to ensure a broad discourse and consensus-based conclusions. To that end the tool presents a process that demanded firstly the submission of a survey instrument by each participant and latterly their consideration and evaluation of peer responses and agreement of a shared perspective. It uses social tools and online survey instruments but has operated as successfully as an offline managed process.

In more specific terms, *CARDIO* requires participants to reflect on data management maturity in thirty individual areas. For each, users must apply a rating of 1 to 5 to reflect their perception of how well their institution performs in that area. Some supporting text provides details indicative of what it means to score 1, 2, 3, 4 or 5 in that area. Users also have the opportunity to respond with a "Don't know" answer, or to declare a particular issue not-applicable. Users may clarify or contextualise their selections with additional information in the form of free text or uploaded documentation.

We developed *CARDIO* to promote institutional discourse and a collaborative approach to data management problem solving; provide reassurance of infrastructural capacity and satisfaction of data management planning commitments; highlight priority areas for resource investment or areas where investment will have greatest impact; relate local data contexts to the wider world of data management via a shared evidence base; and facilitate engagement with senior management over data management responsibilities and shortcomings. These are of course complementary to the goals of *PORRO*. The notion of a shared knowledge base in particular is a reference to the role of the ontology in presenting a structured view of digital preservation or data management best practice.

CARDIO has established a role in promoting widespread analysis across multiple perspectives within a single institutional setting. It has been a core utility within the *Digital Curation Centre*'s institutional engagement program, whereby around eighteen UK HE institutions have benefited from consultancy services aimed at boosting data management capacity and understanding[1].

CARDIO is manifested as both an online interactive resource [DCC, 2011] and a traditional methodology tool that entails interviews, focus groups and collaborative reporting. Managed deployments in 2012 included studies at London School of Economics and Queen Mary University London. More details are included as part of Chapter 5, which focuses on *PORRO*'s evaluation.

CARDIO and PORRO

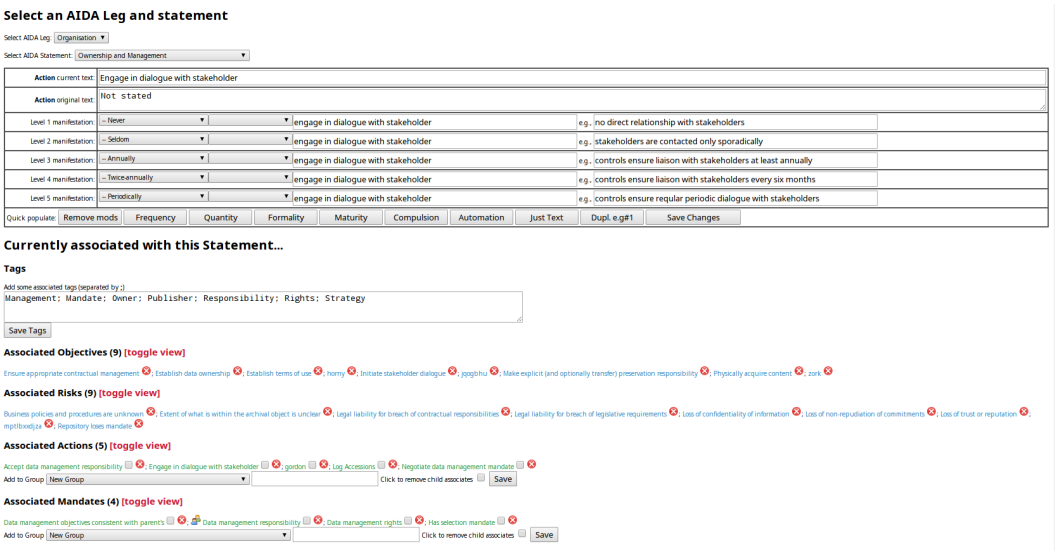
CARDIO enables comparison with a consolidated collection of real world data, via the *DCC CARDIO Knowledge Base* which is founded upon the *PORRO* ontology. *PORRO* has been encoded to correspond with the thirty infrastructural facets that comprise the *CARDIO* assessment, ranging between issues of organization, resources and technology. The mapping is actually to the University of London Computer Centre's *Assessing Institutional Digital Assets* model for digital preservation [AIDA, 2010] which in turn has its origins in the University of Cornell's [Kenney and McGovern, 2003] three leg / five stage model for digital

[1] See http://www.dcc.ac.uk/tailored-support/institutional-engagements

preservation. This enables the provision of relevant considerations in each specific area, as well as the illustration of more complex risk relationships. It means that users can identify areas of perceived weakness and explore opportunities for effective resolution, or alternatively challenge perceived strength with robust gap analysis.

A bespoke web tool was developed using *PHP/MySQL* and *AJAX* to facilitate the creation of mappings (see Figure 4.6 which shows the tool developed to map AIDA stages with PORRO classes and in turn link to five stages of organisational maturity). The tool enables the association of *PORRO* elements to *CARDIO* statements (survey concepts). Furthermore it supports the streamlining of five levels of manifestation for each. This means that mappings between *CARDIO* and *PORRO* concepts can be enriched with detail about the extent to which *PORRO* concepts are realised. Built in tiers of manifestation are based on frequency (e.g., how often particular *PORRO* actions are undertaken), quantity (e.g., how much of a particular resource an organisation has), formality (e.g., the extent to which policies are formally enacted and documented), maturity (e.g., a measure of how well established a policy is) and compulsion (e.g. a measure of how serious a particular responsibility is).

Figure 4.6: Example CARDIO Mapper Application



Mappings between *PORRO* classes and *CARDIO* concepts are intuitive and straightforward. An example of CARDIO's mappings to *PORRO* follows. Given the close association between *CARDIO*, *AIDA* and the Cornell model this mapping additionally serves as a connection between these prior resources and *PORRO*.

Example CARDIO Mapping - AIDA Element "Technological Infrastructure"

Corresponding Goals

- E053_Establish_Appropriate_Hardware_Infrastructure

- E057_Establish_Appropriate_Software_Infrastructure
- E064_Establish_Assurances_Of_Sufficiency_Of_Staff_Skills_And_Capacity
- E074_Establish_Hardware_Upgrade_Policy
- E093_Establish_Software_Upgrade_Policy

Corresponding Parameters

- E182_Media_Refreshment
- E239_Rights_And_Ownership_Definitions
- E260_Supported_Systems_And_Applications
- E261_Systems_Development_Management
- E264_Technology_Licensing
- E265_Technology_Skills_Development

Corresponding Activities

- E318_Develop_Technical_Training_And_Induction
- E353_Liaise_With_Technology_Provider
- E393_Plan_And_Execute_System_Upgrades
- E401_Record_System_Changes
- E404_Refresh_Media_Or_Hardware
- E409_Report_Technical_Status
- E418_Review_Technical_Provision

Corresponding Resources

- E479_Custodial_History_Record
- E480_Custodial_History_Records
- E502_Formal_Contracts_And_Terms
- E587_Update_And_Upgrade_Prompts

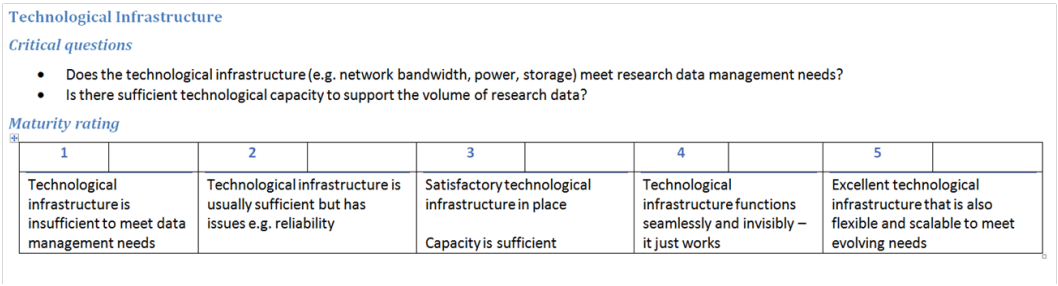
Corresponding Risks

- E596_Authentication_Subsystem_Fails
- E597_Authorisation_Subsystem_Fails
- E620_Hardware_Failure_Or_Incompatibility
- E631_Ingest_Subsystem_Fails
- E657_Non-Availability_Of_Core_Utilities
- E658_Non-Availability_Of_Information_Delivery_Services
- E667_Software_Failure_Or_Incompatibility

The power of the associations is evident when referred to following a typical *CARDIO* assessment. According to its established workflow, participants are given little insight into the

meaning of particular categories, instead referred to short excerpts of text aligned to each of five nominal maturity levels and asked to select the one that they think is most representative of their own context's circumstances.

Figure 4.7: Example CARDIO User Prompt



Technological Infrastructure

Critical questions

- Does the technological infrastructure (e.g. network bandwidth, power, storage) meet research data management needs?
- Is there sufficient technological capacity to support the volume of research data?

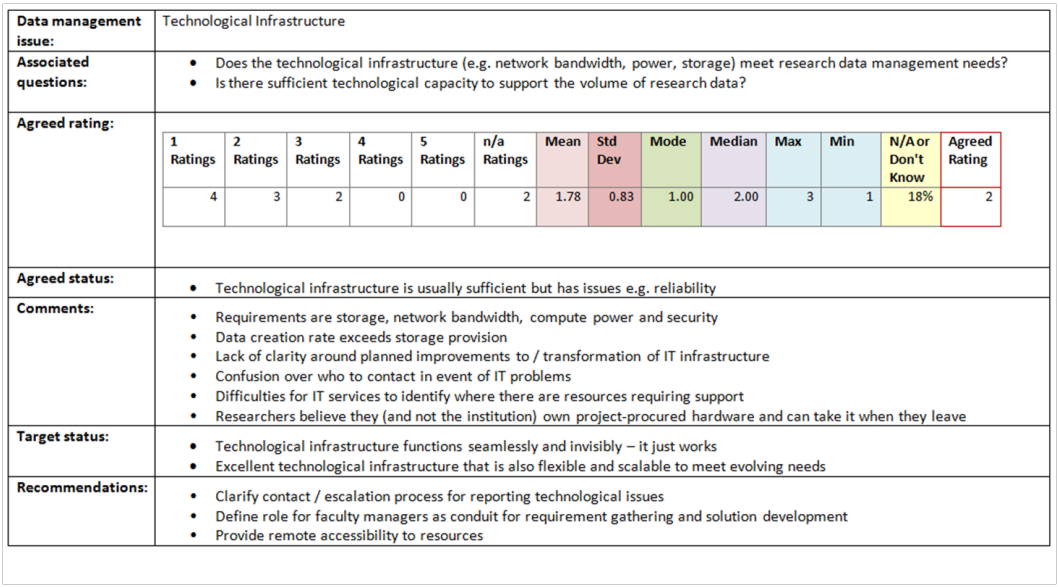
Maturity rating

1	2	3	4	5
Technological infrastructure is insufficient to meet data management needs	Technological infrastructure is usually sufficient but has issues e.g. reliability	Satisfactory technological infrastructure in place Capacity is sufficient	Technological infrastructure functions seamlessly and invisibly – it just works	Excellent technological infrastructure that is also flexible and scalable to meet evolving needs

Figure 4.7 is an example form with an illustrative range of options.

Where the process highlighted shortcomings the tool is intended to provide guidance on what should be implemented in order to improve and to assist in the construction of a case for why this was necessary. Linking risk to each CARDIO element immediately provides a platform upon which the latter can be achieved. Making a case to senior management is demonstrably more successful with a coherent and succinct alignment of possible risks associated with inaction in any given area. Furthermore, the association illustrates appropriate responses to shortcomings and makes explicit gaps in existing provision.

Figure 4.8: CARDIO Example Report Excerpt



Data management issue:	Technological Infrastructure
Associated questions:	• Does the technological infrastructure (e.g. network bandwidth, power, storage) meet research data management needs? • Is there sufficient technological capacity to support the volume of research data?
Agreed rating:	

1 Ratings	2 Ratings	3 Ratings	4 Ratings	5 Ratings	n/a Ratings	Mean	Std Dev	Mode	Median	Max	Min	N/A or Don't Know	Agreed Rating
4	3	2	0	0	2	1.78	0.83	1.00	2.00	3	1	18%	2

Agreed status:	• Technological infrastructure is usually sufficient but has issues e.g. reliability
Comments:	• Requirements are storage, network bandwidth, compute power and security • Data creation rate exceeds storage provision • Lack of clarity around planned improvements to / transformation of IT infrastructure • Confusion over who to contact in event of IT problems • Difficulties for IT services to identify where there are resources requiring support • Researchers believe they (and not the institution) own project-procured hardware and can take it when they leave
Target status:	• Technological infrastructure functions seamlessly and invisibly – it just works • Excellent technological infrastructure that is also flexible and scalable to meet evolving needs
Recommendations:	• Clarify contact / escalation process for reporting technological issues • Define role for faculty managers as conduit for requirement gathering and solution development • Provide remote accessibility to resources

Figure 4.8 is an exerpt from a report illustrating the views collected from one study about the issue of *Technological Infrastructure*. Tracing corresponding PORRO classes and named individuals via the ontology browser (see above) makes explicit such recommendations where

there are shortcomings. Indeed, we can take an example corresponding risk "E658_Non-Availability_Of_Information_Delivery_Services" and traverse the ontology to identify a range of associated risk causation and mitigation factors. An excerpt of the links follows, and reveals the value of a holistic mapping of information system and organisational components.

E651_Non-Availability_Of_Information_Delivery_Services
- P12inv_is_made_less_likely_by [E666_Preservation_Risk_Influences]
- E353_Maintain_Access_Platform
- - P05inv_is_supported_by [E438_Preservation_Support_Resources]
- - E528_Network
- - E562_Security_Platform
- - E499_General_Hardware
- - E500_General_Software
- - E497_Format_Support
- - E441_Access_Platform
- - - Database of affiliate stakeholders
- - - Means for logging access attempts
- - - Secure read/write access from affiliates via FTP
- - - Software infrastructure for encrypting data for transfer
- - - Software infrastructure for personalisation of user experience (e.g. MyData system)
- - - Suitable FTP server (e.g. ProFTP)
- - - Suitable web server software (e.g Apache)
- - - System for controlling information access
- - - Tivoli access manager
- - - Web browsing interface
- - - ...
- - - Another, e.g., remote access arrangements
- - - - P07inv_is_enhanced_by[E296_Preservation_Activity]
- - - - E399_Regulate_Access_To_Data
- - - - - P04inv_is_directed_by[E127_Preservation_Parameter]
- - - - - - E135_Content_Closure
- - - - - - E148_Cost_Model_For_Access_Provision
- - - - - - E198_Policy_On_Access_Control
- - - - - - E260_Terms_Of_Access

A customised graph can be generated to illustrate related information facets - looping through the hierarchy reveals deeper lying considerations that if resolved may in turn benefit multiple identified issues. These implicitly provide a more systematic and coherent basis for resolving preservation or data management shortcomings and offer a constructive, tailored

mechanism to plan improvement that is independent of repository maturity or priorities. Not only risk resolution can be managed in this way. It is similarly straightforward to select a given preservation goal for example and trace it to identify associated drivers and facilitating factors.

4.7 Summary of the Work

In the course of this thesis we have built a preservation process and metric for evaluation around the concept of risk. A given risk (or its inverse) characterises one or more corresponding preservation objectives - completing an objective means avoiding a risk - and risk management is manifested in terms of those organisational, infrastructural and policy facets that accomplish something. We have delivered a structured knowledge base for approaching repository development and evaluation as well as several ancillary and complementary results.

We start with an objective identified within a range of institutions: to establish and understand best practice for taking custodial responsibility for digital information. This increasingly prominent goal is overshadowed by widespread uncertainty. We developed *DRAMBORA* to provide practical support to those organisations aspiring to best practice, but underserved by highly prescriptive metrics that were available. In fact, *DRAMBORA* revealed as much to its users as it did to our own research. Its development entailed a number of exploratory audits within a range of organisations, enabling us to shape an appropriate methodology for self and supported evaluation, but also exposed a set of practical evidence of how information is preserved, and of those factors that contribute to success and failure. The systematic analyses comprising this work were unprecedented and in a context often characterised by organisations' uncertainty about the future and hesitation to invite scrutiny, the data that was collected was uniquely valuable.

The assessments, and *DRAMBORA* itself also revealed a shortcoming of a wholly bottom-up approach. Although clearly more customisable to a given set of organisational objectives (essential in such a heterogeneous arena) the lack of a clear objective benchmark was frustrating to users. When we supported self assessments in the development of our methodology feedback clearly identified the value of an expert contributor, whose function became not only to advise on methodological aspects of self assessment, but also to relate perceptions of maturity to a broader context. How then to take our established body of best practice and expose it as part of self assessment, essentially enabling that information to perform the role of an expert facilitator.

Our interactive tool takes the core methodology of reflective introspection and adds a means by which users can relate their efforts to those elsewhere, although still short of a complete

inter-connected knowledge base of best practice. That almost one hundred and fifty organisations have made extensive use of the tool in one or more self-assessments speaks to its value - this usage far exceeds that associated with any other systematised preservation assessment tool or instrument. The variety of countries and associated disciplines evident in these statistics (described in full in chapter 2) lends further credibility to claims of the tools success.

This methodology is a contribution of critical importance for the preservation community. Many of the efforts to date associated with evaluating digital preservation have been driven by the pursuit of validation of the competencies of a given preserving organisation. This has yielded certification standards but little in the way of formal process for their application. Our method is firmly aimed at those doing preservation, and provides the means for them to be confident in the suitability and sufficiency of their work, and if appropriate to facilitate a later exposure of their efforts to external certification. It also ensures that the dynamic nature of preservation (which may change based on new technological innovations, such as the increased dependency on cloud computing) continues to be reflected in best practice.

As *DRAMBORA* benefited from its deployment in formal audits, it similarly equipped us to pursue our second research objective, a continuation of the work of surveying preservation contexts. *DRAMBORA* Interactive reveals data corresponding to assessments that exhibit diversity in geographic location, legal context, types of digital collection, mandate and budgetary model. *DRAMBORA* requires participants to describe their preservation efforts in terms of objectives, activities, resources and risks, providing immediate evidence of how leading organisations go about ensuring the longevity of digital materials. We undertook a series of systematic audits based on evolving methodological and intellectual criteria (themselves developed iteratively based on our findings). Good and bad practice were both identified, characterised and related within these assessments, providing evidence that would form the basis for lines of enquiry and evaluation feedback in subsequent audit exercises. Collectively, the assessments yielded several benefits. The first, taken at face value was that the participating organisations were given the opportunity to better understand their successes and shortcomings, and to adapt to a critique based on comparison with a combination of objectively conceived and empirical real world best practice. Secondly, we could refine our approach both in terms of process and more importantly in the intellectual basis upon which our evaluation was conducted. As we learned more about how preservation takes place we became better equipped to identify opportunities for improvement elsewhere. Our knowledge base was developing, and it became clear that the perspectives we had been granted by exhaustively assessing a series of operations were unique. The existing instruments that we were using were highlighted as being occasionally incomplete, more commonly at least lacking in terms of specificity or applicability to real world motivations and approaches.

The outcomes of each audit were recorded, accomplishing our third research objective. We

iteratively developed a taxonomy of concepts from these audit reports, characterising our evidence base in terms of what organisations wish to achieve (agnostic of any given objective standard) and the processes, tools, policies and mandates that inform and/or support that. These were in turn related to a developing catalogue of risks, whether caused by or mitigated by these factors. Each facet was recorded in two forms; a higher level, more generic expression was intended to be immune to issues of applicability across context and time, complemented by specific examples the particular implementations or manifestations we observed and recorded in our audit experiences. Injecting semantic qualities to the data is of tremendous importance, as it allows the data to be interrogated by applications or human users and enables the conception of a network of relatable factors that contribute to preservation outcomes.

Fourthly we evaluated existing methodologies for undertaking preservation assessment. A critical dimension of this work has been to establish where our outcomes are positioned within an existing international preservation certification landscape. We are not content to seek to replace wholesale the existing provisions, several of which have enjoyed formal standardisation. Instead, we seek to identify and fill the gaps in what currently exists. We considered the value and applicability of several standards and de facto standards and offered a critical assessment of each. Several leading examples are collectively encapsulated within the *European Framework for Audit and Certification of Digital Repositories*, which presents a series of increasingly onerous certification tiers that correspond with the *Data Seal of Approval*, *ISO 16363* and the equivalent German standard. The first two tiers require just documented self-assessment while the most involved requires a full externally administered audit to be conducted within any organisation seeking certified status. We reflected on the many positive aspects of these resources. Each has at least some intellectual basis in the *Trusted Repository Audit and Certification Criteria and Checklist* which can be considered a seminal resource in this area. By extension, each can be considered a valuable expression of generic aspects of preservation practice. *TRAC*'s formal standardisation can be considered a more practical expression of - and companion resource to - the equally seminal *Reference Model for an Open Archival Information System*. However, their shortcomings are mainly in terms of their utility and practical applicability. Even though self-assessment comprises two-thirds of the Framework there remains little explicit emphasis on processes to guide a prospective repository administrators seeking to evaluate his or her efforts. While the *Data Seal of Approval* presents as an advantage its low barrier to entry this is accompanied by shortcomings, principally in terms of lack of granularity of coverage. The *TRAC* and ISO standards, pursuing exhaustiveness, extend to many, many criteria; preservation is a complicated business with implications spanning every aspect of an organisation's administration, technology and information management process. But they in turn expose themselves to criticism as impractically conceived, beset by uncertainties in terms of how metrics can be

satisfied and based upon a set of preservation requirements so generic that it doesn't really exist within any single organisational context.

In the area of ontology evaluation there are a number of possible alternative methods to assess the quality and correctness of ontologies. There are likewise several criteria upon which ontologies can be evaluated. Hlomani and Stacey [Hlomani and Stacey, 2014] offer ten determining factors, which are quite compelling. These include accuracy (does the ontology reflect expert knowledge about the domain); adaptability (the extent to which the ontology supports specialisation or extensibility); clarity (how well it communicates its terms); cohesion (the extent to which classes are related); completeness (can it answer all appropriate competency questions); computational efficiency (relating to the speed that tools can work with the ontology); conciseness (limiting redundant or unnecessary elements in the context of the domain); consistency (does not include or allow for contradictions); coupling (connectivity with existing ontologies); and coverage (how well is the modelled domain represented).

In order to determine the extent to which any given ontology demonstrates these traits the literature offers four primary evaluation approaches. The first is the gold standard comparison [Maedche and Staab, 2002] whereby a given ontology is compared with a definitive or authoritative ontology within the same domain. In our case there is no such gold standard and therefore this approach is of limited value. In fact, we aspire for PORRO to ultimately represent a gold standard ontology for encapsulating issues associated with the delivery of preservation services.

A second approach for evaluation is based on the evaluation of tools that use a given ontology [Porzel and Malaka, 2004]. We have employed this method with respect to the *Collaborative Assessment of Research Data Infrastructure and Objectives* (CARDIO) tool developed with colleagues in the context of the Digital Curation Centre's work with UK Higher Education institutions. Reassurances were offered in terms of the ontology's clarity, cohesion and computational efficiency by a combination of end user feedback and several information outcomes, whereby meaning and implications of user outcomes were clarified and gaps and shortcomings made more evident through the use of the tool.

We utilised further ontology evaluation approaches in our assessment of other *DRAMBORA* assessments that were undertaken within the context of several digital libraries. A corpus based approach, involving comparisons with a source of data (e.g. a collection of documents) about the domain to be covered by the ontology [Brewster et al., 2004] revealed the ontology's accuracy and coverage, as well as its adaptability to this specific sub-domain.

Finally, we leveraged a range of existing top-down approaches (such as *TRAC* and the *Data Seal of Approval* to assess how well the ontology meets a set of predefined criteria, standards, requirements [Lozano-Tello and Gómez-Pérez, 2004], providing evidence of completeness and conciseness. A full account of all these evaluation approaches is provided in Chapter 5.

We have conceived the *PORRO* ontology as a structured expression of preservation best practice, collated from over a dozen full-scale audit exercises in addition to around one hundred and fifty online self-assessments conducted using *DRAMBORA Interactive*. This qualifies this data as a legitimate consolidation of overall preservation practice, a unique dataset that was both conceived and validated by lengthy exposure to real world preservation efforts undertaken by experts in the field. We have sought to take advantage of the resources that are available and position our efforts in a fashion that ensures their compatibility. *DRAMBORA* was our direct response to the difficulties posed by a wholly top-down approach, a process-driven methodology that requires self-assessors to reflect on their own priorities and their associated strengths and shortcomings. *PORRO* enhances this process by providing pliable hooks to best-practice that are customisable to any given preservation context. This satisfies our fourth objective, a presentation of best practice in a taxonomical and ontological format. Like *DRAMBORA*, its design has been principally motivated by its associated use cases.

In isolation the value of an ontology is difficult to convey and therefore a suite of indicative tools that use the ontology as their intellectual foundation is an important step. The effectiveness of our prototype tool portfolio, the delivery of which is our final research objective, is evident within two operational contexts. The first is research data management, where its adoption as a data source for the *CARDIO* collaborative data curation evaluation tool has been demonstrably useful. That is a process that builds consensus of a given organisation's data management capacity. At the point where individual contributors have agreed upon the status of their existing efforts reference is made to *PORRO* to identify potential approaches to improve existing provisions. This can be considered *PORRO*'s preservation planning application. Within the *3D Coform Repository Infrastructure* (RI), *PORRO* is used in the identification of risk, whereby a given set of real world circumstances are identified within the ontology and traced to potential associated risk factors. This is what we mean by bidirectionality within *PORRO*. Its applicability is such that it can be used to provoke the development of preservation activities or resource acquisition, or to warn of threats associated with existing or proposed systems.

Further practical evidence of *PORRO*'s value can be seen by referring to the core use cases that *PORRO* is capable of satisfying. Through its adoption in the applications referenced earlier we can say that *PORRO* supports the identification of risk (whereby users or user agents can traverse the knowledge base to identify linked concepts based on common identified contextual characteristics); the facilitation of risk resolution (whereby risks that have been identified externally or using the ontology are mapped to appropriate mitigation measures); performance of gap analysis (whereby real world generic goals are represented in the ontology, as are prescribed criteria which in turn are mapped to *PORRO* concepts, both fleshed out by their correspondence to required or appropriate parameter considerations, actions or resources); and validation of approaches (whereby particular policy, resource or

activity prioritisation can be traced to corresponding objectives and risks which illustrate the appropriateness of investment).

Chapter 5

Evaluation

5.1 Introduction

Our evaluation seeks to demonstrate the extent to which our proposed ontology and its associated applications are sufficiently expressive, meaningful and usable to support institutional self-assessment. As noted below, evaluation of preservation approaches is not easy, given an inherent implicit temporal dimension, whereby the shortcomings of any approach may not be completely evident until years later.

We approach the challenge by referencing existing tools and ensuring comparable expressiveness, and by deploying the ontology in a range of institutional settings to verify its perceived usefulness and utility.

5.2 Long Term Preservation Evaluation

5.2.1 The Challenges of Preservation Evaluation

To date, the preservation community has wrestled with the challenge of empirically evaluating its efforts, with variable success. Unlike, for example, the information retrieval community [NIST, 2013] there is no widely accepted resource or infrastructure to empirically evaluate and compare results. This is partially a consequence of digital preservation's temporal dimension - it is difficult to evaluate success when its realization is not immediate. Nevertheless, recent years have seen the emergence of a range of evaluation and validation approaches spanning both infrastructural and more focused elements of preservation. The former can be further subdivided into top-down and bottom-up approaches. The *Trustworthy Repository Audit and Certification Criteria and Checklist* and the associated *ISO 16363* standard each detail characteristics that should be demonstrable in trustworthy preservation

environments. Our *Digital Repository Audit Method Based on Risk Assessment* (DRAMBORA) is a more subjective means of determining repositories' fitness for purpose, based on their own specific priorities and responsibilities. Both approaches have been rigorously applied in a range of contexts, and therefore a considerable quantity of data exists describing not only instances of repository conformity, but also identified risks, which may be associated with both infrastructural and information characteristics.

Elsewhere, in the context of more micro-level evaluation of specific preservation actions, *Plato's* evaluation process employs utility analysis to determine the suitability and viability of specific migration or emulation approaches given a particular organizations data and organizational requirements and obligations. Similarly, the Planets Testbed provides a common data corpus and an experimental environment to facilitate more objective, comparable and re-creatable evaluation of preservation tools and processes. We can also look to criteria published by Library of Congress for influential factors for file format evaluation, corresponding to issues of sustainability, fidelity and functionality, and work exploring formats' vulnerability to information loss from file corruption [Thaller et al., 2008].

Much of this work has focused on support for efficient retention of content properties, and its collective success is dependent on a number of factors. One must establish the capacity to characterize content properties, in unambiguous terms that are suitably comprehensive. Secondly, one must be able to measure and subsequently validate these properties, in order to determine their prolonged existence or availability during a preservation timeline. Finally, one must establish an understanding of infrastructural and contextual influences on both materials themselves, and any proposed preservation interventions. *PORRO* can be used to ensure that chosen properties of content and context are comprehensive and that their area of influence (and that of any proposed preservation intervention) is sufficiently well understood.

5.3 Comparison with Best Practice

Our proposal implies a capacity to represent and illustrate a full range of digital preservation facets and their relationships irrespective of discipline or domain and support and inform validation and planning activities (and ultimately their automation). It is in these terms that evaluation is undertaken.

Our objectives can be distilled into core desirable qualities of completeness, applicability and usefulness. The first is evaluated by comparison with evidence of existing practice. To date, discounting "spambot" and other erroneous registrations, around three hundred and fifty repositories are registered as users of *DRAMBORA*'s interactive online tool, representative of institutions including national libraries and archives, academic research reposito-

ries, commercial data centers and financial services institutions. Of these, the database and logging systems reveal high usage activity from around one hundred repositories. Mapping user submissions to *PORRO* has revealed the ontology's breadth is sufficient. Its dual tiered approach to recording information facets (with generalized entries linking to more specific example 'implementations') ensures its scalability to encapsulate emerging trends while maintaining its generic qualities and without becoming skewed in any specific disciplinary direction. Likewise, its alignment with de facto standards such as TRAC, and by extension to forthcoming standards like *ISO 16363* (as described above, the ontology contains mappings to *TRAC* criteria) provide further reassurance of its completeness, at least in terms of scope.

The validity of the ontology, and the extent to which it is representative of the real world is assessed by exposure to real world scenarios. *PORRO*'s integration within novel tools and its increasingly prominent role in high profile Digital Curation Centre engagement activities provide some assurances of its effectiveness. Likewise, this provides some evidence of its applicability to diverse domains. Within *3D-Coform* mappings have been made between ontology elements and discipline-specific terminology with success.

We turn to testimonials of stakeholders involved in order to further validate the ontology. Reflecting upon its role in supporting checklist-based audits provides further evidence. The *CARDIO* tool uses *PORRO* to inform breadth of data management requirements and responsibilities. It relates tangible provisions to a conceptual model for data management and therefore provides meaningful information to support improved implementation. Pilot collaborations with colleagues at London School of Economics and the Queen Mary University, London have generated excellent feedback based on the ontology's role not only in the evaluation of existing data management infrastructures, but also in the subsequent development of strategies, resources and approaches. With mappings established to *CARDIO*'s thirty focal areas users have been able to straightforwardly interpret risks and associated causal and remedial factors.

Full evaluation case studies are not particularly widely available within the preservation and data management context, but brief reports from the Center for Research Libraries' *Certification of Digital Archives* and *Certification and Assessment of Digital Repositories projects* [CRL, 2012a] were published via its website. These projects included assessments of Portico (on two separate occasions); the Inter-University Consortium for Political and Social Research; the *LOCKSS* distributed archiving system; and *HathiTrust* at the University of Michigan. Aligning the broad findings of these to *PORRO* reveals broader recommendations than issued within these brief reports.

The latter series of CRL audits align identified concerns with corresponding TRAC criteria, reflecting the methodology and process adopted. Portico was audited twice in 2006 and 2010 as part of both CRL projects. The most recent evaluation concluded that the reposi-

tory had some shortcomings in its succession planning (TRAC A1.2); its definition of roles, responsibilities and job descriptions (A2.2); its policy documentation (several policies suffered from inconsistencies and contradictions) (A3.2); availability of documentation (A3.6); its definitions of "understandability" or "usability" of preserved content (B2.10); its systems for auditing collections and determining completeness (B2.12); its procedures for software and hardware upgrade (A3.6 and C1.10); and its capacity to deliver content (owing primarily to its status as a dark archive) (C2.2).

HathiTrust's 2011 audit describes concerns associated with a lack of succession planning (A1.2), ambiguity over content and system ownership and control (A3.3, A3.7, A4.3); and unclear quality assurance standards (A3.8, B1.1, B1.7, B1.8, B2.4), which is particularly critical since the repository aims to aggregate content from third party libraries.

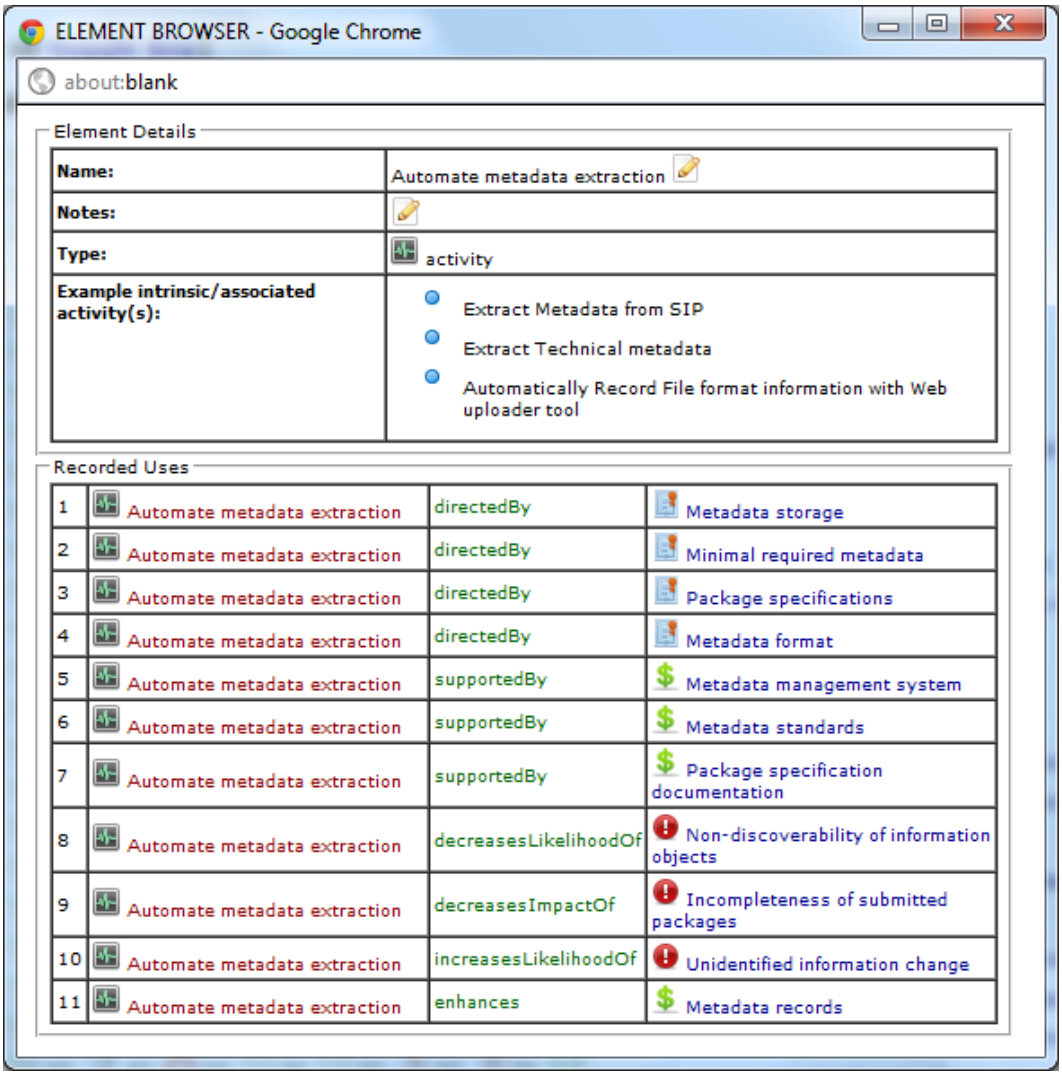
The *LOCKSS* distributed archiving system was criticized in a corresponding audit report for lacking means to determine when content which should be ingested was being withheld by depositing publishers and once more for a lack of succession planning. While CRL did not align these evaluations explicitly with individual criteria within *TRAC*, it is possible to do so via with the relationships within *PORRO*. Doing so reveals thirteen relevant criteria on the first point (implying a wide range of related considerations), and only criterion A1.2 on the issue of succession planning.

ICPSR was criticized in the report documenting its evaluation for lacking strict controls in the execution of its media migration policy, succession agreements or partnerships, an explicit policy for preservation, a traceable collection history, a policy for documenting system changes, explicit assurances of preservation rights and for failing to transfer copyright on ingest and monitor physical access by master key users.

PORRO has been retrospectively used in association with these reports to simulate the ways in which the ontology may support the evaluation process. The first conclusion is that the ontology supports the straightforward mapping of these findings to its implicit information elements, most obviously via the corresponding *TRAC* criteria, but also with ease via corresponding activities, resources, policies or rights whether by virtue of their existence or omission. Its added value is the straightforward revelation of associated issues and additional risk exposure. Taking the issue of succession planning which consistently appears we can look to one of two corresponding objectives within *PORRO*, "Establish relationships with succession partners" or "Establish appropriate strategies for facilitating succession of organization or content". Traversing the ontology we see that the "Succession arrangement" parameter which helps characterize the first objective in turn directs the activity "Establish succession arrangements". While this appears perfectly intuitive it is at this point that we begin to attain greater insights as illustrated in Figure 5.1, where we see that this activity may be supported with resources such as "Membership of partners' network". We also reveal

additional motivations for pursuing this activity, since it can limit the impact of risks such as "Loss of mandate", "Budgetary reduction" or "Enforced cessation of repository activities".

Figure 5.1: PORRO Ontology Browser



Clearly this example is simple, but illustrative of the value of a related network of elements one may more intuitively use these links to explore from a starting point of risk exposure to find the elements best suited to their resolution. The ontology is more useful still when used in tandem with *TRAC*. The generic *TRAC* criteria are accompanied by examples of evidence, intended to illustrate what must be demonstrable to achieve conformity. *PORRO* not only reflects this example evidence (which typically amounts to types of documentation within which evidence might be found), it exceeds it, with details of all the associated contextual arrangements which may indicate the criteria's satisfaction. *TRAC* criterion B1.2 (a random selection) is entitled "Repository clearly specifies the information that needs to be associated with digital material at the time of its deposit (i.e., SIP)". Suggested example evidence for auditors includes transfer requirements and producer-archive agreements. This is mapped

to *PORRO* via the goal E029_Define_Ingest_Package_Specification and provides means for interpretation of *TRAC*'s frequently misunderstood provisions.

Table 5.1: Goal: Define Ingest Package Specification

Define ingest package specification	
characterisedBy	Policy on relationship between ingest, archival and dissemination packages
characterisedBy	Minimal required metadata
characterisedBy	Package specifications
characterisedBy	Metadata creation responsibility
characterisedBy	Metadata creation workflow
legitimisedBy	Has prescribed minimal metadata requirements
threatenedBy	Extent of what is within the archival object is unclear
threatenedBy	Shortcomings in semantic or technical understandability of information
threatenedBy	Archival information cannot be traced to a received package
threatenedBy	Loss of authenticity of information
threatenedBy	Incompleteness of submitted packages
threatenedBy	Structural non-validity or malformedness of received packages
threatenedBy	Destruction of primary documentation
threatenedBy	Loss of information provenance

In terms of its ability to accommodate diverse information facets and support the expression of myriad associations, *PORRO* is successful. *PORRO*'s content is navigable, relatable and intuitive, capable of illustrating close and distant relationships between various system and information components. Even in isolation, the ontology enables digital libraries to reference the encapsulated knowledge in order to support their own risk assessment and preservation planning exercises. Since the ontology is intended to present a holistic vision of managed risks one can determine risk exposure by reference to infrastructural components that are lacking in an example institution, or focus on risks threatening vital provisions in priority areas. In the context of risk management, respondents' confidence in their perceived organizational maturity would be challenged by exposure to possible risks (with real world precedent) that may pose threats. For example, if respondents consider elements of their legal infrastructure to be very mature they can traverse a small number of relationships to confront possible risk scenarios concerning IPR infringement, Freedom of Information liabilities or contractual breach. If satisfied that these risks are adequately countered they will have greater faith in their assertion. Conversely, the process may prompt an awareness of shortcomings that were not previously well understood. The ontology is expected to scale

to reflect the very latest perspectives in preservation decision making, and with additional population, to present further insights.

5.4 Methodology

Our evaluation methodology for *PORRO* has two primary components. The first is a scrutiny of existing assertions of preservation capability and their comparison with *PORRO*. The evaluations are limited to those undertaken by ostensibly trusted organisations and therefore we focus on those institutions awarded the *Data Seal of Approval* [DSA, 2012]. Each submission presented in support of an application for a *DSA* is available publicly and structured in a fashion that facilitates comparison. To simplify, the correspondence between *DSA* and *PORRO* entities is determined by mapping via the *TRAC* criteria. Individual responses provided to satisfy individual criteria are in turn analysed and their alignment with the *PORRO* criteria assessed.

Secondly, we completed a further series of institutional assessments using *DRAMBORA* and *CARDIO* tool and refer to their results to in order to reveal evidence of *PORRO*'s value, as well as further insights into its applicability to a range of audit contexts.

5.4.1 Comparison with Other Metrics

The first phase of evaluation is principally intended to reveal the top-down utility of a *PORRO* supported repository evaluation. We focus on the popularly deployed *Data Seal of Approval* [Harmsen and de Leeuw, 2010] and explore the extent to which *PORRO* is capable of reflecting and supporting its criteria and methodology.

The *Data Seal of Approval* comprises of sixteen individual guidelines that collectively describe a conforming organisation - one that is demonstrably capable of providing digital preservation services. We take twenty-two successful awards of the Data Seal (granted between 2010 and 2015) and employ a combination of document research, textual content analysis and participant observation to map evidence of conformity with properties encoded within the *PORRO* ontology. By doing so we demonstrate the completeness of *PORRO* and its ability to map to existing criteria and provide further granularity of meaning to those high level, broadly expressed guidelines.

5.4.2 Comparison with Organisational Typologies

The second phase of evaluation is intended to illustrate *PORRO*'s success in providing an adaptable framework for evaluation based on the individual priorities of a given institution

or set of institutions. We focus on digital libraries and research data management systems, distinct types within an overall spectrum of digital custodial organisations.

Employing case study, textual analysis and participant observation we established risk profiles for digital libraries. We utilised the semantic structure of *PORRO* to undertake a series of assessments in a set of international digital libraries, identifying and aligning *PORRO* with the common objective, activities and challenges.

Secondly, we operated case studies in two research data management environments, London School of Economics and Political Sciences and Queen Mary, University of London. Using the *CARDIO* tool and the *PORRO* semantic framework we demonstrate organisational capacity to manage data and generated agreed intervention recommendations by relating identified issues with ontology elements, and ultimately with corresponding risk mitigation steps.

5.5 Evaluation Participants

Organisations awarded the *Data Seal of Approval* to date are as follows:

- 3TU.Datacentrum (2010)
- Archaeology Data Service (2014-2015)
- BABS - Long Term Preservation at the Bavarian State Library- Library Archiving and Access System (2010)
- Banco de Informacin para la Investigacin Aplicada en Ciencias Sociales (BIIACS) (2014-2015)
- BAS CLARIN (2010)
- CLARIN-D Resource Center Leipzig (2010)
- CLARIND-UDS (2010)
- DANS: Electronic Archiving SYstem (EASY) (2014-2015)
- Deutsches Textarchiv (2010)
- German National Library/ Deutsche Nationalbibliothek (DNB) (2010)
- HZSK Repository (2010)
- IDS Repository (2010)
- IMS Repository (2010)

- Inter-university Consortium for Political and Social Research (ICPSR) (2010)
- LASA (2010)
- LISS panel data (2010
- Odum Institute Data Archive (2014-2015)
- Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) (2010)
- Platform for Archiving CINES (PAC) (2010)
- The Language Archive - Max Planck Institute for Psycholinguistics (2010)
- Tbingen CLARIN-D Repository (2010)
- UK Data Archive (2010)

Participants in the Case Study sections were London School of Economics and Political Science, The *Michigan-Google Digitization Project* and *MBooks* at the University of Michigan Library, *Gallica* at the Bibliothque nationale de France, the Digital Library of the National Library of Sweden and CERN's *Document Server*.

5.6 Results Against Certification Process

5.6.1 Overview of Mapping Between PORRO and DSA

As described in a Chapter 2, the *Data Seal of Approval* (DSA) [Harmsen and de Leeuw, 2010] is a set of requirements and certification developed in the Netherlands by the Dutch Data Archiving and Networked Services (DANS) [DANS, 2012]. The *DSA* comprises sixteen broadly expressed guidelines that despite their brevity aim to cover questions of preservation capacity as inclusively as detailed standards such as *TRAC* and *ISO 16363*. In fact, the criteria extend beyond the archive's responsibilities to also encapsulate required commitments of data producers and data consumers. Nevertheless, for practical purposes it is the archive that is considered the "primary implementer" of the guidelines, and should assume responsibility for verifying and demonstrating evidence of the other actors' commitment and capacity. To that end, a mapping between *DSA* and *PORRO* is feasible and intuitive.

The mappings below were produced to illustrate the applicability of *PORRO* to a range of certification contexts. Each *DSA* guideline is mapped to corresponding *PORRO* goals which

can in turn be traced to corresponding parameters, resources, activities and risks. The mapping incorporates some redundancy but there are notably sixteen *PORRO* goals considered not explicitly mappable to the *DSA* criteria, as follows:

- E105_Maintain_Business_Planning_Autonomy
- E046_Establish_Appropriate_Business_Planning
- E048_Establish_Appropriate_Contingency_Funding
- E057_Establish_Appropriate_Strategies_For_Facilitating_Succession_Of_Organisation_Or_Content
- E091_Establish_Relationships_With_Succession_Partners
- E063_Establish_Assurances_Of_Sufficiency_Of_Staff_Skills_And_Capacity
- E051_Establish_Appropriate_Financial_Accounting_Infrastructure
- E064_Establish_Assurances_That_All_Costs_Are_And_Will_Continue_To_Be_Covered
- E066_Establish_Budgetary_Protection_Assurances
- E104_Maintain_Budget_Carry-Over_Facility
- E106_Maintain_Comprehensive_Costings_Breakdown
- E049_Establish_Appropriate_Coordination_And_Steering_Platform
- E096_Evaluate_And_Certify_Activities
- E047_Establish_Appropriate_Categories_Of_Staff
- E065_Establish_Budget_Dedicated_To_Training_Provision
- E087_Establish_Portfolio_Of_Internal_Or_External_Staff_Training_Provisions

These mainly correspond to aspects of organisational sustainability and staffing but their omission (or rather lack of explicit inclusion) is problematic. Technical aspects of preservation are only as effective as the organisation that oversees them. The *Data Seal of Approval* would be more compelling if organisations were required to demonstrate that even if their organisation faces risks of continuity an appropriate succession arrangements have been made. There are no aspects of *DSA* that cannot be comfortably accommodated within *PORRO*'s set of goals.

Each successful application for the *Data Seal of Approval* is documented on a corresponding website where the evidence of satisfaction of each criteria is reproduced [DSA, 2012]. This information is itself a useful reference resource for subsequent applicants seeking the seal of approval for their own institution. It is argued that by illustrating *PORRO*'s encapsulation of applicants' evidence, PORRO's own applicability as a tool to support wider evaluation is validated.

A mapping between each of the *Data Seal*'s sixteen guidelines follows below. For three of the guidelines (numbers two, seven and ten), those that correspond most closely to the

major preservation functions of ingest, preservation and dissemination, a more comprehensive mapping is provided. The corresponding justificatory texts submitted in support of each of the existing twenty two data seals awarded to date have been collated and parsed, with elements mapped to individual, related *PORRO* elements. That is, rather than simply including a correspondence between high level goals, the *PORRO* resource, parameter and right/responsibility elements that define and serve to accomplish these goals are included. This validates the mappings and illustrates the expressiveness of the ontology.

Guideline 1: The data producer deposits the data in a data repository with sufficient information for others to assess the quality of the data, and compliance with disciplinary and ethical norms.

- E043_Establish_And_Maintain_Terms_Of_Deposit
- E108_Maintain_Depositor_Dialogue
- E112_Make_Explicit_Preservation_Responsibility
- E122_Physically_Acquire_Content
- E041_Establish_And_Exercise_Ingest_Policy
- E042_Establish_And_Exercise_Selection_Policy
- E086_Establish_Policy_Transparency
- E069_Establish_Criteria_For_Data_Review
- E128_Select_And_Appraise_Ingested_Content
- E030_Authenticate_Source_Of_Ingested_Packages
- E116_Monitor_And_Fulfil_Freedom_Of_Information_Responsibilities
- E118_Monitor_And_Fulfil_Other_Legislative_And_Legal_Responsibilities
- E117_Monitor_And_Fulfil_Ipr_Responsibilities
- E035_Define_Ingest_Package_Specification
- E110_Maintain_Link_Between_Data_And_Metadata
- E124_Process_Ingested_Content
- E132_Verify_Ingest_Package_Conformity_With_Specification
- E076_Establish_List_Of_Supported_Formats
- E080_Establish_Means_For_Data_Review

Guideline 2: The data producer provides the data in formats recommended by the data repository.

This first fuller mapping illustrates the complexity of the challenge of information ingest, providing a means to identify comprehensiveness of provisions. Collectively, the *Data Seal*

applications reflect the full range of interactions associated with this aspect of deposit. Notable again though is a failure to relate the functional aspects of the repository with its operational or adminstrative aspects. The *Data Seal* application process appears to encourage applicants to disregard organisational aspects. *PORRO*'s linked approach is validated with these explicit relationships, which makes its mappings to *TRAC* for instance (which is similarly holistic in its approach) more balanced.

x - E043_Establish_And_Maintain_Terms_Of_Deposit
x - - P03inv_is_characterised_by E138_Compliance_Responsibility
x - - - P06inv_is_evidenced_by E457_Deposit_Agreement
x - - - P06inv_is_evidenced_by E542_Mandate_Definition
x - - - P04_directs E344_Exchange_Transfer_Documentation
x - - - P04_directs E355_Log_Accessions
– - - - P04_directs E388_Negotiate_Data_Management_Mandate
x - - P03inv_is_characterised_by E166_Exemptions_To_Preservation_Responsibility
x - - - P06inv_is_evidenced_by E542_Mandate_Definition
x - - - P04_directs E330_Engage_In_Dialogue_With_Stakeholder
x - - P03inv_is_characterised_by E203_Policy_Governing_Withdrawal_Of_Data_Management_Responsibility
x - - - P06inv_is_evidenced_by E457_Deposit_Agreement
x - - - P04_directs E330_Engage_In_Dialogue_With_Stakeholder
x - - P03inv_is_characterised_by E201_Policy_For_Negotiation_Of_Preservation_Responsibility
– - - - P04_directs E303_Accept_Data_Management_Responsibility
x - - - P04_directs E330_Engage_In_Dialogue_With_Stakeholder
x - - P03inv_is_characterised_by E238_Rights_And_Ownership_Definitions
x - - - P06inv_is_evidenced_by E457_Deposit_Agreement
x - - - P04_directs E344_Exchange_Transfer_Documentation
x - - - P04_directs E355_Log_Accessions
x - - P03inv_is_characterised_by E244_Selection
x - - - P06inv_is_evidenced_by E491_Acquisition_Tracking_System
x - - - P06inv_is_evidenced_by E452_Content_Processing_Forms
x - - - P04_directs E304_Aggregate_Data_Referenced_By_Or_Contextual_To_Dataset
– - - - P04_directs E320_Dispose_Of_Non-Ingested_Content
x - - - P04_directs E404_Refuse_Content_Ingest
x - - - P04_directs E409_Request_Data_Deposit
x - - - P04_directs E411_Retrieve_Content
x - - P03inv_is_characterised_by E255_Supported_Acquisition_Methods
x - - - P06inv_is_evidenced_by E504_Content_Processing_System
x - - - P06inv_is_evidenced_by E505_Content_Retriever
x - - - P04_directs E304_Aggregate_Data_Referenced_By_Or_Contextual_To_Dataset

– - - - P04_directs E318_Digitise_Analogue_Content
x - - - P04_directs E320_Dispose_Of_Non-Ingested_Content
x - - - P04_directs E389_Notify_Data_Originator_Of_Data_Receipt
x - - - P04_directs E411_Retrieve_Content
– - - P02inv_is_legitimised_by E274_Data_Management_Objectives_Consistent_With_Parent
x - - P02inv_is_legitimised_by E275_Data_Management_Responsibility
x - - P02inv_is_legitimised_by E276_Data_Management_Rights
– - - P02inv_is_legitimised_by E282_Has_Mandate_To_Aggregate_Published_Data
x - E100_Initiate_Stakeholder_Dialogue
x - - P03inv_is_characterised_by E203_Policy_Governing_Withdrawal_Of_Data_Management_Responsibility
x - - - P06inv_is_evidenced_by E457_Deposit_Agreement
x - - - P04_directs E330_Engage_In_Dialogue_With_Stakeholder
x - - P03inv_is_characterised_by E201_Policy_For_Negotiation_Of_Preservation_Responsibility
– - - - P04_directs E303_Accept_Data_Management_Responsibility
x - - - P04_directs E330_Engage_In_Dialogue_With_Stakeholder
x - - P02inv_is_legitimised_by E275_Data_Management_Responsibility
x - - P02inv_is_legitimised_by E276_Data_Management_Rights
x - E108_Maintain_Depositor_Dialogue
x - - P03inv_is_characterised_by E203_Policy_Governing_Withdrawal_Of_Data_Management_Responsibility
x - - - P06inv_is_evidenced_by E457_Deposit_Agreement
x - - - P04_directs E330_Engage_In_Dialogue_With_Stakeholder
x - - P03inv_is_characterised_by E201_Policy_For_Negotiation_Of_Preservation_Responsibility
x - - - P04_directs E303_Accept_Data_Management_Responsibility
x - - - P04_directs E330_Engage_In_Dialogue_With_Stakeholder
x - - P02inv_is_legitimised_by E275_Data_Management_Responsibility
x - - P02inv_is_legitimised_by E276_Data_Management_Rights
x - E112_Make_Explicit_Preservation_Responsibility
x - - P03inv_is_characterised_by E166_Exemptions_To_Preservation_Responsibility
x - - - P06inv_is_evidenced_by E542_Mandate_Definition
x - - - P04_directs E330_Engage_In_Dialogue_With_Stakeholder
x - - P03inv_is_characterised_by E203_Policy_Governing_Withdrawal_Of_Data_Management_Responsibility
x - - - P06inv_is_evidenced_by E457_Deposit_Agreement
x - - - P04_directs E330_Engage_In_Dialogue_With_Stakeholder
x - - P03inv_is_characterised_by E201_Policy_For_Negotiation_Of_Preservation_Responsibility
x - - - P04_directs E303_Accept_Data_Management_Responsibility
x - - - P04_directs E330_Engage_In_Dialogue_With_Stakeholder
x - - P02inv_is_legitimised_by E275_Data_Management_Responsibility
x - E029_Adopt_Appropriate_Preservation_Formats

x - - P03inv_is_characterised_by E155_Data_Representation
x - - - P06inv_is_evidenced_by E559_Preservation_Plan
x - - - P06inv_is_evidenced_by E471_Package_Relationship_Documentation
x - - - P04_directs E333_Establish_Preservation_Plan
x - - - P04_directs E334_Establish_Referential_Integrity
x - - - P04_directs E348_Identify_Data_Properties
x - - - P04_directs E354_Link_Preserved_Content_With_Original
x - - - P04_directs E430_Verify_Characteristics_Of_Data
x - - P03inv_is_characterised_by E222_Preservation_Level_Implications
x - - - P06inv_is_evidenced_by E559_Preservation_Plan
x - - - P06inv_is_evidenced_by E474_Preservation_Policy
x - - - P06inv_is_evidenced_by E462_Format_Documentation
x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation
– - - - P06inv_is_evidenced_by E551_Obsolescence_Metric
– - - - P04_directs E333_Establish_Preservation_Plan
– - - - P04_directs E402_Reference_External_Sources_During_Data_Management_Planning
x - - - P04_directs E339_Evaluate_Format_And_Media_Risk
x - - - P04_directs E372_Manage_Format_And_Media_Support
x - - - P04_directs E376_Migrate_Format_Or_Media
x - - P03inv_is_characterised_by E224_Preservation_Package_Structure
x - - - P06inv_is_evidenced_by E559_Preservation_Plan
x - - - P06inv_is_evidenced_by E471_Package_Relationship_Documentation
x - - - P04_directs E333_Establish_Preservation_Plan
x - - - P04_directs E334_Establish_Referential_Integrity
x - - - P04_directs E348_Identify_Data_Properties
– - - - P04_directs E402_Reference_External_Sources_During_Data_Management_Planning
x - - P03inv_is_characterised_by E227_Preservation_Strategy
x - - - P06inv_is_evidenced_by E559_Preservation_Plan
x - - - P06inv_is_evidenced_by E474_Preservation_Policy
x - - - P04_directs E333_Establish_Preservation_Plan
x - - - P04_directs E345_Execute_Preservation_Plan
– - - - P04_directs E402_Reference_External_Sources_During_Data_Management_Planning
x - - P03inv_is_characterised_by E256_Supported_Dissemination_Formats
– - - - P06inv_is_evidenced_by E488_Access_Platform
– - - - P06inv_is_evidenced_by E527_Format_Support
x - - - P06inv_is_evidenced_by E560_Preservation_Platform
x - - - P06inv_is_evidenced_by E462_Format_Documentation
x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation

– - - - P04_directs E359_Maintain_Access_Platform

– - - - P04_directs E360_Maintain_Administration_Platform

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - - P04_directs E431_Verify_Data_Formats

x - - P03inv_is_characterised_by E257_Supported_Ingest_Formats

x - - - P06inv_is_evidenced_by E527_Format_Support

x - - - P06inv_is_evidenced_by E535_Ingest_Platform

x - - - P06inv_is_evidenced_by E545_Media_Support

x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation

– - - - P04_directs E360_Maintain_Administration_Platform

x - - - P04_directs E366_Maintain_Ingest_Platform

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - - P04_directs E431_Verify_Data_Formats

x - - P03inv_is_characterised_by E258_Supported_Preservation_Formats

x - - - P06inv_is_evidenced_by E527_Format_Support

x - - - P06inv_is_evidenced_by E545_Media_Support

x - - - P06inv_is_evidenced_by E560_Preservation_Platform

x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation

– - - - P04_directs E360_Maintain_Administration_Platform

x - - - P04_directs E368_Maintain_Preservation_Platform

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - - P04_directs E431_Verify_Data_Formats

x - - P03inv_is_characterised_by E168_Format_Migration

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - - P04_directs E431_Verify_Data_Formats

x - - P03inv_is_characterised_by E188_Obsolescence_Risk_Tolerance

x - - - P06inv_is_evidenced_by E462_Format_Documentation

– - - - P06inv_is_evidenced_by E551_Obsolescence_Metric

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - P03inv_is_characterised_by E239_Risk_Assessment_Validation

x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation

– - - - P06inv_is_evidenced_by E551_Obsolescence_Metric

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - P02inv_is_legitimised_by E290_Has_Preservation_Policy_Discretion

x - - P02inv_is_legitimised_by E291_Has_Preservation_Responsibility

x - - P02inv_is_legitimised_by E292_Has_Preservation_Rights

x - E035_Define_Ingest_Package_Specification

x - - P03inv_is_characterised_by E214_Policy_On_Relationship_Between_Ingest_And_Archival_And_Dissemination_Packages

x - - - P06inv_is_evidenced_by E559_Preservation_Plan

x - - - P06inv_is_evidenced_by E558_Preservation_Management_System

x - - - P06inv_is_evidenced_by E471_Package_Relationship_Documentation

x - - - P06inv_is_evidenced_by E514_Data_Transformation_Plans

x - - - P06inv_is_evidenced_by E472_Package_Specification_Documentation

x - - - P04_directs E334_Establish_Referential_Integrity

x - - - P04_directs E354_Link_Preserved_Content_With_Original

x - - - P04_directs E357_Log_Object_Lifecycle

x - - - P04_directs E306_Assign_A_Processing_Record_To_Data

x - - - P04_directs E322_Document_Interactions_Surrounding_Dataset

x - - - P04_directs E323_Document_Package_Content

x - - - P04_directs E324_Document_Package_Structure

x - - - P04_directs E399_Record_Media_Movement

– - - - P04_directs E379_Monitor_Data_Citations_And_Reuse

x - - - P04_directs E380_Monitor_Dataset_Usage

x - - - P04_directs E312_Define_Package_Specifications

x - - - P04_directs E396_Publish_Package_Specifications

x - - P03inv_is_characterised_by E187_Minimal_Required_Metadata

x - - - P06inv_is_evidenced_by E472_Package_Specification_Documentation

x - - - P06inv_is_evidenced_by E467_Metadata_Creation_Guidelines

x - - - P06inv_is_evidenced_by E469_Metadata_Schema

x - - - P04_directs E323_Document_Package_Content

x - - - P04_directs E324_Document_Package_Structure

– - - - P04_directs E308_Automate_Metadata_Extraction

x - - - P04_directs E312_Define_Package_Specifications

x - - - P04_directs E415_Review_Metadata

x - - - P04_directs E310_Create_Object_Metadata

x - - - P04_directs E311_Create_Package_Descriptor

x - - P03inv_is_characterised_by E190_Package_Specifications

x - - - P06inv_is_evidenced_by E472_Package_Specification_Documentation

– - - - P04_directs E308_Automate_Metadata_Extraction

x - - - P04_directs E312_Define_Package_Specifications

x - - - P04_directs E353_Link_Metadata_To_Corresponding_Data

x - - - P04_directs E396_Publish_Package_Specifications

x - - P03inv_is_characterised_by E182_Metadata_Creation_Responsibility

x - - - P06inv_is_evidenced_by E467_Metadata_Creation_Guidelines

x - - - P04_directs E323_Document_Package_Content

x - - - P04_directs E324_Document_Package_Structure

x - - - P04_directs E312_Define_Package_Specifications

x - - - P04_directs E310_Create_Object_Metadata

x - - - P04_directs E311_Create_Package_Descriptor

x - - P03inv_is_characterised_by E183_Metadata_Creation_Workflow

x - - - P06inv_is_evidenced_by E467_Metadata_Creation_Guidelines

x - - - P04_directs E323_Document_Package_Content

x - - - P04_directs E324_Document_Package_Structure

x - - - P04_directs E312_Define_Package_Specifications

x - - - P04_directs E310_Create_Object_Metadata

x - - - P04_directs E311_Create_Package_Descriptor

x - - P02inv_is_legitimised_by E289_Has_Prescribed_Minimal_Metadata_Requirements

x - E132_Verify_Ingest_Package_Conformity_With_Specification

x - - P03inv_is_characterised_by E272_Validation_Checks_And_Requirements

x - - - P06inv_is_evidenced_by E588_Validation_System

x - - - P04_directs E428_Validate_Content

x - - - P04_directs E347_Generate_Fixity_Information

– - - - P04_directs E418_Scan_For_Viruses

x - - - P04_directs E429_Validate_Media_And_Storage

x - - P03inv_is_characterised_by E173_Ingest_Specification

x - - - P06inv_is_evidenced_by E471_Package_Relationship_Documentation

x - - - P06inv_is_evidenced_by E472_Package_Specification_Documentation

x - - - P04_directs E428_Validate_Content

x - - - P04_directs E373_Manage_Package_Specifications

x - - P03inv_is_characterised_by E250_Specification_Relationships

x - - - P06inv_is_evidenced_by E471_Package_Relationship_Documentation
x - - - P04_directs E428_Validate_Content
x - - - P04_directs E373_Manage_Package_Specifications
x - E076_Establish_List_Of_Supported_Formats
x - - P03inv_is_characterised_by E256_Supported_Dissemination_Formats
– - - - P06inv_is_evidenced_by E488_Access_Platform
– - - - P06inv_is_evidenced_by E527_Format_Support
x - - - P06inv_is_evidenced_by E560_Preservation_Platform
x - - - P06inv_is_evidenced_by E462_Format_Documentation
x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation
– - - - P04_directs E359_Maintain_Access_Platform
– - - - P04_directs E360_Maintain_Administration_Platform
x - - - P04_directs E339_Evaluate_Format_And_Media_Risk
x - - - P04_directs E372_Manage_Format_And_Media_Support
x - - - P04_directs E376_Migrate_Format_Or_Media
x - - - P04_directs E431_Verify_Data_Formats
x - - P03inv_is_characterised_by E257_Supported_Ingest_Formats
x - - - P06inv_is_evidenced_by E527_Format_Support
x - - - P06inv_is_evidenced_by E535_Ingest_Platform
x - - - P06inv_is_evidenced_by E545_Media_Support
x - - - P06inv_is_evidenced_by E462_Format_Documentation
x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation
– - - - P04_directs E360_Maintain_Administration_Platform
x - - - P04_directs E366_Maintain_Ingest_Platform
x - - - P04_directs E339_Evaluate_Format_And_Media_Risk
x - - - P04_directs E372_Manage_Format_And_Media_Support
x - - - P04_directs E376_Migrate_Format_Or_Media
x - - - P04_directs E431_Verify_Data_Formats
x - - P03inv_is_characterised_by E258_Supported_Preservation_Formats
x - - - P06inv_is_evidenced_by E527_Format_Support
x - - - P06inv_is_evidenced_by E545_Media_Support
x - - - P06inv_is_evidenced_by E560_Preservation_Platform
x - - - P06inv_is_evidenced_by E462_Format_Documentation
x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation
– - - - P04_directs E360_Maintain_Administration_Platform
x - - - P04_directs E368_Maintain_Preservation_Platform
x - - - P04_directs E339_Evaluate_Format_And_Media_Risk
x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - - P04_directs E431_Verify_Data_Formats

x - - P03inv_is_characterised_by E168_Format_Migration

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - - P04_directs E431_Verify_Data_Formats

x - - P03inv_is_characterised_by E188_Obsolescence_Risk_Tolerance

x - - - P06inv_is_evidenced_by E462_Format_Documentation

– - - - P06inv_is_evidenced_by E551_Obsolescence_Metric

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - P03inv_is_characterised_by E239_Risk_Assessment_Validation

x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation

– - - - P06inv_is_evidenced_by E551_Obsolescence_Metric

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

Guideline 3: The data producer provides the data together with the metadata requested by the data repository.

- E043_Establish_And_Maintain_Terms_Of_Deposit
- E100_Initiate_Stakeholder_Dialogue
- E108_Maintain_Depositor_Dialogue
- E112_Make_Explicit_Preservation_Responsibility
- E035_Define_Ingest_Package_Specification
- E037_Document_Archival_Data
- E110_Maintain_Link_Between_Data_And_Metadata
- E125_Record_And_Maintain_Descriptive_Metadata
- E126_Record_And_Maintain_Representation_Information
- E127_Record_Appropriate_Metadata
- E132_Verify_Ingest_Package_Conformity_With_Specification

Guideline 4: The data repository has an explicit mission in the area of digital archiving and promulgates it.

- E100_Initiate_Stakeholder_Dialogue
- E088_Establish_Ratification_Of_Preservation_Mission_From_Parent_Or_Governing_Entity
- E108_Maintain_Depositor_Dialogue
- E109_Maintain_End_User_Dialogue

Guideline 5: The data repository uses due diligence to ensure compliance with legal regulations and contracts including, when applicable, regulations governing the protection of human subjects.

- E039_Ensure_Appropriate_Contractual_Management
- E116_Monitor_And_Fulfil_Freedom_Of_Information_Responsibilities
- E118_Monitor_And_Fulfil_Other_Legislative_And_Legal_Responsibilities
- E117_Monitor_And_Fulfil_Ipr_Responsibilities
- E111_Maintain_Risk_Awareness
- E103_Maintain_Best_Practice_Awareness

Guideline 6: The data repository applies documented processes and procedures for managing data storage.

- E034_Define_Disaster_Recovery_Policy
- E036_Define_Policy_And_Procedures_For_Undertaking_Backups
- E041_Establish_And_Exercise_Ingest_Policy
- E042_Establish_And_Exercise_Selection_Policy
- E073_Establish_Hardware_Upgrade_Policy
- E074_Establish_Information_Security_Policy
- E082_Establish_Media_Refreshment_Policy
- E086_Establish_Policy_Transparency
- E092_Establish_Software_Upgrade_Policy
- E085_Establish_Policy_Review_Policy
- E031_Backup_Documentation
- E058_Establish_Appropriate_Technical_Documentation_Base
- E038_Document_Software_Dependencies

Guideline 7: The data repository has a plan for long-term preservation of its digital assets.

As seen below in this fuller mapping, the ontology corresponds very closely with the *Data Seal* applications. Notable is *PORRO*'s explicit association between rights issues and preservation approaches. No Data Seal application describes the rights implications of their preservation approaches within its response to the requirements of Guideline 7. This is illustrative of PORRO's more comprehensive approach (as indicated too with its coverage of issues of organisational infrastructure and sustainability which are not addressed with the Data Seal of Approval.

Despite *DSA*'s obvious and quite explicit focus on preservation several applicants provided less evidence to support satisfaction of this guideline than any other (if only in terms of number of words). *PORRO*'s preservation entities and associated relationships are demonstrably fuller.

x - E112_Make_Explicit_Preservation_Responsibility
x - - P03inv_is_characterised_by E166_Exemptions_To_Preservation_Responsibility
x - - - P06inv_is_evidenced_by E542_Mandate_Definition
x - - - P04_directs E330_Engage_In_Dialogue_With_Stakeholder
x - - P03inv_is_characterised_by E203_Policy_Governing_Withdrawal_Of_Data_Management_Responsibility
x - - - P06inv_is_evidenced_by E457_Deposit_Agreement
x - - - P04_directs E330_Engage_In_Dialogue_With_Stakeholder
x - - P03inv_is_characterised_by E201_Policy_For_Negotiation_Of_Preservation_Responsibility
x - - - P04_directs E303_Accept_Data_Management_Responsibility
x - - - P04_directs E330_Engage_In_Dialogue_With_Stakeholder
x - - P02inv_is_legitimised_by E275_Data_Management_Responsibility
x - E029_Adopt_Appropriate_Preservation_Formats
x - - P03inv_is_characterised_by E155_Data_Representation
x - - - P06inv_is_evidenced_by E559_Preservation_Plan
x - - - P06inv_is_evidenced_by E471_Package_Relationship_Documentation
x - - - P04_directs E333_Establish_Preservation_Plan
x - - - P04_directs E334_Establish_Referential_Integrity
x - - - P04_directs E348_Identify_Data_Properties
x - - - P04_directs E354_Link_Preserved_Content_With_Original
x - - - P04_directs E430_Verify_Characteristics_Of_Data
x - - P03inv_is_characterised_by E222_Preservation_Level_Implications
x - - - P06inv_is_evidenced_by E559_Preservation_Plan
x - - - P06inv_is_evidenced_by E474_Preservation_Policy
x - - - P06inv_is_evidenced_by E462_Format_Documentation
x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation
x - - - P06inv_is_evidenced_by E551_Obsolescence_Metric

x - - - P04_directs E333_Establish_Preservation_Plan

– - - - P04_directs E402_Reference_External_Sources_During_Data_Management_Planning

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - P03inv_is_characterised_by E224_Preservation_Package_Structure

x - - - P06inv_is_evidenced_by E559_Preservation_Plan

x - - - P06inv_is_evidenced_by E471_Package_Relationship_Documentation

x - - - P04_directs E333_Establish_Preservation_Plan

x - - - P04_directs E334_Establish_Referential_Integrity

x - - - P04_directs E348_Identify_Data_Properties

– - - - P04_directs E402_Reference_External_Sources_During_Data_Management_Planning

x - - P03inv_is_characterised_by E227_Preservation_Strategy

x - - - P06inv_is_evidenced_by E559_Preservation_Plan

x - - - P06inv_is_evidenced_by E474_Preservation_Policy

x - - - P04_directs E333_Establish_Preservation_Plan

x - - - P04_directs E345_Execute_Preservation_Plan

– - - - P04_directs E402_Reference_External_Sources_During_Data_Management_Planning

x - - P03inv_is_characterised_by E256_Supported_Dissemination_Formats

– - - - P06inv_is_evidenced_by E488_Access_Platform

x - - - P06inv_is_evidenced_by E527_Format_Support

x - - - P06inv_is_evidenced_by E560_Preservation_Platform

x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation

– - - - P04_directs E359_Maintain_Access_Platform

– - - - P04_directs E360_Maintain_Administration_Platform

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - - P04_directs E431_Verify_Data_Formats

x - - P03inv_is_characterised_by E257_Supported_Ingest_Formats

x - - - P06inv_is_evidenced_by E527_Format_Support

x - - - P06inv_is_evidenced_by E535_Ingest_Platform

– - - - P06inv_is_evidenced_by E545_Media_Support

x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation

– - - - P04_directs E360_Maintain_Administration_Platform

– - - - P04_directs E366_Maintain_Ingest_Platform

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - - P04_directs E431_Verify_Data_Formats

x - - P03inv_is_characterised_by E258_Supported_Preservation_Formats

x - - - P06inv_is_evidenced_by E527_Format_Support

x - - - P06inv_is_evidenced_by E545_Media_Support

x - - - P06inv_is_evidenced_by E560_Preservation_Platform

x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation

– - - - P04_directs E360_Maintain_Administration_Platform

x - - - P04_directs E368_Maintain_Preservation_Platform

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - - P04_directs E431_Verify_Data_Formats

x - - P03inv_is_characterised_by E168_Format_Migration

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - - P04_directs E431_Verify_Data_Formats

x - - P03inv_is_characterised_by E188_Obsolescence_Risk_Tolerance

x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E551_Obsolescence_Metric

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - P03inv_is_characterised_by E239_Risk_Assessment_Validation

x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation

x - - - P06inv_is_evidenced_by E551_Obsolescence_Metric

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - P02inv_is_legitimised_by E290_Has_Preservation_Policy_Discretion

x - - P02inv_is_legitimised_by E291_Has_Preservation_Responsibility

x - - P02inv_is_legitimised_by E292_Has_Preservation_Rights

x - E075_Establish_Levels_Of_Preservation

x - - P03inv_is_characterised_by E222_Preservation_Level_Implications

x - - - P06inv_is_evidenced_by E559_Preservation_Plan

x - - - P06inv_is_evidenced_by E474_Preservation_Policy

x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation

x - - - P06inv_is_evidenced_by E551_Obsolescence_Metric

x - - - P04_directs E333_Establish_Preservation_Plan

– - - - P04_directs E402_Reference_External_Sources_During_Data_Management_Planning

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - P03inv_is_characterised_by E227_Preservation_Strategy

x - - - P06inv_is_evidenced_by E559_Preservation_Plan

x - - - P06inv_is_evidenced_by E474_Preservation_Policy

x - - - P04_directs E333_Establish_Preservation_Plan

x - - - P04_directs E345_Execute_Preservation_Plan

– - - - P04_directs E402_Reference_External_Sources_During_Data_Management_Planning

x - - P03inv_is_characterised_by E221_Preservation_Level_Assignment

x - - - P06inv_is_evidenced_by E559_Preservation_Plan

x - - - P06inv_is_evidenced_by E474_Preservation_Policy

x - - - P04_directs E345_Execute_Preservation_Plan

– - - - P04_directs E402_Reference_External_Sources_During_Data_Management_Planning

x - - P03inv_is_characterised_by E244_Selection

– - - - P06inv_is_evidenced_by E491_Acquisition_Tracking_System

– - - - P06inv_is_evidenced_by E452_Content_Processing_Forms

x - - - P04_directs E304_Aggregate_Data_Referenced_By_Or_Contextual_To_Dataset

– - - - P04_directs E320_Dispose_Of_Non-Ingested_Content

– - - - P04_directs E404_Refuse_Content_Ingest

– - - - P04_directs E409_Request_Data_Deposit

– - - - P04_directs E411_Retrieve_Content

x - - P02inv_is_legitimised_by E290_Has_Preservation_Policy_Discretion

x - - P02inv_is_legitimised_by E291_Has_Preservation_Responsibility

x - - P02inv_is_legitimised_by E292_Has_Preservation_Rights

x - - P02inv_is_legitimised_by E282_Has_Mandate_To_Aggregate_Published_Data

x - E081_Establish_Means_To_Track_Data_Object_Through_Preservation_Workflow_And_Lifecycle

x - - P03inv_is_characterised_by E214_Policy_On_Relationship_Between_Ingest_And...

x - - - P06inv_is_evidenced_by E559_Preservation_Plan

x - - - P06inv_is_evidenced_by E558_Preservation_Management_System

x - - - P06inv_is_evidenced_by E471_Package_Relationship_Documentation

x - - - P06inv_is_evidenced_by E514_Data_Transformation_Plans

x - - - P06inv_is_evidenced_by E472_Package_Specification_Documentation

x - - - P04_directs E334_Establish_Referential_Integrity

x - - - P04_directs E354_Link_Preserved_Content_With_Original

– - - - P04_directs E357_Log_Object_Lifecycle

x - - - P04_directs E306_Assign_A_Processing_Record_To_Data

x - - - P04_directs E322_Document_Interactions_Surrounding_Dataset

x - - - P04_directs E323_Document_Package_Content

x - - - P04_directs E324_Document_Package_Structure

– - - - P04_directs E399_Record_Media_Movement

– - - - P04_directs E379_Monitor_Data_Citations_And_Reuse

– - - - P04_directs E380_Monitor_Dataset_Usage

x - - - P04_directs E312_Define_Package_Specifications

x - - - P04_directs E396_Publish_Package_Specifications

x - - P03inv_is_characterised_by E140_Content_Change

x - - - P06inv_is_evidenced_by E453_Custodial_History_Record

x - - - P06inv_is_evidenced_by E475_Processing_Record

x - - - P04_directs E306_Assign_A_Processing_Record_To_Data

x - - - P04_directs E322_Document_Interactions_Surrounding_Dataset

x - - - P04_directs E323_Document_Package_Content

x - - - P04_directs E324_Document_Package_Structure

– - - - P04_directs E399_Record_Media_Movement

x - - P03inv_is_characterised_by E143_Content_Removal_And_Deletion

– - - - P06inv_is_evidenced_by E491_Acquisition_Tracking_System

x - - - P06inv_is_evidenced_by E453_Custodial_History_Record

x - - - P06inv_is_evidenced_by E475_Processing_Record

x - - - P04_directs E306_Assign_A_Processing_Record_To_Data

x - - - P04_directs E322_Document_Interactions_Surrounding_Dataset

x - - - P04_directs E323_Document_Package_Content

x - - - P04_directs E324_Document_Package_Structure

– - - - P04_directs E399_Record_Media_Movement

x - - P03inv_is_characterised_by E156_Data_Review

x - - - P06inv_is_evidenced_by E453_Custodial_History_Record

x - - - P04_directs E306_Assign_A_Processing_Record_To_Data

x - - - P04_directs E307_Audit_Collections_And_Procedures

x - - P03inv_is_characterised_by E164_Documentation_Review

x - - - P06inv_is_evidenced_by E453_Custodial_History_Record

x - - - P06inv_is_evidenced_by E468_Metadata_Records

x - - - P04_directs E306_Assign_A_Processing_Record_To_Data

x - - - P04_directs E307_Audit_Collections_And_Procedures

x - - - P04_directs E323_Document_Package_Content

x - - - P04_directs E390_Perform_Metadata_Format_Conversion

x - - - P04_directs E415_Review_Metadata

x - - P03inv_is_characterised_by E230_Process_And_Infrastructure_Review

x - - - P06inv_is_evidenced_by E475_Processing_Record

x - - - P04_directs E307_Audit_Collections_And_Procedures

– - - - P04_directs E400_Record_System_Changes

x - - P02inv_is_legitimised_by E290_Has_Preservation_Policy_Discretion

x - - P02inv_is_legitimised_by E291_Has_Preservation_Responsibility

x - - P02inv_is_legitimised_by E292_Has_Preservation_Rights

– - - P02inv_is_legitimised_by E301_Sufficiency_And_Suitability_Of_Audit_Practice

x - E097_Exercise_Preservation_Plans

x - - P03inv_is_characterised_by E155_Data_Representation

x - - - P06inv_is_evidenced_by E559_Preservation_Plan

x - - - P06inv_is_evidenced_by E471_Package_Relationship_Documentation

x - - - P04_directs E333_Establish_Preservation_Plan

x - - - P04_directs E334_Establish_Referential_Integrity

x - - - P04_directs E348_Identify_Data_Properties

x - - - P04_directs E354_Link_Preserved_Content_With_Original

x - - - P04_directs E430_Verify_Characteristics_Of_Data

x - - P03inv_is_characterised_by E222_Preservation_Level_Implications

x - - - P06inv_is_evidenced_by E559_Preservation_Plan

x - - - P06inv_is_evidenced_by E474_Preservation_Policy

x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation

x - - - P06inv_is_evidenced_by E551_Obsolescence_Metric

x - - - P04_directs E333_Establish_Preservation_Plan

– - - - P04_directs E402_Reference_External_Sources_During_Data_Management_Planning

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - P03inv_is_characterised_by E224_Preservation_Package_Structure

x - - - P06inv_is_evidenced_by E559_Preservation_Plan

x - - - P06inv_is_evidenced_by E471_Package_Relationship_Documentation

x - - - P04_directs E333_Establish_Preservation_Plan

x - - - P04_directs E334_Establish_Referential_Integrity

x - - - P04_directs E348_Identify_Data_Properties

x - - - P04_directs E402_Reference_External_Sources_During_Data_Management_Planning

x - - P03inv_is_characterised_by E227_Preservation_Strategy

x - - - P06inv_is_evidenced_by E559_Preservation_Plan

x - - - P06inv_is_evidenced_by E474_Preservation_Policy

x - - - P04_directs E333_Establish_Preservation_Plan

x - - - P04_directs E345_Execute_Preservation_Plan

x - - - P04_directs E402_Reference_External_Sources_During_Data_Management_Planning

x - - P03inv_is_characterised_by E221_Preservation_Level_Assignment

x - - - P06inv_is_evidenced_by E559_Preservation_Plan

x - - - P06inv_is_evidenced_by E474_Preservation_Policy

x - - - P04_directs E345_Execute_Preservation_Plan

x - - - P04_directs E402_Reference_External_Sources_During_Data_Management_Planning

x - - P03inv_is_characterised_by E181_Media_Refreshment

x - - - P06inv_is_evidenced_by E586_Update_And_Upgrade_Prompts

– - - - P06inv_is_evidenced_by E479_System_Maintenance_And_Support_Agreement

x - - - P06inv_is_evidenced_by E462_Format_Documentation

– - - - P04_directs E403_Refresh_Media_Or_Hardware

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - P03inv_is_characterised_by E256_Supported_Dissemination_Formats

– - - - P06inv_is_evidenced_by E488_Access_Platform

x - - - P06inv_is_evidenced_by E527_Format_Support

x - - - P06inv_is_evidenced_by E560_Preservation_Platform

x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation

– - - - P04_directs E359_Maintain_Access_Platform

– - - - P04_directs E360_Maintain_Administration_Platform

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - - P04_directs E431_Verify_Data_Formats

x - - P03inv_is_characterised_by E257_Supported_Ingest_Formats

x - - - P06inv_is_evidenced_by E527_Format_Support

x - - - P06inv_is_evidenced_by E535_Ingest_Platform

– - - - P06inv_is_evidenced_by E545_Media_Support

x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation

– - - - P04_directs E360_Maintain_Administration_Platform

– - - - P04_directs E366_Maintain_Ingest_Platform

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - - P04_directs E431_Verify_Data_Formats

x - - P03inv_is_characterised_by E258_Supported_Preservation_Formats

x - - - P06inv_is_evidenced_by E527_Format_Support

x - - - P06inv_is_evidenced_by E545_Media_Support

x - - - P06inv_is_evidenced_by E560_Preservation_Platform

x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation

– - - - P04_directs E360_Maintain_Administration_Platform

x - - - P04_directs E368_Maintain_Preservation_Platform

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - - P04_directs E431_Verify_Data_Formats

x - - P03inv_is_characterised_by E168_Format_Migration

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - - P04_directs E431_Verify_Data_Formats

x - - P03inv_is_characterised_by E239_Risk_Assessment_Validation

x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation

x - - - P06inv_is_evidenced_by E551_Obsolescence_Metric

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - P02inv_is_legitimised_by E290_Has_Preservation_Policy_Discretion

x - - P02inv_is_legitimised_by E291_Has_Preservation_Responsibility

x - - P02inv_is_legitimised_by E292_Has_Preservation_Rights

x - E123_Plan_For_Preservation

x - - P03inv_is_characterised_by E222_Preservation_Level_Implications

x - - - P06inv_is_evidenced_by E559_Preservation_Plan

x - - - P06inv_is_evidenced_by E474_Preservation_Policy

x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation

x - - - P06inv_is_evidenced_by E551_Obsolescence_Metric

x - - - P04_directs E333_Establish_Preservation_Plan

– - - - P04_directs E402_Reference_External_Sources_During_Data_Management_Planning

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - P03inv_is_characterised_by E227_Preservation_Strategy

x - - - P06inv_is_evidenced_by E559_Preservation_Plan

x - - - P06inv_is_evidenced_by E474_Preservation_Policy

x - - - P04_directs E333_Establish_Preservation_Plan

x - - - P04_directs E345_Execute_Preservation_Plan

x - - - P04_directs E402_Reference_External_Sources_During_Data_Management_Planning

x - - P03inv_is_characterised_by E181_Media_Refreshment

x - - - P06inv_is_evidenced_by E586_Update_And_Upgrade_Prompts

– - - - P06inv_is_evidenced_by E479_System_Maintenance_And_Support_Agreement

x - - - P06inv_is_evidenced_by E462_Format_Documentation

– - - - P04_directs E403_Refresh_Media_Or_Hardware

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - P03inv_is_characterised_by E256_Supported_Dissemination_Formats

– - - - P06inv_is_evidenced_by E488_Access_Platform

x - - - P06inv_is_evidenced_by E527_Format_Support

x - - - P06inv_is_evidenced_by E560_Preservation_Platform

x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation

– - - - P04_directs E359_Maintain_Access_Platform

– - - - P04_directs E360_Maintain_Administration_Platform

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - - P04_directs E431_Verify_Data_Formats

x - - P03inv_is_characterised_by E257_Supported_Ingest_Formats

x - - - P06inv_is_evidenced_by E527_Format_Support

x - - - P06inv_is_evidenced_by E535_Ingest_Platform

– - - - P06inv_is_evidenced_by E545_Media_Support

x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation

– - - - P04_directs E360_Maintain_Administration_Platform

– - - - P04_directs E366_Maintain_Ingest_Platform
x - - - P04_directs E339_Evaluate_Format_And_Media_Risk
x - - - P04_directs E372_Manage_Format_And_Media_Support
x - - - P04_directs E376_Migrate_Format_Or_Media
x - - - P04_directs E431_Verify_Data_Formats
x - - P03inv_is_characterised_by E258_Supported_Preservation_Formats
x - - - P06inv_is_evidenced_by E527_Format_Support
x - - - P06inv_is_evidenced_by E545_Media_Support
x - - - P06inv_is_evidenced_by E560_Preservation_Platform
x - - - P06inv_is_evidenced_by E462_Format_Documentation
x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation
– - - - P04_directs E360_Maintain_Administration_Platform
x - - - P04_directs E368_Maintain_Preservation_Platform
x - - - P04_directs E339_Evaluate_Format_And_Media_Risk
x - - - P04_directs E372_Manage_Format_And_Media_Support
x - - - P04_directs E376_Migrate_Format_Or_Media
x - - - P04_directs E431_Verify_Data_Formats
x - - P03inv_is_characterised_by E168_Format_Migration
x - - - P04_directs E372_Manage_Format_And_Media_Support
x - - - P04_directs E376_Migrate_Format_Or_Media
x - - - P04_directs E431_Verify_Data_Formats
x - - P03inv_is_characterised_by E188_Obsolescence_Risk_Tolerance
x - - - P06inv_is_evidenced_by E462_Format_Documentation
x - - - P06inv_is_evidenced_by E551_Obsolescence_Metric
x - - - P04_directs E339_Evaluate_Format_And_Media_Risk
x - - - P04_directs E372_Manage_Format_And_Media_Support
x - - - P04_directs E376_Migrate_Format_Or_Media
x - - P03inv_is_characterised_by E239_Risk_Assessment_Validation
x - - - P06inv_is_evidenced_by E462_Format_Documentation
x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation
x - - - P06inv_is_evidenced_by E551_Obsolescence_Metric
x - - - P04_directs E339_Evaluate_Format_And_Media_Risk
x - - - P04_directs E372_Manage_Format_And_Media_Support
x - - - P04_directs E376_Migrate_Format_Or_Media
x - - P02inv_is_legitimised_by E290_Has_Preservation_Policy_Discretion
x - - P02inv_is_legitimised_by E291_Has_Preservation_Responsibility
x - - P02inv_is_legitimised_by E292_Has_Preservation_Rights
x - E129_Select_Preservation_Strategies

x - - P03inv_is_characterised_by E222_Preservation_Level_Implications
x - - - P06inv_is_evidenced_by E559_Preservation_Plan
x - - - P06inv_is_evidenced_by E474_Preservation_Policy
x - - - P06inv_is_evidenced_by E462_Format_Documentation
x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation
x - - - P06inv_is_evidenced_by E551_Obsolescence_Metric
x - - - P04_directs E333_Establish_Preservation_Plan
– - - - P04_directs E402_Reference_External_Sources_During_Data_Management_Planning
x - - - P04_directs E339_Evaluate_Format_And_Media_Risk
x - - - P04_directs E372_Manage_Format_And_Media_Support
x - - - P04_directs E376_Migrate_Format_Or_Media
x - - P03inv_is_characterised_by E227_Preservation_Strategy
x - - - P06inv_is_evidenced_by E559_Preservation_Plan
x - - - P06inv_is_evidenced_by E474_Preservation_Policy
x - - - P04_directs E333_Establish_Preservation_Plan
x - - - P04_directs E345_Execute_Preservation_Plan
x - - - P04_directs E402_Reference_External_Sources_During_Data_Management_Planning
x - - P03inv_is_characterised_by E228_Preservation_Validation
x - - - P06inv_is_evidenced_by E559_Preservation_Plan
x - - - P06inv_is_evidenced_by E474_Preservation_Policy
x - - - P06inv_is_evidenced_by E561_Preservation_Validation_System
x - - - P04_directs E342_Evaluate_Preservation_Plan
x - - - P04_directs E402_Reference_External_Sources_During_Data_Management_Planning
x - - - P04_directs E430_Verify_Characteristics_Of_Data
x - - P03inv_is_characterised_by E181_Media_Refreshment
x - - - P06inv_is_evidenced_by E586_Update_And_Upgrade_Prompts
– - - - P06inv_is_evidenced_by E479_System_Maintenance_And_Support_Agreement
x - - - P06inv_is_evidenced_by E462_Format_Documentation
– - - - P04_directs E403_Refresh_Media_Or_Hardware
x - - - P04_directs E372_Manage_Format_And_Media_Support
x - - - P04_directs E376_Migrate_Format_Or_Media
x - - P03inv_is_characterised_by E223_Preservation_Mechanism
x - - - P06inv_is_evidenced_by E527_Format_Support
x - - - P06inv_is_evidenced_by E582_Storage_Platform
x - - - P06inv_is_evidenced_by E545_Media_Support
x - - - P06inv_is_evidenced_by E560_Preservation_Platform
– - - - P04_directs E360_Maintain_Administration_Platform
– - - - P04_directs E364_Maintain_Backup_Platform

– - - - P04_directs E367_Maintain_Network_Protocol_Support
x - - - P04_directs E368_Maintain_Preservation_Platform
x - - P03inv_is_characterised_by E256_Supported_Dissemination_Formats
– - - - P06inv_is_evidenced_by E488_Access_Platform
x - - - P06inv_is_evidenced_by E527_Format_Support
x - - - P06inv_is_evidenced_by E560_Preservation_Platform
x - - - P06inv_is_evidenced_by E462_Format_Documentation
x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation
– - - - P04_directs E359_Maintain_Access_Platform
– - - - P04_directs E360_Maintain_Administration_Platform
x - - - P04_directs E339_Evaluate_Format_And_Media_Risk
x - - - P04_directs E372_Manage_Format_And_Media_Support
x - - - P04_directs E376_Migrate_Format_Or_Media
x - - - P04_directs E431_Verify_Data_Formats
x - - P03inv_is_characterised_by E257_Supported_Ingest_Formats
x - - - P06inv_is_evidenced_by E527_Format_Support
x - - - P06inv_is_evidenced_by E535_Ingest_Platform
– - - - P06inv_is_evidenced_by E545_Media_Support
x - - - P06inv_is_evidenced_by E462_Format_Documentation
x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation
– - - - P04_directs E360_Maintain_Administration_Platform
– - - - P04_directs E366_Maintain_Ingest_Platform
x - - - P04_directs E339_Evaluate_Format_And_Media_Risk
x - - - P04_directs E372_Manage_Format_And_Media_Support
x - - - P04_directs E376_Migrate_Format_Or_Media
x - - - P04_directs E431_Verify_Data_Formats
x - - P03inv_is_characterised_by E144_Content_Representation
x - - - P06inv_is_evidenced_by E535_Ingest_Platform
x - - - P06inv_is_evidenced_by E582_Storage_Platform
x - - - P06inv_is_evidenced_by E560_Preservation_Platform
– - - - P04_directs E360_Maintain_Administration_Platform
x - - - P04_directs E365_Maintain_Generic_And_Shared_Technology
x - - - P04_directs E366_Maintain_Ingest_Platform
x - - - P04_directs E368_Maintain_Preservation_Platform
x - - - P04_directs E371_Maintain_Storage_Platform
x - - P03inv_is_characterised_by E185_Metadata_Representation
x - - - P06inv_is_evidenced_by E527_Format_Support
– - - - P06inv_is_evidenced_by E492_Administration_Platform

x - - - P06inv_is_evidenced_by E535_Ingest_Platform

x - - - P06inv_is_evidenced_by E582_Storage_Platform

x - - - P06inv_is_evidenced_by E560_Preservation_Platform

– - - - P04_directs E360_Maintain_Administration_Platform

x - - P03inv_is_characterised_by E258_Supported_Preservation_Formats

x - - - P06inv_is_evidenced_by E527_Format_Support

x - - - P06inv_is_evidenced_by E545_Media_Support

x - - - P06inv_is_evidenced_by E560_Preservation_Platform

x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation

– - - - P04_directs E360_Maintain_Administration_Platform

x - - - P04_directs E368_Maintain_Preservation_Platform

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - - P04_directs E431_Verify_Data_Formats

x - - P03inv_is_characterised_by E168_Format_Migration

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - - P04_directs E431_Verify_Data_Formats

x - - P03inv_is_characterised_by E188_Obsolescence_Risk_Tolerance

x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E551_Obsolescence_Metric

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - P03inv_is_characterised_by E239_Risk_Assessment_Validation

x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation

x - - - P06inv_is_evidenced_by E551_Obsolescence_Metric

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - P02inv_is_legitimised_by E290_Has_Preservation_Policy_Discretion

x - - P02inv_is_legitimised_by E291_Has_Preservation_Responsibility

x - - P02inv_is_legitimised_by E292_Has_Preservation_Rights

x - E113_Make_Explicit_Preservation_Rights

– - - P03inv_is_characterised_by E153_Copyright_And_Access_Restrictions

– - - - P06inv_is_evidenced_by E511_Copyrighting_Mechanism
– - - - P04_directs E338_Evaluate_Data_Copyright_Status
– - - - P04_directs E378_Monitor_Copyright_Status
– - - P03inv_is_characterised_by E157_Data_Rights_Transfer
– - - - P06inv_is_evidenced_by E571_Rights_Database
– - - - P04_directs E338_Evaluate_Data_Copyright_Status
– - - - P04_directs E378_Monitor_Copyright_Status
– - - P03inv_is_characterised_by E195_Policy_Covering_Distribution_Of_Copyright_Material
– - - - P06inv_is_evidenced_by E571_Rights_Database
– - - - P04_directs E338_Evaluate_Data_Copyright_Status
– - - - P04_directs E378_Monitor_Copyright_Status
x - - P02inv_is_legitimised_by E283_Has_Mandate_To_Manage_And_Distribute_Copyright_Materials
– - - P02inv_is_legitimised_by E294_Has_Restrictions_On_Data_Management_Or_Distribution_Based_On_Copyright_Status
x - E032_Classify_Archival_Data
x - - P03inv_is_characterised_by E222_Preservation_Level_Implications
x - - - P06inv_is_evidenced_by E559_Preservation_Plan
x - - - P06inv_is_evidenced_by E474_Preservation_Policy
x - - - P06inv_is_evidenced_by E462_Format_Documentation
x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation
x - - - P06inv_is_evidenced_by E551_Obsolescence_Metric
x - - - P04_directs E333_Establish_Preservation_Plan
– - - - P04_directs E402_Reference_External_Sources_During_Data_Management_Planning
x - - - P04_directs E339_Evaluate_Format_And_Media_Risk
x - - - P04_directs E372_Manage_Format_And_Media_Support
x - - - P04_directs E376_Migrate_Format_Or_Media
x - - P03inv_is_characterised_by E239_Risk_Assessment_Validation
x - - - P06inv_is_evidenced_by E462_Format_Documentation
x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation
x - - - P06inv_is_evidenced_by E551_Obsolescence_Metric
x - - - P04_directs E339_Evaluate_Format_And_Media_Risk
x - - - P04_directs E372_Manage_Format_And_Media_Support
– - - - P04_directs E376_Migrate_Format_Or_Media
x - E120_Monitor_File_Format_Obsolescence
x - - P03inv_is_characterised_by E256_Supported_Dissemination_Formats
– - - - P06inv_is_evidenced_by E488_Access_Platform
x - - - P06inv_is_evidenced_by E527_Format_Support
x - - - P06inv_is_evidenced_by E560_Preservation_Platform
x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation
– - - - P04_directs E359_Maintain_Access_Platform
– - - - P04_directs E360_Maintain_Administration_Platform
x - - - P04_directs E339_Evaluate_Format_And_Media_Risk
x - - - P04_directs E372_Manage_Format_And_Media_Support
x - - - P04_directs E376_Migrate_Format_Or_Media
x - - - P04_directs E431_Verify_Data_Formats
x - - P03inv_is_characterised_by E257_Supported_Ingest_Formats
x - - - P06inv_is_evidenced_by E527_Format_Support
x - - - P06inv_is_evidenced_by E535_Ingest_Platform
– - - - P06inv_is_evidenced_by E545_Media_Support
x - - - P06inv_is_evidenced_by E462_Format_Documentation
x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation
– - - - P04_directs E360_Maintain_Administration_Platform
– - - - P04_directs E366_Maintain_Ingest_Platform
x - - - P04_directs E339_Evaluate_Format_And_Media_Risk
x - - - P04_directs E372_Manage_Format_And_Media_Support
x - - - P04_directs E376_Migrate_Format_Or_Media
x - - - P04_directs E431_Verify_Data_Formats
x - - P03inv_is_characterised_by E258_Supported_Preservation_Formats
x - - - P06inv_is_evidenced_by E527_Format_Support
x - - - P06inv_is_evidenced_by E545_Media_Support
x - - - P06inv_is_evidenced_by E560_Preservation_Platform
x - - - P06inv_is_evidenced_by E462_Format_Documentation
x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation
– - - - P04_directs E360_Maintain_Administration_Platform
x - - - P04_directs E368_Maintain_Preservation_Platform
x - - - P04_directs E339_Evaluate_Format_And_Media_Risk
x - - - P04_directs E372_Manage_Format_And_Media_Support
x - - - P04_directs E376_Migrate_Format_Or_Media
x - - - P04_directs E431_Verify_Data_Formats
x - - P03inv_is_characterised_by E168_Format_Migration
x - - - P04_directs E372_Manage_Format_And_Media_Support
x - - - P04_directs E376_Migrate_Format_Or_Media
x - - - P04_directs E431_Verify_Data_Formats
x - - P03inv_is_characterised_by E188_Obsolescence_Risk_Tolerance
x - - - P06inv_is_evidenced_by E462_Format_Documentation
x - - - P06inv_is_evidenced_by E551_Obsolescence_Metric

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

x - - P03inv_is_characterised_by E239_Risk_Assessment_Validation

x - - - P06inv_is_evidenced_by E462_Format_Documentation

x - - - P06inv_is_evidenced_by E543_Means_For_Format_And_Media_Representation

x - - - P06inv_is_evidenced_by E551_Obsolescence_Metric

x - - - P04_directs E339_Evaluate_Format_And_Media_Risk

x - - - P04_directs E372_Manage_Format_And_Media_Support

x - - - P04_directs E376_Migrate_Format_Or_Media

Guideline 8: Archiving takes place according to explicit work flows across the data life cycle.

- E089_Establish_Relationship_Between_Access_And_Archival_Packages
- E070_Establish_Criteria_For_Disposal
- E081_Establish_Means_To_Track_Data_Object_Through_Preservation_Workflow_And_Lifecycle
- E035_Define_Ingest_Package_Specification
- E059_Establish_Archival_Packages_Configuration
- E102_Maintain_Archival_Package_Referential_Integrity
- E114_Manage_Formation_Of_Dissemination_Package
- E078_Establish_Means_For_Data_Disposal
- E040_Ensure_Synchronisation_Of_Data_Separated_By_Time_Or_Space
- E132_Verify_Ingest_Package_Conformity_With_Specification
- E095_Establish_Transformation_Procedure_From_Ingest_To_Archival_Packages
- E090_Establish_Relationship_Between_Ingest_And_Archival_Packages

Guideline 9: The data repository assumes responsibility from the data producers for access and availability of the digital objects.

- E071_Establish_Data_Ownership
- E043_Establish_And_Maintain_Terms_Of_Deposit
- E067_Establish_Conditions_For_Access
- E072_Establish_Designated_Community
- E084_Establish_Physical_And_Logical_Provisions_For_Access
- E098_Implement_Access_Controls
- E099_Implement_Categories_Of_Access

- E114_Manage_Formation_Of_Dissemination_Package

Guideline 10: The data repository enables the users to discover and use the data and refer to them in a persistent way.

The provision of permanent access once more sees broad correspondence between *PORRO* and the *DSA* applications.

x - E068_Establish_Criteria_For_Data_Identification
x - - P03inv_is_characterised_by E155_Data_Representation
x - - - P06inv_is_evidenced_by E559_Preservation_Plan
x - - - P06inv_is_evidenced_by E471_Package_Relationship_Documentation
x - - - P04_directs E333_Establish_Preservation_Plan
x - - - P04_directs E334_Establish_Referential_Integrity
x - - - P04_directs E348_Identify_Data_Properties
x - - - P04_directs E354_Link_Preserved_Content_With_Original
x - - - P04_directs E430_Verify_Characteristics_Of_Data
x - - P03inv_is_characterised_by E214_Policy_On_Relationship_Between_Ingest_And...
x - - - P06inv_is_evidenced_by E559_Preservation_Plan
x - - - P06inv_is_evidenced_by E558_Preservation_Management_System
x - - - P06inv_is_evidenced_by E471_Package_Relationship_Documentation
x - - - P06inv_is_evidenced_by E514_Data_Transformation_Plans
x - - - P06inv_is_evidenced_by E472_Package_Specification_Documentation
x - - - P04_directs E334_Establish_Referential_Integrity
x - - - P04_directs E354_Link_Preserved_Content_With_Original
x - - - P04_directs E357_Log_Object_Lifecycle
– - - - P04_directs E306_Assign_A_Processing_Record_To_Data
– - - - P04_directs E322_Document_Interactions_Surrounding_Dataset
x - - - P04_directs E323_Document_Package_Content
x - - - P04_directs E324_Document_Package_Structure
x - - - P04_directs E399_Record_Media_Movement
x - - - P04_directs E379_Monitor_Data_Citations_And_Reuse
x - - - P04_directs E380_Monitor_Dataset_Usage
x - - - P04_directs E312_Define_Package_Specifications
x - - - P04_directs E396_Publish_Package_Specifications
x - - P02inv_is_legitimised_by E290_Has_Preservation_Policy_Discretion
x - - P02inv_is_legitimised_by E291_Has_Preservation_Responsibility
x - - P02inv_is_legitimised_by E292_Has_Preservation_Rights
x - E079_Establish_Means_For_Data_Identification
x - - P03inv_is_characterised_by E186_Metadata_Storage

x - - - P06inv_is_evidenced_by E472_Package_Specification_Documentation

– - - - P04_directs E308_Automate_Metadata_Extraction

x - - - P04_directs E353_Link_Metadata_To_Corresponding_Data

x - - P03inv_is_characterised_by E190_Package_Specifications

x - - - P06inv_is_evidenced_by E472_Package_Specification_Documentation

x - - - P04_directs E308_Automate_Metadata_Extraction

x - - - P04_directs E312_Define_Package_Specifications

x - - - P04_directs E353_Link_Metadata_To_Corresponding_Data

x - - - P04_directs E396_Publish_Package_Specifications

x - - P03inv_is_characterised_by E184_Metadata_Format

x - - - P06inv_is_evidenced_by E470_Metadata_Standards

x - - - P06inv_is_evidenced_by E472_Package_Specification_Documentation

x - - - P06inv_is_evidenced_by E468_Metadata_Records

– - - - P04_directs E308_Automate_Metadata_Extraction

x - - - P04_directs E353_Link_Metadata_To_Corresponding_Data

x - - - P04_directs E390_Perform_Metadata_Format_Conversion

x - - P03inv_is_characterised_by E146_Content_Versioning

x - - - P06inv_is_evidenced_by E470_Metadata_Standards

x - - - P06inv_is_evidenced_by E472_Package_Specification_Documentation

x - - - P04_directs E312_Define_Package_Specifications

– - - - P04_directs E353_Link_Metadata_To_Corresponding_Data

x - - P03inv_is_characterised_by E182_Metadata_Creation_Responsibility

– - - - P06inv_is_evidenced_by E467_Metadata_Creation_Guidelines

x - - - P04_directs E323_Document_Package_Content

x - - - P04_directs E324_Document_Package_Structure

x - - - P04_directs E312_Define_Package_Specifications

x - - - P04_directs E310_Create_Object_Metadata

x - - - P04_directs E311_Create_Package_Descriptor

x - - P03inv_is_characterised_by E183_Metadata_Creation_Workflow

x - - - P06inv_is_evidenced_by E467_Metadata_Creation_Guidelines

x - - - P04_directs E323_Document_Package_Content

x - - - P04_directs E324_Document_Package_Structure

x - - - P04_directs E312_Define_Package_Specifications

x - - - P04_directs E310_Create_Object_Metadata

x - - - P04_directs E311_Create_Package_Descriptor

x - - P02inv_is_legitimised_by E289_Has_Prescribed_Minimal_Metadata_Requirements

x - E125_Record_And_Maintain_Descriptive_Metadata

x - - P03inv_is_characterised_by E186_Metadata_Storage

x - - - P06inv_is_evidenced_by E472_Package_Specification_Documentation

x - - - P04_directs E308_Automate_Metadata_Extraction

x - - - P04_directs E353_Link_Metadata_To_Corresponding_Data

x - - P03inv_is_characterised_by E187_Minimal_Required_Metadata

x - - - P06inv_is_evidenced_by E472_Package_Specification_Documentation

x - - - P06inv_is_evidenced_by E467_Metadata_Creation_Guidelines

x - - - P06inv_is_evidenced_by E469_Metadata_Schema

x - - - P04_directs E323_Document_Package_Content

x - - - P04_directs E324_Document_Package_Structure

– - - - P04_directs E308_Automate_Metadata_Extraction

x - - - P04_directs E312_Define_Package_Specifications

x - - - P04_directs E415_Review_Metadata

x - - - P04_directs E310_Create_Object_Metadata

x - - - P04_directs E311_Create_Package_Descriptor

x - - P03inv_is_characterised_by E184_Metadata_Format

x - - - P06inv_is_evidenced_by E470_Metadata_Standards

x - - - P06inv_is_evidenced_by E472_Package_Specification_Documentation

x - - - P06inv_is_evidenced_by E468_Metadata_Records

– - - - P04_directs E308_Automate_Metadata_Extraction

x - - - P04_directs E353_Link_Metadata_To_Corresponding_Data

x - - - P04_directs E390_Perform_Metadata_Format_Conversion

x - - P03inv_is_characterised_by E255_Supported_Acquisition_Methods

– - - - P06inv_is_evidenced_by E504_Content_Processing_System

– - - - P06inv_is_evidenced_by E505_Content_Retriever

– - - - P04_directs E304_Aggregate_Data_Referenced_By_Or_Contextual_To_Dataset

– - - - P04_directs E318_Digitise_Analogue_Content

– - - - P04_directs E320_Dispose_Of_Non-Ingested_Content

– - - - P04_directs E389_Notify_Data_Originator_Of_Data_Receipt

– - - - P04_directs E411_Retrieve_Content

x - - P03inv_is_characterised_by E235_Repository_Integration

x - - - P04_directs E304_Aggregate_Data_Referenced_By_Or_Contextual_To_Dataset

x - - P03inv_is_characterised_by E182_Metadata_Creation_Responsibility

x - - - P06inv_is_evidenced_by E467_Metadata_Creation_Guidelines

x - - - P04_directs E323_Document_Package_Content

x - - - P04_directs E324_Document_Package_Structure

x - - - P04_directs E312_Define_Package_Specifications

x - - - P04_directs E310_Create_Object_Metadata

x - - - P04_directs E311_Create_Package_Descriptor

– - - P03inv_is_characterised_by E183_Metadata_Creation_Workflow
x - - - P06inv_is_evidenced_by E467_Metadata_Creation_Guidelines
x - - - P04_directs E323_Document_Package_Content
x - - - P04_directs E324_Document_Package_Structure
x - - - P04_directs E312_Define_Package_Specifications
x - - - P04_directs E310_Create_Object_Metadata
x - - - P04_directs E311_Create_Package_Descriptor
x - - P02inv_is_legitimised_by E289_Has_Prescribed_Minimal_Metadata_Requirements
– - - P02inv_is_legitimised_by E282_Has_Mandate_To_Aggregate_Published_Data
x - E083_Establish_Naming_Convention
x - - P03inv_is_characterised_by E187_Minimal_Required_Metadata
x - - - P06inv_is_evidenced_by E472_Package_Specification_Documentation
x - - - P06inv_is_evidenced_by E467_Metadata_Creation_Guidelines
x - - - P06inv_is_evidenced_by E469_Metadata_Schema
x - - - P04_directs E323_Document_Package_Content
x - - - P04_directs E324_Document_Package_Structure
x - - - P04_directs E308_Automate_Metadata_Extraction
x - - - P04_directs E312_Define_Package_Specifications
x - - - P04_directs E415_Review_Metadata
x - - - P04_directs E310_Create_Object_Metadata
x - - - P04_directs E311_Create_Package_Descriptor
x - - P03inv_is_characterised_by E182_Metadata_Creation_Responsibility
x - - - P06inv_is_evidenced_by E467_Metadata_Creation_Guidelines
x - - - P04_directs E323_Document_Package_Content
x - - - P04_directs E324_Document_Package_Structure
x - - - P04_directs E312_Define_Package_Specifications
x - - - P04_directs E310_Create_Object_Metadata
x - - - P04_directs E311_Create_Package_Descriptor
x - - P03inv_is_characterised_by E183_Metadata_Creation_Workflow
x - - - P06inv_is_evidenced_by E467_Metadata_Creation_Guidelines
x - - - P04_directs E323_Document_Package_Content
x - - - P04_directs E324_Document_Package_Structure
x - - - P04_directs E312_Define_Package_Specifications
x - - - P04_directs E310_Create_Object_Metadata
x - - - P04_directs E311_Create_Package_Descriptor
x - - P02inv_is_legitimised_by E289_Has_Prescribed_Minimal_Metadata_Requirements

Guideline 11: The data repository ensures the integrity of the digital objects and the metadata.

- E036_Define_Policy_And_Procedures_For_Undertaking_Backups
- E101_Limit_Data_Loss_Incidence
- E061_Establish_Assurances_Of_Recoverability_Of_Any_Lost_Data
- E093_Establish_Suitability_Of_Backup_Infrastructure_Through_Testing
- E031_Backup_Documentation
- E102_Maintain_Archival_Package_Referential_Integrity
- E044_Establish_Appropriate_Backup_Redundancy_Provisions
- E045_Establish_Appropriate_Backup_Remoteness_Provisions
- E050_Establish_Appropriate_Database_Backup_Infrastructure
- E055_Establish_Appropriate_Provisions_For_Backup
- E033_Continuously_Validate_Data_Integrity
- E107_Maintain_Data_Integrity
- E130_Validate_Data_Integrity
- E131_Validate_Integrity_Of_Backups

Guideline 12: The data repository ensures the authenticity of the digital objects and the metadata.

- E029_Adopt_Appropriate_Preservation_Formats
- E081_Establish_Means_To_Track_Data_Object_Through_Preservation_Workflow_And_Lifecycle
- E123_Plan_For_Preservation
- E129_Select_Preservation_Strategies
- E030_Authenticate_Source_Of_Ingested_Packages
- E102_Maintain_Archival_Package_Referential_Integrity
- E110_Maintain_Link_Between_Data_And_Metadata
- E114_Manage_Formation_Of_Dissemination_Package

Guideline 13: The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS.

- E052_Establish_Appropriate_Hardware_Infrastructure
- E056_Establish_Appropriate_Software_Infrastructure
- E060_Establish_Assurances_Of_Availability_Of_Appropriate_Technical_Skills
- E077_Establish_Logical_Storage_Provisions
- E050_Establish_Appropriate_Database_Backup_Infrastructure
- E058_Establish_Appropriate_Technical_Documentation_Base

- E062_Establish_Assurances_Of_Site_Stability

Guideline 14: The data consumer complies with access regulations set by the data repository.

- E094_Establish_Terms_Of_Use
- E067_Establish_Conditions_For_Access
- E084_Establish_Physical_And_Logical_Provisions_For_Access
- E098_Implement_Access_Controls
- E099_Implement_Categories_Of_Access
- E115_Monitor_Access_Behaviours
- E121_Monitor_Unauthorised_Access
- E109_Maintain_End_User_Dialogue
- E054_Establish_Appropriate_Physical_Security_Provisions
- E053_Establish_Appropriate_Logical_Security_Provisions

Guideline 15: The data consumer conforms to and agrees with any codes of conduct that are generally accepted in the relevant sector for the exchange and proper use of knowledge and information.

- E039_Ensure_Appropriate_Contractual_Management
- E094_Establish_Terms_Of_Use
- E067_Establish_Conditions_For_Access
- E109_Maintain_End_User_Dialogue
- E119_Monitor_And_Respond_To_Designated_Community_Evolution
- E116_Monitor_And_Fulfil_Freedom_Of_Information_Responsibilities
- E118_Monitor_And_Fulfil_Other_Legislative_And_Legal_Responsibilities
- E117_Monitor_And_Fulfil_Ipr_Responsibilities

Guideline 16: The data consumer respects the applicable licences of the data repository regarding the use of the data.

- E039_Ensure_Appropriate_Contractual_Management
- E094_Establish_Terms_Of_Use
- E067_Establish_Conditions_For_Access
- E098_Implement_Access_Controls
- E099_Implement_Categories_Of_Access
- E119_Monitor_And_Respond_To_Designated_Community_Evolution
- E116_Monitor_And_Fulfil_Freedom_Of_Information_Responsibilities
- E118_Monitor_And_Fulfil_Other_Legislative_And_Legal_Responsibilities
- E117_Monitor_And_Fulfil_Ipr_Responsibilities

5.7 Results Against Evaluatory Deployments

5.7.1 DELOS Digital Library Audits

The purpose of the case study that follows is to compare the best practice on display at four mature, international digital libraries with the contents of the *PORRO* ontology and to verify its completeness and applicability to these contexts.

During November and December of 2007 the *DELOS* Projects Digital Preservation Cluster undertook a series of evaluative facilitated assessments of a series digital library infrastructures. From the perspective of this thesis, the studies principal objective was the validation of a common set of criteria that may be applied to digital libraries irrespective of their organisational spacing, scale or the specific characteristics of their collections. These criteria are realised as *PORRO*.

Four digital libraries participated in the pilot assessments, hand picked to reflect the diversity that exists within this highly active field. The Michigan-*Google Digitization Project* and *MBooks* at the University of Michigan Library, *Gallica* at the Bibliothque nationale de France, the Digital Library of the National Library of Sweden and CERNs *Document Server* exhibit a range of organisational and functional characteristics representative of most of that which is conceivable within a digital library context. The conclusions that followed each assessment would be distilled into a broadly applicable generic template, focussing not on diversity, but the fundamental commonalities that distinguish digital libraries.

Each assessment incorporated an onsite visit that took an average of three days, preceded by a lengthy period of dialogue and information exchange between project facilitators and institutional participants, and considerable desk-based research. *DRAMBORA* presented an explicit 6 step method for performing assessment and during the onsite activities this was conformed with closely; initial stages built towards the development of a comprehensive organisational profile which incorporated detailed and documented expressions of organisational purpose and process. Taking the organisation's mission or mandate as a starting point, a process of hierarchical analysis, investigation and expansion resulted in the formalised expression of organisational objectives, implicit activities, regulatory and technological influences and fundamental roles and responsibilities. This provided an input to the latter stages of risk identification, assessment and management, where threats posed to the organisational infrastructure and the continued delivery of services were defined and evaluated, with plans for their ongoing management formulated and prioritised.

The process was insightful and highlighted opportunities for improvement of the *DRAMBORA* methodology, as well as a range of generic objectives, functions and concerns common to digital libraries. A greater understanding of the practical ways in which organisations as-

sess their own risks was reached; a generic risk profile for digital libraries was established and provides a means to verify *PORRO*'s applicability.

Finally, from the perspective of each of the audited institutions the process was overwhelmingly successful; testimonials from representatives of each described in detail the benefits of formally scrutinizing the organisational characteristics and implicit challenges faced within their own digital library.

Case Study Summary Findings

A number of conclusions were drawn from these audits about the current state of digital libraries, capable of usefully informing an emerging and more general profile for these institutions. Determining the expressiveness of *PORRO* requires comparison with common objectives, activities and challenges being faced and embraced by leading custodial institutions. An initial assertion that was quickly affirmed was the presence, at the heart of digital libraries, of a digital repository.

Common Objectives A common core goal that was shared by each of the audited digital libraries was to facilitate access to digital materials. In general terms, much of the digital content overseen by the four participating digital libraries were derivatives of physical information assets. Digital formats and dissemination infrastructures enabled these organisations to reach an increasingly wide audience, and to more effectively pursue their stated mandates, which uniformly prioritised not just the collection but also the distribution of information.

As described at length above, preservation appeared to be a less high priority objective, at least in terms more familiar to digital archivists. Format stabilising is performed at ingest in most cases, but there appeared little commitment to more complex preservation measures. Only in anticipated service extensions, such as those mooted at *Gallica* and the University of Michigan, or where born-digital materials were being managed, as in the case of the CERN *Document Server*, did semantic information preservation appear to be occupying a more central focus.

More fundamental organisational objectives were similarly common. A pivotal part of the business operations of all four audit participants was revenue generation; across each institution the specific methods employed to secure resource varied. Notwithstanding the natural and predictable sentiment that more resource would be welcomed, none of the participating digital libraries appeared to be overly concerned about the availability of sufficient finances, and all had made reasonably robust provisions to ensure adequate sums would continue to be available to support their endeavours.

All libraries appeared to have shared objectives with regard to employing suitably qualified staff, although the extent to which the digital infrastructure was 'mainstreamed' within the overall library context varied. University of Michigan and BnF were explicitly committed to blurring the lines that separated digital and traditional physical materials, and this appeared to be a model that others were moving towards. By leveraging the skill sets of a range of library staff objectives could be more straightforwardly achieved, with minimal need to conceive and construct additional facilities and infrastructures where the challenges of managing digital and analogue materials were comparable.

All libraries were committed to the provision of secure technological infrastructures, and each had established robust facilities to support widespread dissemination of content while limiting threats to data integrity posed by remote or physical intrusions.

Common Constraints Relatively few generalisations could be made about the constraints affecting each of the participating digital libraries; a range of organisational context types is represented within the four, and each was subject to distinct legislative requirements. These often subjective constraints tended to add extra definition to the broadly stated aims and objectives that appeared to exhibit greater commonality. This is a validation of *PORRO*'s fundamental structure which relates individual goals to one or more parameters, enabling a single generic goal to be interpreted flexibly depending on context.

From a legal perspective the most pervasive influence appeared to be intellectual property law, most specifically copyright, which largely affected the terms within which content could be distributed. Even those organisations that seek to ingest public domain materials acknowledged and were required to account for associated responsibilities. Other laws provided more context-specific influence; for example, legal deposit laws varied substantially between jurisdictions and could have the effect of adding additional responsibilities or empowering libraries, legitimising and validating strategic intentions.

Technical constraints were extremely specific to individual organisations, and corresponded to available expertise and to existing technical provisions and infrastructures. The most obvious common areas are the standards with which each of the participants conformed with. MARC and METS were both relied upon by the majority of these digital libraries, and were likely to find a place in most mainstream environments. Image formats such as TIFF were frequently utilised to encode archival image content. Other standards commonly conformed to include web accessibility standards, which are, notwithstanding local variations and specificities, broadly international in scope, and must be satisfied by web based access systems. No standards appear to be consistently utilised for ingest, other than where particular formats, such as METS, are relied upon as wrappers for ingest packages.

Common Roles Roles within each digital library varied in terms of nomenclature, although sufficient similarities could be identified to derive a sense of those which are common. In general terms each digital library appeared to operate under several layers of management. Ultimate responsibility could be traced to the highest organisational tiers of the environment within which the digital library exists, or even beyond where additional umbrella accountabilities could be traced. Of greater interest were the more hands-on roles with responsibilities for the maintenance and running of the digital libraries. Each of the examples assessed in this study relied upon the efforts of a single digital library manager, with duties to oversee related efforts, facilitate communication with higher levels of management and ensure the organisationally suitability and viability in terms of staff allocations, and overall resource availability. Selecting content, facilitating creation or digitisation efforts, overseeing ingest, ensuring quality control, managing access (including verifying the legal status of digital content) and managing technological and information security infrastructures were other commonly identified roles. Needless to say, in different libraries the relationship between roles and individuals was not necessarily one to one; sometimes individual roles were occupied by several staff, and at other times a single staff member might have occupied multiple roles. No explicit preservation role was generally identified, although this appeared likely to evolve as ambitions increasingly encompass overt preservation objectives.

Common Activities and Assets Activities generally corresponded closely with the common roles identified above; the more functional activities amounted to a generic workflow for digital libraries, beginning with selection of content, followed by digitisation, ingest, quality control, and finally access provision. Numerous resources were relied upon to facilitate the satisfactory completion of each; intellectual assets included selection criteria, standard digitisation parameters, ingest and metadata schema (describing the structure of a submission package for example), quality control standards and access authorisation systems. More technologically-oriented assets included systems to support each of these stages, such as digitisation software and hardware, ingest systems, scripts to automate the process of quality control, and authentication and authorisation subsystems, each of which often relied on additional data held in separate databases. Human resources were of course necessary to support each of these stages; even where automation was possible, human interaction was generally required to perform or support validation or system monitoring. More granular aspects of activities and assets naturally tended to be more specialist to individual organisational contexts, and associated with specific characteristics or objectives. Organisational-oriented activities were similarly common; revenue generation, staff training, legal monitoring and policy development were all fundamental activities within each of the participating organisations.

Common Risks Some of the foremost risks that appeared to be almost completely generic are in fact not particularly unique to digital libraries. Threats to resource availability, organisational cohesion, retention of key skills and legal conformance face all organisations, irrespective of their business or the context or domain area within which they operate. The specifics of these challenges of course varied; within the small sample in the *DELOS* project for example a range of organisational settings were observed, each with practical consequences. There were difficulties in defining common risks for all organisations, even those that claimed common status as digital libraries, because the practical variances are potentially significant. Nevertheless, even a general understanding of applicable risks can be useful for digital libraries. These represent broadly daubed strokes, given meaningful definition on a canvas of risks by the addition of finer detail that relates to more subjective aspects of the implicit threats.

Process-oriented risks are most generically applicable and immediately meaningful within the digital library landscape. For example, threats relating to adequacy and completeness of metadata to facilitate ingest, preservation and, most notably, discovery are faced by all digital library infrastructures. The adoption of common library and other information interchange standards such as MARC and METS was evident throughout each of the audited institutions; systems generally demanded valid and well formed examples of metadata to function appropriately. Digital object acquisition carried similarly widely evident risks; a common one related to the quality assurance of digital materials and metadata created and ingested in often highly automated processes. For all of the participating digital libraries there were risks associated with dissemination of content. Perhaps the most universal was the threat of liability for breach of intellectual property law caused by circulating copyrighted materials. For some this was more dangerous, as digital collections included known copyright materials, with access privileges calculated at the point of dissemination. All the digital libraries acknowledged that there may be in-copyright material within their collections, even where the intention was to digitise and provide access to only public domain materials.

Digital preservation appeared to have been vocally embraced by each of the audited libraries, but there were evident risks associated with the current adopted approaches. In some respects the potential impacts of information loss were negligible, given that in most cases redigitisation would entirely alleviate the impact; within the participating organisations digital objects could seldomly be described as original, or non-reproducible. A bigger preservation problem related to a lack of appropriately comprehensive and formally defined policies. Indeed, this omission appears to be responsible for many of the most pertinent common risks faced by digital libraries. Each of the libraries involved in this study was well established, and functionally effective. But their maturity in terms of policy infrastructure was at times questionable. It was thought that in order to formalise their objectives, procedures, and indeed their very legitimacy, digital libraries must make extra efforts to document their policies

in a transparent fashion.

Digital Library Risk Profile Conclusions Applying risk analysis based auditing methodology to digital libraries identified both common strengths and weaknesses in their work. While digital libraries were highly efficient in automating the ingest of digitised content, and providing flexible access to their collections, the acquisition of born-digital content posed more difficult requirements that needed bespoke solutions and often semi-automatic processing. For metadata management and provision of access digital libraries relied on existing library standards and electronic catalogues that could be linked to simple storage solutions. Relying primarily on standard formats had introduced some complacency, exaggerated further because within such contexts digitised collections represent little more than access-facilitating surrogates of their analogue collections. The observed technical infrastructure was adequate and secure for the purposes of the digital library services.

The areas within which the audited digital libraries were identifiably falling short included:

- policy management and policy/procedure documentation and maintaining the knowledge-base of the organisation on the whole;
- creation and management of preservation metadata;
- documentation of systems in use and maintenance of audit trails of processing applied to digital objects in librarys care;
- stakeholder transparency and participation;
- management of assigning responsibility for preservation planning and effective preservation strategy building.

All participating libraries were in the process of expanding and changing their services, which would bring these weaknesses increasingly to the fore.

Case Study Conclusions

The most overwhelming response from the audited institutions was that the audit process yielded numerous benefits, and provided insights that would undoubtedly prompt further investigation and probable response. However, a general response that appeared to be consistent from each of the audited organisations was the value of the process would be lessened if the facilitators were not present. If organisations were incapable of exploring their own risks independently then the potential benefits of the process may not be fully exploitable.

This reaction may simply be a methodological consequence of the way these audits were undertaken. Generally speaking, facilitators elected to refrain from impressing upon library staff shortcomings that they regarded as self evident, instead preferring to lead them to their own independent realization via the various stages of the *DRAMBORA* process. At times this worked well but there were several opportunities where the *DELOS* facilitators shared their experiences of visiting other organisations; they were well positioned to comment on and compare systems to those in place elsewhere. This reveals the value of an external perspective in what is really an internal, reflective process, and is a critical validation of *PORRO*'s value. Individuals responsible for self assessment have an implicit understanding of their own organisations mandate, objectives and fundamental activities as a direct consequence of their personal and professional association. These can be systematically explored in *PORRO* to verify their completeness.

The value of *PORRO*'s implicit flexbility is similarly reflected in these conclusions. With the ontology model success can be verified by reference to specific aims and circumstances of a given repository. Only relevant constraints and contextual influences need even be considered, and significant additional information is available to reveal the meaning and interrelationships between functional and contextual building blocks. A tailored approach has unquestionable applicability and value - repositories within this case study exhibited considerable diversity (notwithstanding their common 'digital library' status) and a customisable approach to interrogation is therefore of value. This contrasts with more objective audit processes that might be criticised as either irrelevant or meaningless on account of their more generally applicable scope. Without a knowledge base, there is an implicit vulnerability in this approach, threatening the extent to which repositories can independently improve. With adequate resource to best practice documentation, customisable to their circumstances, self assessing repositories can only reasonably identify problems within the bounds of what they believe that they should be doing. A knowledge base illustrates the boundaries of best practice. It presents the precise implications of particular actions, and enables a user to feel reassured that they are doing everything necessary to accomplish a particular activity, support a particular resource or mitigate a particular risk. Those cases where organisations are oblivious to their shortcomings, or unaware of the available possibilities that they might usefully seize (which could happen using *DRAMBORA* in isolation) are dramatically reduced.

Such information can form the basis for repository profiles. Just as on can map existing criteria to *PORRO*, as has been done with the *CARDIO*, *TRAC* and *Data Seal of Approval* guidelines, one can do this with repository classes, geographic or legislative jurisdictions or strategic priorities. Core roles, responsibilities, functions and risks for a variety of repository types can be displayed.

Another value of *PORRO*'s evidence base is its linked nature, which combines roles, activities and responsibilities that may be wide ranging organisationally. All of the role holders

and individuals involved in the repositorys business must engage and be engaged with to ensure the success of an evaluation, because digital preservation, as its characterised, comprises so many different aspects, organisationally and technologically. In order to be of real value to the organisation, everyone with any relevant responsibilities or concerns ought to be involved and *PORRO* facilitates this by linking the at times diverse contributions into a single ontology. The audit process is in reality little more than a formalised means of facilitating dialogue and discussion between the stakeholders and *PORRO* models the lines across which this dialogue can meaningfully take place. In those organisations that did invest time and effort from every functional and organisational unit there were visible benefits, as everything from minor confusions to more long-standing concerns were raised, discussed and generally resolved. Communication on an organisation-wide basis is always acknowledged as vital, but all too often overlooked or underemphasised. The self-audit represents an invaluable opportunity to develop a shared and globally acceptable interpretation and understanding of overall strengths, weakness, opportunities and threats. This benefit should be more explicitly expressed within *DRAMBORA*.

Following the identification of risks, a considerable part of the time spent on site during the *DELOS* audits was committed to risk assessment; for each risk repository staff discussed the severity of the threat and provided impact and probability scores. The original *DRAMBORA* text adopts a fairly granular scale for both impact and probability, although during the assessments it was generally felt that this complexity presented unnecessary additional barriers to the process. Again, *PORRO* supports this process, by indicating not only where risks occur in terms of interactions, but also the practical steps or resource investments that can be made to mitigate or reduced their impact. In isolation a risk can be identified more straightforwardly than it can be quantified. Historical data can inform where risks are likely to occur multiple times, but preservation is typically littered with risks that may have no history of occurrence. Identifying which explicitly identified risk management measures are implemented reveals corresponding probability and potential impact. When *DRAMBORA* is utilised to support a self assessment process, its results are of most value for internal use - it seems likely that in isolation risks will be considered in terms of their relative severity against those already identified. *PORRO* is constructed 'bottom-up' from real world environments but its combination of a wide range of perspectives lends it an objective weight. This can inform risk assessment results that have considerably greater objective weight, and may then be the basis for a more global comparison, in the same way that deploying a consistent group of individuals to assess multiple organisations lends a global applicability to the results of each.

A vital commodity when describing risk is a means to determine, or express risk impact. It appeared that the perception of challenge associated with preservation within digital library contexts is quite distinct from that of those dealing with born digital or otherwise unique digital assets. In most cases within the audited institutions, the value of digital con-

tent was mainly surrogacy for physical assets. Libraries remain primarily access-focussed and digitised content is considerably more plentiful than born-digital materials. Preservation is naturally prioritised lower since, notwithstanding the significant cost of rescanning large quantities of content, anything that is lost can generally be digitised again. The original *DRAMBORA* text describes risk impact in terms of only loss of digital object authenticity and understandability. Initial concerns with this limited definition of impact were to some extent met with subsequent reference to the loss of organisations ability to ensure authenticity and understandability of their digital collections. However, the experiences of the *DELOS* assessments revealed that even this slightly broader definition was too narrow to be either universally usable or applicable. *PORRO* by contrast introduces a far greater granularity of impact, where the outcomes of individual risks can be traced to each explicitly threatened goal or resource. Also modelled are the potentially viral pathways risks can take where if a risk occurs another is made more likely or more destructive. Many valid risks could be only loosely related to digital holdings and the consequent loss of digital information, rendering any attempt to quantitatively express the extent of potential impact in such terms unfeasible. An objective risk impact scoring system that considers only one manifestation of success or failure is unnecessarily restrictive. *DRAMBORA* sought to extend the grammar of risk impact by enabling users to select from four classes of risk impact, which were 'Reputation and Intangibles', 'Organisational Viability', 'Service Delivery' and 'Technology', but this still renders risks disconnected and their wider implications unknown. Since *PORRO* enables the association of risk with any number of goals, risks or resources cause and effect can be much better understood within an organisational setting.

The four organisations that participated in this process were all in a state of transition. New services were being developed, expansions being planned to other areas, new contracts being signed and new responsibilities embraced as novel legislation emerged. In light of the almost constant development that characterises the repository and digital library community it becomes difficult to say at any particular moment whether a particular organisation satisfies a requirement to be trustworthy. This is especially true when assessment is based upon heavyweight, monolithic standards with significant associated audit costs. *PORRO*'s metric is much more focussed on facilitating improvement than on the imposition of transitory judgements. Its model is compatible with an ongoing process of maturity modelling. Concerned with not only validating the effectiveness of existing infrastructures, but also determining the suitability of proposed developments, the ontology effectively reflects the dynamic characteristics of the repository domain. In general terms, it is easier to isolate and accredit individual services that the repository is offering, irrespective of their maturity, and then make some conclusions about the organisation as a whole, aimed at its overall development. If the aim of the audit is simply to judge the entire organisation at once, any verdict will have to be accompanied by numerous caveats. This will not really assist those stakeholders concerned

about the sustainability or effectiveness of the repository in question, whereas a more general expression of maturity, structured according to available services and measured against mandate and objectives has considerably greater value.

5.7.2 CARDIO Evaluation

Introduction to the CARDIO Evaluations

As described above, the *Collaborative Assessment of Research Data Infrastructure and Objectives* (CARDIO) process and associated tool was developed by the Digital Curation Centre in 2012 to support evaluation of institutional provisions to support management, sharing and long term preservation of research data. Following its initial release, *PORRO* was introduced as an integral part of *CARDIO*, providing an intellectual context to support the provision of responses to its thirty individual sections. The purpose of the case study that follows is to illustrate the value of PORRO in contextualising and informing the process of institutional assessment for a wide range of user types and backgrounds. *CARDIO*/*PORRO* was the basis for institutional assessments undertaken as part of the Digital Curation Centre's programme of institutional engagements. These were largely prompted by institutional concern surrounding increasing demands from Research Councils UK regarding the sharing and management of research data. This case study focuses on the assessment that took place at the London School of Economics (LSE) in summer 2012.

Evaluation Summary (LSE)

As part of its programme of institutional engagements the Digital Curation Centre supported ongoing efforts at LSE to develop research data management capability and capacity. This assessment and subsequent report respectively sought to capture and document the status and perceptions across a range of areas influential to research data management. These were subdivided into matters of organisation, technology and resources.

Following the survey and programme of interviews described in this thesis first chapter, six LSE employees (including both researchers and support staff) completed a full-scale "*CARDIO*" assessment, considering 30 individual areas influential to research data management capacity / capability and scoring LSE provisions from 1 to 5. Their composition was intended to reflect the range of service providers and data-supported research methods undertaken within the institution. This provided useful insight about strengths and shortcomings of current research data management provisions.

A number of methods were used to administer the *CARDIO* workflow during the assessments, with lessons having been learned from earlier, internal pilot exercises conducted by

colleagues at the University of Bath and by the author during an assessment of Queen Mary University of London. Prior to the introduction of *PORRO*, *CARDIO*'s initial test user base had reported some criticism. *CARDIO* leads users through a process that requires them to evaluate the performance of their institution with respect to data management in thirty individual areas. Reflecting earlier work these areas ranged from organisational to technological aspects. The tool was deliberately succinct but according to feedback was prone to ambiguity. User testimonials revealed confusion from some users of the meaning of particular questions [Ball and Darlington, 2012]. *PORRO* was introduced to lend context and was demonstrably successful.

Agreed ratings across each individual area were provided in detail in the subsequent pages. Most LSE provisions were rated between 1 and 3 out of 5. The surveys illustrated the widespread opportunities for improvement, and prompted some thoughts on what actions should be prioritised in order to meet emerging research data management requirements.

The key recommendations were as follows:

1. Publish data policy that defines responsibilities and clarifies a data definition and ensure its widespread circulation, adoption and systematic review.

2. Identify and promote institutional and funder requirements for sharing.

3. Develop guidance for selecting and preparing data for long term accessibility (e.g. metadata, format choices, migration).

4. Encourage registration of LSE and LSE-licensed data in central systems and promote repository data storage functionality.

5. Promote implications and applicability of appropriate legislation through staff training and publicity.

6. Offer legal guidance, providing greater clarity on data ownership and IPR, especially with respect to reconfigured licensed datasets.

7. Re-evaluate service (e.g., IT service) portfolio to ensure that it meets researchers requirements in order to make a more compelling case for charging a proportion of research income. As part of this, ensure IT services are involved in project application process.

8. Promote risk awareness and the resourcing of mitigating measures via data management planning during grant application, project and post-project processes.

9. Offer explicit data management training for doctoral students / early career researchers.

10. Develop and promote experts directory to clarify who should be contacted with specific types of request/query.

5.7.3 Case Study Conclusions

At QMUL the *CARDIO* process took the form of an interactive workshop; participants reflected upon a starter set of institutional assessments provided by a small cross-section of representatives from University service and academic departments. A reference manual contextualised individual sections for participants. At the example assessment at LSE a cohort provided their responses in conversation with the author who was coordinating the assessment. At times the discussions labored as the author attempted to explain to participants what was being asked of them. Greater clarity was clearly offered with the introduction of contextualising *PORRO* mappings. Others used the online tool and had the opportunity to identify corresponding *PORRO* classes via the *CARDIO* reference manual [McHugh, 2011]. Both approaches were successful in further contextualising the process.

In addition, the provision of the linked *PORRO* entities was thought to offer an additional range of entry points to consider the respective capacity and capabilities in each area. Responses could be provoked by identification of a pertinent risk or an omitted, failing or demonstrably successful system characteristic.

Perhaps more significant was *PORRO*'s role in forecasting or planning future developments. *CARDIO* is partly about diagnosing issues and concerns but has most value in supporting the definition of plans to ensure their resolution. The recommendations above were wholly conceived using *PORRO*'s implicit relationships and both welcomed and endorsed by participating members of LSE's *Research Data Management Committee*.

CARDIO's workflow requires, following the submission of individual perspectives, the establishment of consensus between participants. The online tool provides a set of social functions (including persistent chat client and notification system) to support this but in the example assessments undertaken at Queen Mary University and London School of Economics physical interaction between respondents was preferred, in a workshop or meeting format. A common semantic model was shown to inform a coherent perspective and enable agreement to take place more straightforwardly, whereby participants were less concerned with understanding the parameters of the question than agreeing their response.

End users' collective perspectives are aggregated into a single institutional view. This immediately directs one towards disproportionately strong or weak areas. Improvement is a critical part of the *CARDIO* process, and therefore one can prioritise at a glance the areas of greatest concern. *PORRO*'s association enabled the identification of facets that can be linked to specific organisational policy, process and assets.

Chapter 6

Conclusion

The experiences of legacy research in preservation and data management risk awareness, coupled with analysis of preservation planning approaches reveal clear opportunities for richer knowledge management structures. Digital preservation is a pressing priority across sectors and disciplines posing many challenges that span conceptual boundaries. How we characterise and share best practice for safeguarding materials threatened by technical obsolescence, organisational failure and physical and logical degradation is pivotal. Success is contingent on a wide range of factors, diverse and variable individual priorities, emphases and contextual circumstances. Each can be tremendously influential in both presenting challenges to and facilitating preservation.

To establish common consensus on the implications of particular preservation choices and environments we must first understand the interrelationships that comprise the preservation context. We have presented PORRO as a means of enabling such expression. It supports the classification of preservation profiles, pockets of objective meaning in a necessarily subjective context. This in turn supports preservation decision making, and the establishment of greater collective awareness of risks implicit in classes of preservation approach and information.

6.1 Performance Against Research Objectives

At the heart of our research objectives is risk: success would establish the concept as a measurable and relatable means of classifying activities associated with the process of digital preservation. Our results complement existing prescriptive standards for preservation with a set of structured criteria and corresponding methodology that reflect a widespread appetite for responsive, adaptable and flexible instruments to support preservation evaluation. We have conceived and built a unique knowledge base of preservation practice based on accounts

directly presented in the course of preservation audits and observed via the *DRAMBORA Interactive* online tool and a further series of facilitated self-assessments.

It is evident, not least through research undertaken at LSE (referenced in Chapter 2) and other externally conducted research such as the surveys undertaken by the Digital Curation Centre [DCC, 2015] that preservation capacity is an ambition shared by practitioners within commercial, cultural heritage and academic sectors. From each community there is an evident perception that while there is value in defining objective metrics to assess the realisation of such goals there is an equally legitimate need for more customisable support systems and tools.

In order to satisfy our first research objective, the establishment of a method and tool for undertaking risk management in preservation contexts, we initially surveyed a range of representative contexts within which data is created, used and preserved. The series of repository audits documented within this thesis is a unique collection of perspectives of preservation practice. These analyses are of demonstrable benefit to prospective and more experienced data custodians. We have successfully characterised the range of elements that comprise a successful preservation system and present these as an illustrative account of practice. More recent audits conducted beyond this research have yielded similar online examples [Rosenthal, 2014, Greenberg and Marks, 2012]. These examples have also illustrated the value of sharing the experiences of the audit process as well as the characteristics of a trusted (or risk prone) preservation system. Their efforts and the associated community response (in common with that which was inspired by our research) offer some additional validation of the benefits of presenting a transparent account of preservation practice. They also affirm technological choices made within the course of this research. For instance, they share our adoption of wiki technology and the alignment of documentation and associated knowledge with a structured certification instrument (in each case the *TRAC* criteria). It is encouraging that they also appear to embrace a similar spirit of transparency as prompted our efforts.

DRAMBORA and its manifestation as an interactive online tool closely reflect the findings of these surveys. Our methodology was subject to an iterative process of development that refined several aspects. Among the most notable was its initial insufficiently prescriptive nature; many of those participating in self-assessments reported difficulties determining their risk exposure, lacking a sufficient comprehensiveness of perspective to understand the extent of shortcomings or optimal approaches for improvement. Our interactive tool takes the core methodology of reflective self-assessment and adds a means by which users can relate their efforts to those underway elsewhere.

Our development of this methodology into an integrated online tool and its development by both direct and indirect exposure to a range of real life preservation scenarios has enhanced the methodology and overcome several of the challenges associated with a primarily self-

assessment driven approach. *DRAMBORA* Interactive is a demonstrably successful online resource, and has enjoyed high impact and considerable usage since its release by a world-wide audience of practitioners and educators.

This methodology is a contribution of critical importance. Many of the efforts to date associated with evaluating digital preservation have been motivated by funders and repository end users seeking affirmation or validation of the competencies of a given preserving organisation. This has yielded certification standards but seldom has an associated process for determining conformity been explored. More recently, a companion standard describing *Requirements for bodies providing audit and certification of candidate trustworthy digital repositories* has emerged but this remains of value and significance mainly to a small elite of accredited individuals and organisations. Our method is firmly aimed at those doing preservation, and provides the means for them to be confident in the suitability and sufficiency of their work, and if appropriate to facilitate a later exposure of their efforts to external certification. It also ensures that the dynamic nature of preservation (which may change based on new technological innovations, such as the increased dependency on cloud computing) continues to be reflected in best practice.

As *DRAMBORA* benefited from its deployment in formal audits, it similarly equipped us to pursue our second research objective, a continuation of the work of surveying preservation contexts. Our online tool *DRAMBORA Interactive* has yielded data corresponding to assessments that exhibit diversity in geographic location, legal context, types of digital collection, mandate and budgetary model. *DRAMBORA* requires participants to describe their preservation efforts in terms of objectives, activities, resources and risks, providing immediate evidence of how leading organisations go about ensuring the longevity of digital materials. Taking this as a starting point we sought to complement this data by engaging directly with a selection of organisations that was similarly representative as those using the tool online. We undertook a series of systematic audits based on evolving methodological and intellectual criteria (themselves developed iteratively based on our findings). Good and bad practice were both identified, characterised and related within these assessments, providing evidence that would form the basis for lines of enquiry and evaluation feedback in subsequent audit exercises. Collectively, the assessments yielded several primary benefits. The first, taken at face value was that the participating organisations were given the opportunity to better understand their successes and shortcomings, and to adapt to a critique based on comparison with a combination of objectively conceived and empirical real world best practice. Secondly, we took the opportunity to refine our approach both in terms of process and perhaps more importantly in the intellectual basis upon which our evaluation was conducted. Simply, as we learned more about how preservation takes place we became better equipped to identify opportunities for improvement elsewhere. Our knowledge base was developing, and it became clear that the perspectives we had been granted by exhaustively assessing a series

of operations were unique. The existing instruments that we were using were highlighted as being occasionally incomplete, more commonly at least lacking in terms of specificity or applicability to real world motivations and approaches.

Subsequent assessments, and analysis of third party undertaken assessments enabled us to produce evidence of the robustness and applicability of our best practice conclusions, detailed in the evaluation chapter of this thesis. Similarly, surveys of other institutions such as the London School of Economics and Political Science provide empirical evidence of the growing institutional appetite for appropriate data management and preservation capacity.

To enjoy these latter benefits the outcomes of the audit were recorded, accomplishing our third research objective. *DRAMBORA Interactive*'s existing data structure provided a means for recording facets of individual preservation efforts. We used this as the basis for initial distillation of survey outcomes to a more structured and easily comparable format, developing our taxonomy on an iterative basis. Maintaining a legacy association with the categories of analysis outlined in the *TRAC* standard we characterised our evidence base in terms of what organisations wish to achieve (agnostic of any given objective standard) and the processes, tools, policies and mandates that inform and/or support them. These were in turn related to a developing catalogue of risks, whether caused by or mitigated by these factors. Each facet was recorded in two forms; a higher level, more generic expression was intended to be immune to issues of applicability across context and time, complemented by specific examples - the particular implementations or manifestations we observed and recorded in our audit experiences. Injecting semantic qualities to the data is of tremendous importance, as it allows the data to be interrogated by applications or human users and enables the conception of a network of relatable factors that contribute to preservation outcomes.

Fourthly we evaluated existing methodologies for undertaking preservation assessment. A critical dimension of this work has been to establish where our outcomes are positioned within an existing international preservation certification landscape. We are not content to seek to replace wholesale the existing provisions, several of which have enjoyed formal standardisation. Instead, we seek to identify and fill the gaps in what currently exists. We considered the value and applicability of several standards and de facto standards and offered a critical assessment of each. Several leading examples are collectively encapsulated within the European Framework for Audit and Certification of Digital Repositories, which presents a series of increasingly onerous certification tiers that correspond with the Data Seal of Approval, ISO 16363 and the equivalent German standard. The first two tiers require just documented self-assessment while the most involved requires a full externally administered audit to be conducted within any organisation seeking certified status. We reflected on the many positive aspects of these resources. Each has at least some intellectual basis in the *Trustworthy Repository Audit and Certification Criteria and Check-list*. By extension, each can be considered a valuable expression of generic aspects of preservation practice.

TRAC's formal standardisation can be considered a more practical expression of - and companion resource to - the equally seminal *Reference Model for an Open Archival Information System*. However, their shortcomings are mainly in terms of their utility and practical applicability. Even though self-assessment comprises two-thirds of the Framework there remains little explicit emphasis on process to guide a prospective repository administrators seeking to evaluate his or her efforts. While the *Data Seal of Approval* presents as an advantage its low barrier to entry this is accompanied by shortcomings, principally in terms of lack of granularity of coverage. The *TRAC* and ISO standards, pursuing exhaustiveness, extend to many, many criteria; preservation is a complicated business with implications spanning every aspect of an organisation's administration, technology and information management process. But they in turn expose themselves to criticism as impractically conceived, beset by uncertainties in terms of how metrics can be satisfied and based upon a set of preservation requirements so generic that it doesn't really exist within any single organisational context.

We have conceived the *PORRO* ontology as a structured expression of preservation best practice, collated from over a dozen full-scale audit exercises in addition to around one hundred and fifty online self-assessments conducted using *DRAMBORA Interactive*. This qualifies this data as a legitimate consolidation of overall preservation practice, a unique dataset that was both conceived and validated by lengthy exposure to real world preservation efforts undertaken by experts in the field. We have sought to take advantage of the resources that are available and position our efforts in a fashion that ensures their compatibility. *DRAMBORA* was our direct response to the difficulties posed by a wholly top-down approach, a process-driven methodology that requires self-assessors to reflect on their own priorities and their associated strengths and shortcomings. *PORRO* enhances this process by providing pliable hooks to best-practice that are customisable to any given preservation context. This satisfies our fourth objective, a presentation of best practice in a taxonomical and ontological format. Like *DRAMBORA*, its design has been principally motivated by its associated use cases.

PORRO is demonstrably successful when applied to a range of contexts, most importantly evidenced by its usefulness in a further series of evaluator deployments, participants expressing satisfactions at the extent to which it is capable of off-setting the challenges of self-assessment. Further validation and evidence of adaptability was obtained by using *PORRO* as part of data management planning exercises at Queen Mary University of London and LSE (used in association with the DCC's *CARDIO* tool). Its completeness and coherence with respect to existing metrics is also evidenced within our further two pronged evaluation, whereby the ontology elements were mapped to facets of systems that have enjoyed community-approved certification and also simulated against third party audit results to reveal their expressiveness.

In isolation the value of an ontology is difficult to convey and therefore a suite of indicative tools that use the ontology as their intellectual foundation is an important step. The

effectiveness of our prototype tool portfolio, the delivery of which is our final research objective, is evident within two operational contexts. The first is research data management, where its adoption as a data source for the *CARDIO* collaborative data curation evaluation tool has been demonstrably useful. That is a process that builds consensus of a given organisation's data management capacity. At the point where individual contributors have agreed upon the status of their existing efforts reference is made to *PORRO* to identify potential approaches to improve existing provisions. This can be considered *PORRO*'s preservation planning application. Within the *3D Coform Repository Infrastructure* (RI), *PORRO* is used in the identification of risk, whereby a given set of real world circumstances are identified within the ontology and traced to potential associated risk factors. This is what we mean by bidirectionality within *PORRO*. Its applicability is such that it can be used to provoke the development of preservation activities or resource acquisition, or to warn of threats associated with existing or proposed systems.

Similarly practical expression of *PORRO*'s value is evidenced by referring to core use cases that *PORRO* is capable of satisfying. Through its adoption in the applications referenced earlier we can say that *PORRO* supports the identification of risk (whereby users or user agents can traverse the knowledge base to identify linked concepts based on common identified contextual characteristics); the facilitation of risk resolution (whereby risks that have been identified externally or using the ontology are mapped to appropriate mitigation measures); performance of gap analysis (whereby real world generic goals are represented in the ontology, as are prescribed criteria which in turn are mapped to *PORRO* concepts, both fleshed out by their correspondence to required or appropriate parameter considerations, actions or resources); and validation of approaches (whereby particular policy, resource or activity prioritisation can be traced to corresponding objectives and risks which illustrate the appropriateness of investment).

6.2 Future Work

Future work associated with *PORRO* has the potential to be exciting and, with certification establishing increasing practical momentum [Giaretta and Lambert, 2012], highly impactful. With the definition of a robust means of storing properties of objects, representation methods, context and risk we aim to develop existing tools to support formation of more sophisticated relationships between object, system and contextual properties and risks that encapsulate their relationships. Automating the processes of repository, object and risk classification will in turn support existing preservation planning and maturity modeling approaches.

In specific terms we hope to take the evaluation activities undertaken here and continue to focus more explicitly on not just preservation infrastructures, but also specific properties as-

sociated with digital objects themselves, and their relationship with the contexts within which they are preserved. More domain specific analysis will provide insights intended to support better understanding of risk relationships at the level of content, relatable to wider aspects of repository and external context. Furthermore, as we have done with publicly available case studies detailing existing preservation infrastructures, we will look to exploit existing data property resources, including the *Plato* preservation planning tool. By documenting information properties and values from its varied implicitly recorded preservation plans we will continue to extend and enrich the *PORRO* knowledge base. In turn we anticipate that this will offer greater insights into preservation optimization and risk awareness.

Meanwhile, we aim to continue to extend institutional provisions, structuring and ingesting activity, resource, risk and liability data from over one hundred newly completed organizational assessments undertaken using the *DRAMBORA interactive* tool. The implicit sensitivity of much of this data limits opportunities to exploit it more widely, but its anonymisation and redistribution in a public tool also remains a critical planned outcome of this work.

Appendices

Appendix A

PORRO Classes

The complete set of PORRO classes follows in this Appendix. Indentation is indicative of superclass/subclass relationships. The .owl file representing the full ontology, including properties, is available from:

`http://mchughontology.hatii.arts.gla.ac.uk/porro.owl`

E001_Porro_Entity
- E002_Preservation_Criterion
- E003_Preservation_Criterion_Evidence
- E004_Preservation_Criteria_Source
- - E005_International_Standard
- - E006_National_Standard
- - E007_Community_Standard
- - E008_Article_Or_Conference_Proceeding
- - E009_Research_Report
- - E010_Audio_Or_Transcript
- - E011_Tool_or_Learning_Resource
- - E012_Audit_Or_Certification_Report
- - E013_Law_Or_Regulation
- E014_Custodial_Entity
- E015_Context_Characteristic
- - E016_Funding_Source
- - E017_Budget
- - E018_Staff
- - E019_Staff_Role
- - E020_Domain
- E021_Functional_Entity
- - E022_Preservation_Goal

- - - E023_Adopt_Appropriate_Preservation_Formats
- - - E024_Authenticate_Source_Of_Ingested_Packages
- - - E025_Backup_Documentation
- - - E026_Classify_Archival_Data
- - - E027_Continuously_Validate_Data_Integrity
- - - E028_Define_Disaster_Recovery_Policy
- - - E029_Define_Ingest_Package_Specification
- - - E030_Define_Policy_And_Procedures_For_Undertaking_Backups
- - - E031_Document_Archival_Data
- - - E032_Document_Software_Dependencies
- - - E033_Ensure_Appropriate_Contractual_Management
- - - E034_Ensure_Synchronisation_Of_Data_Separated_By_Time_Or_Space
- - - E035_Establish_And_Exercise_Ingest_Policy
- - - E036_Establish_And_Exercise_Selection_Policy
- - - E037_Establish_And_Maintain_Terms_Of_Deposit
- - - E038_Establish_Appropriate_Backup_Redundancy_Provisions
- - - E039_Establish_Appropriate_Backup_Remoteness_Provisions
- - - E040_Establish_Appropriate_Business_Planning
- - - E041_Establish_Appropriate_Categories_Of_Staff
- - - E042_Establish_Appropriate_Contingency_Funding
- - - E043_Establish_Appropriate_Coordination_And_Steering_Platform
- - - E044_Establish_Appropriate_Database_Backup_Infrastructure
- - - E045_Establish_Appropriate_Financial_Accounting_Infrastructure
- - - E046_Establish_Appropriate_Hardware_Infrastructure
- - - E047_Establish_Appropriate_Logical_Security_Provisions
- - - E048_Establish_Appropriate_Physical_Security_Provisions
- - - E049_Establish_Appropriate_Provisions_For_Backup
- - - E050_Establish_Appropriate_Software_Infrastructure
- - - E051_Establish_Appropriate_Strategies_For_Facilitating_Succession_Of_Organisation_Or_Content
- - - E052_Establish_Appropriate_Technical_Documentation_Base
- - - E053_Establish_Archival_Packages_Configuration
- - - E054_Establish_Assurances_Of_Availability_Of_Appropriate_Technical_Skills
- - - E055_Establish_Assurances_Of_Recoverability_Of_Any_Lost_Data
- - - E056_Establish_Assurances_Of_Site_Stability
- - - E057_Establish_Assurances_Of_Sufficiency_Of_Staff_Skills_And_Capacity
- - - E058_Establish_Assurances_That_All_Costs_Are_And_Will_Continue_To_Be_Covered
- - - E059_Establish_Budget_Dedicated_To_Training_Provision
- - - E060_Establish_Budgetary_Protection_Assurances

- - - E061_Establish_Conditions_For_Access
- - - E062_Establish_Criteria_For_Data_Identification
- - - E063_Establish_Criteria_For_Data_Review
- - - E064_Establish_Criteria_For_Disposal
- - - E065_Establish_Data_Ownership
- - - E066_Establish_Designated_Community
- - - E067_Establish_Hardware_Upgrade_Policy
- - - E068_Establish_Information_Security_Policy
- - - E069_Establish_Levels_Of_Preservation
- - - E070_Establish_List_Of_Supported_Formats
- - - E071_Establish_Logical_Storage_Provisions
- - - E072_Establish_Means_For_Data_Disposal
- - - E073_Establish_Means_For_Data_Identification
- - - E074_Establish_Means_For_Data_Review
- - - E075_Establish_Means_To_Track_Data_Object_Through_Preservation_Workflow_And_Lifecycle
- - - E076_Establish_Media_Refreshment_Policy
- - - E077_Establish_Naming_Convention
- - - E078_Establish_Physical_And_Logical_Provisions_For_Access
- - - E079_Establish_Policy_Review_Policy
- - - E080_Establish_Policy_Transparency
- - - E081_Establish_Portfolio_Of_Internal_Or_External_Staff_Training_Provisions
- - - E082_Establish_Ratification_Of_Preservation_Mission_From_Parent_Or_Governing_Entity
- - - E083_Establish_Relationship_Between_Access_And_Archival_Packages
- - - E084_Establish_Relationship_Between_Ingest_And_Archival_Packages
- - - E085_Establish_Relationships_With_Succession_Partners
- - - E086_Establish_Software_Upgrade_Policy
- - - E087_Establish_Suitability_Of_Backup_Infrastructure_Through_Testing
- - - E088_Establish_Terms_Of_Use
- - - E089_Establish_Transformation_Procedure_From_Ingest_To_Archival_Packages
- - - E090_Evaluate_And_Certify_Activities
- - - E091_Exercise_Preservation_Plans
- - - E092_Implement_Access_Controls
- - - E093_Implement_Categories_Of_Access
- - - E094_Initiate_Stakeholder_Dialogue
- - - E095_Limit_Data_Loss_Incidence
- - - E096_Maintain_Archival_Package_Referential_Integrity
- - - E097_Maintain_Best_Practice_Awareness
- - - E098_Maintain_Budget_Carry-Over_Facility

- - - E099_Maintain_Business_Planning_Autonomy
- - - E100_Maintain_Comprehensive_Costings_Breakdown
- - - E101_Maintain_Data_Integrity
- - - E102_Maintain_Depositor_Dialogue
- - - E103_Maintain_End_User_Dialogue
- - - E104_Maintain_Link_Between_Data_And_Metadata
- - - E105_Maintain_Risk_Awareness
- - - E106_Make_Explicit_Preservation_Responsibility
- - - E107_Make_Explicit_Preservation_Rights
- - - E108_Manage_Formation_Of_Dissemination_Package
- - - E109_Monitor_Access_Behaviours
- - - E110_Monitor_And_Fulfil_Freedom_Of_Information_Responsibilities
- - - E111_Monitor_And_Fulfil_Ipr_Responsibilities
- - - E112_Monitor_And_Fulfil_Other_Legislative_And_Legal_Responsibilities
- - - E113_Monitor_And_Respond_To_Designated_Community_Evolution
- - - E114_Monitor_File_Format_Obsolescence
- - - E115_Monitor_Unauthorised_Access
- - - E116_Physically_Acquire_Content
- - - E117_Plan_For_Preservation
- - - E118_Process_Ingested_Content
- - - E119_Record_And_Maintain_Descriptive_Metadata
- - - E120_Record_And_Maintain_Representation_Information
- - - E121_Record_Appropriate_Metadata
- - - E122_Select_And_Appraise_Ingested_Content
- - - E123_Select_Preservation_Strategies
- - - E124_Validate_Data_Integrity
- - - E125_Validate_Integrity_Of_Backups
- - - E126_Verify_Ingest_Package_Conformity_With_Specification
- - E127_Preservation_Parameter
- - - E128_Alignment_Of_Roles_And_Skills_And_Function
- - - E129_Backup_Strategy
- - - E130_Budgetary_Separation_And_Autonomy
- - - E131_Business_Prioritisation_Areas
- - - E132_Compliance_Responsibility
- - - E133_Content_Access_Levels
- - - E134_Content_Change
- - - E135_Content_Closure
- - - E136_Content_Modification

- - - E137_Content_Removal_And_Deletion
- - - E138_Content_Representation
- - - E139_Content_Selection_And_Acceptance
- - - E140_Content_Versioning
- - - E141_Content_And_System_Redundancy
- - - E142_Contract_Types
- - - E143_Contract_And_Mandate_Cessation
- - - E144_Coordination_And_Steering
- - - E145_Copyright_Challenge_Response
- - - E146_Copyright_In_Collection
- - - E147_Copyright_And_Access_Restrictions
- - - E148_Cost_Model_For_Access_Provision
- - - E149_Data_Representation
- - - E150_Data_Review
- - - E151_Data_Rights_Transfer
- - - E152_Designated_Community_Definition
- - - E153_Disaster_Planning
- - - E154_Discontinuing_Preservation
- - - E155_Dissemination_Specification
- - - E156_Documentation_Availability
- - - E157_Documentation_Requirements
- - - E158_Documentation_Review
- - - E159_Evaluation_Metrics_And_Participants
- - - E160_Exemptions_To_Preservation_Responsibility
- - - E161_External_Skills_Procurement
- - - E162_Format_Migration
- - - E163_Funding_Sources
- - - E164_Identification_And_Naming
- - - E165_Income_And_Expenditure
- - - E166_Ingest_Mechanism
- - - E167_Ingest_Specification
- - - E168_Internal_Budgetary_Allocation
- - - E169_Legal_Requirements_For_Due_Process
- - - E170_Legal_Responsibilities
- - - E171_Logical_Authorisation
- - - E172_Logical_Security_Measures
- - - E173_Logical_Security_Responsibility
- - - E174_Logical_Storage

- - - E175_Media_Refreshment
- - - E176_Metadata_Creation_Responsibility
- - - E177_Metadata_Creation_Workflow
- - - E178_Metadata_Format
- - - E179_Metadata_Representation
- - - E180_Metadata_Storage
- - - E181_Minimal_Required_Metadata
- - - E182_Obsolescence_Risk_Tolerance
- - - E183_Oversight_For_Policy_Review
- - - E184_Package_Specifications
- - - E185_Physical_Access_Authorisation
- - - E186_Physical_Security_Measures
- - - E187_Physical_Security_Responsibility
- - - E188_Physical_Storage
- - - E189_Policy_Covering_Distribution_Of_Copyright_Material
- - - E190_Policy_Describing_Designated_Community
- - - E191_Policy_Development_Traceability
- - - E192_Policy_Development_Triggers
- - - E193_Policy_Flexibility
- - - E194_Policy_For_Documenting_Change
- - - E195_Policy_For_Negotiation_Of_Preservation_Responsibility
- - - E196_Policy_For_Wider_Data_Management_Integration
- - - E197_Policy_Governing_Withdrawal_Of_Data_Management_Responsibility
- - - E198_Policy_On_Access_Control
- - - E199_Policy_On_Accountability
- - - E200_Policy_On_Backup_Frequency
- - - E201_Policy_On_Backup_Location
- - - E202_Policy_On_Budgetary_Management
- - - E203_Policy_On_Budgetary_Planning
- - - E204_Policy_On_Business_Planning
- - - E205_Policy_On_Circumstances_That_Provoke_Change
- - - E206_Policy_On_Content_Availability
- - - E207_Policy_On_Contents_Of_Backup_Package
- - - E208_Policy_On_Relationship_Between_Ingest_And_Archival_And_Dissemination_Packages
- - - E209_Policy_On_Supported_Access_Types
- - - E210_Policy_Responsibility
- - - E211_Policy_Review_Due_Process
- - - E212_Policy_Steering

- - - E213_Policy_Transparency
- - - E214_Preservation_Commitment
- - - E215_Preservation_Level_Assignment
- - - E216_Preservation_Level_Implications
- - - E217_Preservation_Mechanism
- - - E218_Preservation_Package_Structure
- - - E219_Preservation_Prioritisation
- - - E220_Preservation_Risk
- - - E221_Preservation_Strategy
- - - E222_Preservation_Validation
- - - E223_Procedure_For_Change_Management
- - - E224_Process_And_Infrastructure_Review
- - - E225_Professional_Membership
- - - E226_Quality_Assurance_Responsibility
- - - E227_Recovery_Drills
- - - E228_Recruitment_And_Retention
- - - E229_Repository_Integration
- - - E230_Required_Redundancy
- - - E231_Review_Of_Designated_Community
- - - E232_Rights_And_Ownership_Definitions
- - - E233_Risk_Assessment_Validation
- - - E234_Risk_Management
- - - E235_Risk_Tolerance
- - - E236_Scalability_Requirements
- - - E237_Security_Failure_Defaults
- - - E238_Selection
- - - E239_Service_Breadth_And_Prioritisation
- - - E240_Service_Business_Model
- - - E241_Service_Level
- - - E242_Service_Level_Parameters
- - - E243_Specification_For_Archival_Packages
- - - E244_Specification_Relationships
- - - E245_Staff_Resource_Scalability_Requirements
- - - E246_Staff_Turnover
- - - E247_Succession_Arrangement
- - - E248_Succession_Responsibilities
- - - E249_Supported_Acquisition_Methods
- - - E250_Supported_Dissemination_Formats

- - - E251_Supported_Ingest_Formats
- - - E252_Supported_Preservation_Formats
- - - E253_Supported_Systems_And_Applications
- - - E254_Systems_Development_Management
- - - E255_Technical_Review
- - - E256_Technological_Contingency
- - - E257_Technology_Licensing
- - - E258_Technology_Skills_Development
- - - E259_Technology_To_Workflow_Mapping
- - - E260_Terms_Of_Access
- - - E261_Terms_Of_Reference
- - - E262_Training
- - - E263_Understandability
- - - E264_Usage_To_Preservation_Level_Relationship
- - - E265_User_Competency_Requirements
- - - E266_Validation_Checks_And_Requirements
- - E267_Preservation_Right_Or_Responsibility
- - - E268_Data_Management_Objectives_Consistent_With_Parent
- - - E269_Data_Management_Responsibility
- - - E270_Data_Management_Rights
- - - E271_Has_Assurance_Of_Financial_Sustainability
- - - E272_Has_Business_Steering
- - - E273_Has_Legal_Responsibility_To_Manage_Data
- - - E274_Has_Legal_Responsibility_To_Share_Data_And_Provide_Access
- - - E275_Has_Limitation_Of_Liabilities
- - - E276_Has_Mandate_To_Aggregate_Published_Data
- - - E277_Has_Mandate_To_Manage_And_Distribute_Copyright_Materials
- - - E278_Has_Mandated_Data_Closure_Responsibilities
- - - E279_Has_Mandated_Data_Sharing_Responsibilities
- - - E280_Has_Mandated_Data_Sharing_Triggers
- - - E281_Has_Mandated_Staff_Development_Requirements
- - - E282_Has_Mandated_Transparency_Requirement
- - - E283_Has_Prescribed_Minimal_Metadata_Requirements
- - - E284_Has_Preservation_Policy_Discretion
- - - E285_Has_Preservation_Responsibility
- - - E286_Has_Preservation_Rights
- - - E287_Has_Responsibility_To_Limit_Access
- - - E288_Has_Restrictions_On_Data_Management_Or_Distribution_Based_On_Copyright_Status

- - - E289_Has_Restrictions_On_Termination_Of_Data_Management_Responsibilities
- - - E290_Has_Rights_To_Defer_Data_Management_Responsibility
- - - E291_Has_Selection_Mandate
- - - E292_Has_Stakeholder_Management_Responsibility
- - - E293_Mandate_For_Policy_And_Procedure_Discretion
- - - E294_Succession_Partnership_Agreement
- - - E295_Sufficiency_And_Suitability_Of_Audit_Practice
- - E296_Preservation_Activity
- - - E297_Accept_Data_Management_Responsibility
- - - E298_Aggregate_Data_Referenced_By_Or_Contextual_To_Dataset
- - - E299_Anonymise_Data
- - - E300_Assign_A_Processing_Record_To_Data
- - - E301_Audit_Collections_And_Procedures
- - - E302_Automate_Metadata_Extraction
- - - E303_Communicate_Service_Disruption
- - - E304_Create_Object_Metadata
- - - E305_Create_Package_Descriptor
- - - E306_Define_Package_Specifications
- - - E307_Define_Policy_And_Procedure_Review_Triggers
- - - E308_Develop_Active_Training_Plans
- - - E309_Develop_Dedicated_Budget
- - - E310_Develop_Income_Streams
- - - E311_Develop_Technical_Training_And_Induction
- - - E312_Digitise_Analogue_Content
- - - E313_Dispose_Of_Content_And_Media_And_Metadata
- - - E314_Dispose_Of_Non-Ingested_Content
- - - E315_Disseminate_Content_And_Metadata
- - - E316_Document_Interactions_Surrounding_Dataset
- - - E317_Document_Package_Content
- - - E318_Document_Package_Structure
- - - E319_Document_Public_Release_Of_Dataset
- - - E320_Duplicate_Content
- - - E321_Duplicate_Metadata
- - - E322_Duplicate_Systems
- - - E323_Enforce_Secure_Logical_Environment
- - - E324_Engage_In_Dialogue_With_Stakeholder
- - - E325_Engage_Internally_On_Policy_Review
- - - E326_Establish_Income_Streams

- - - E327_Establish_Preservation_Plan
- - - E328_Establish_Referential_Integrity
- - - E329_Establish_Succession_Arrangements
- - - E330_Evaluate_And_Reform_Policy
- - - E331_Evaluate_And_Reform_Procedures
- - - E332_Evaluate_Data_Copyright_Status
- - - E333_Evaluate_Format_And_Media_Risk
- - - E334_Evaluate_Logical_Security_Threats
- - - E335_Evaluate_Physical_Security_Threats
- - - E336_Evaluate_Preservation_Plan
- - - E337_Evaluate_Risk_Exposure
- - - E338_Exchange_Transfer_Documentation
- - - E339_Execute_Preservation_Plan
- - - E340_Expose_Data_To_Access
- - - E341_Generate_Fixity_Information
- - - E342_Identify_Data_Properties
- - - E343_Incentivise_And_Retain_Staff
- - - E344_Justify_Resources
- - - E345_Liaise_With_Security_Provider
- - - E346_Liaise_With_Technology_Provider
- - - E347_Link_Metadata_To_Corresponding_Data
- - - E348_Link_Preserved_Content_With_Original
- - - E349_Log_Accessions
- - - E350_Log_Actions_And_Interactions
- - - E351_Log_Object_Lifecycle
- - - E352_Log_Unauthorized_Access_Attempts
- - - E353_Maintain_Access_Platform
- - - E354_Maintain_Administration_Platform
- - - E355_Maintain_Appropriate_Documentation
- - - E356_Maintain_Authentication_Platform
- - - E357_Maintain_Authorisation_Platform
- - - E358_Maintain_Backup_Platform
- - - E359_Maintain_Generic_And_Shared_Technology
- - - E360_Maintain_Ingest_Platform
- - - E361_Maintain_Network_Protocol_Support
- - - E362_Maintain_Preservation_Platform
- - - E363_Maintain_Redundant_Systems_And_Data
- - - E364_Maintain_Risk_Register

- - - E365_Maintain_Storage_Platform
- - - E366_Manage_Format_And_Media_Support
- - - E367_Manage_Package_Specifications
- - - E368_Manage_Policy_Revision
- - - E369_Manage_Unique_Identification
- - - E370_Migrate_Format_Or_Media
- - - E371_Monitor_Access
- - - E372_Monitor_Copyright_Status
- - - E373_Monitor_Data_Citations_And_Reuse
- - - E374_Monitor_Dataset_Usage
- - - E375_Monitor_Designated_Community_Evolution
- - - E376_Monitor_Security_Status
- - - E377_Monitor_Skills_Gaps
- - - E378_Monitor_Training_Opportunities
- - - E379_Monitor_Training_Requirements
- - - E380_Monitor_User_Requirements
- - - E381_Monitor_User_Satisfaction
- - - E382_Negotiate_Data_Management_Mandate
- - - E383_Notify_Data_Originator_Of_Data_Receipt
- - - E384_Perform_Metadata_Format_Conversion
- - - E385_Perform_Test_System_Recoveries
- - - E386_Plan_And_Execute_System_Upgrades
- - - E387_Plan_Expenditure
- - - E388_Plan_For_Risk_Mitigation_And_Avoidance
- - - E389_Procure_External_Expertise
- - - E390_Publish_Package_Specifications
- - - E391_Pursue_Dedicated_Research_Funding
- - - E392_Record_Changes
- - - E393_Record_Media_Movement
- - - E394_Record_System_Changes
- - - E395_Recruit_Skilled_Staff
- - - E396_Reference_External_Sources_During_Data_Management_Planning
- - - E397_Refresh_Media_Or_Hardware
- - - E398_Refuse_Content_Ingest
- - - E399_Regulate_Access_To_Data
- - - E400_Regulate_Dataset_Closure
- - - E401_Renegotiate_Legal_Mandate
- - - E402_Report_Technical_Status

- - - E403_Request_Data_Deposit
- - - E404_Respond_To_Ipr_Challenge
- - - E405_Retrieve_Content
- - - E406_Review_Business_Performance
- - - E407_Review_Business_Priorities
- - - E408_Review_Legal_Responsibilities_And_Rights
- - - E409_Review_Metadata
- - - E410_Review_Partnerships_And_Alignments
- - - E411_Review_Technical_Provision
- - - E412_Scan_For_Viruses
- - - E413_Securely_Store_Data_And_Media
- - - E414_Seek_Budgetary_Assurances
- - - E415_Self-Evaluate_Activities
- - - E416_Synchronise_Redundant_Data
- - - E417_Technological_Training_And_Induction
- - - E418_Test_Effects_Of_Changes
- - - E419_Transfer_Skills
- - - E420_Undertake_Independent_Audit
- - - E421_Undertake_Test_Recovery
- - - E422_Validate_Content
- - - E423_Validate_Media_And_Storage
- - - E424_Verify_Characteristics_Of_Data
- - - E425_Verify_Data_Formats
- - E426_Preservation_Resource
- - - E427_Preserved_Resource
- - - - - E428_Preserved_Source
- - - - - - E429_Preserved_Source_Bitstream
- - - - - - E430_Preserved_Source_Metadata
- - - - - - E431_Preserved_Source_Fixity
- - - - - E432_Preserved_Process
- - - - - - E433_Preserved_Process_Bitstream
- - - - - - E434_Preserved_Process_Metadata
- - - - - - E435_Preserved_Process_Fixity
- - - - - E436_Preserved_Performance
- - - - - - E437_Preserved_Performance_Metadata
- - - E438_Preservation_Support_Resource
- - - - - E439_Access_Control_System
- - - - - E440_Access_Personalisation_System

- - - - E441_Access_Platform
- - - - E442_Access_Processing_System
- - - - E443_Access_Validation_System
- - - - E444_Acquisition_Tracking_System
- - - - E445_Administration_Platform
- - - - E446_Alarm_System
- - - - E447_Ambient_Environment_Sensors
- - - - E448_Authentication_Subsystem
- - - - E449_Authorisation_Subsystem
- - - - E450_Awarded_Certifications
- - - - E451_Backup_And_Recovery_Management_System
- - - - E452_Backup_Media
- - - - E453_Backup_Platform
- - - - E454_Budgetary_Assurances
- - - - E455_Business_And_Organisation_Documentation
- - - - E456_Business_Plan
- - - - E457_Catalogue
- - - - E458_Change_Management_System
- - - - E459_Changelog
- - - - E460_Closed_Data_Policy
- - - - E461_Communication_Channels
- - - - E462_Communication_Records
- - - - E463_Content_Processing_Forms
- - - - E464_Content_Processing_System
- - - - E465_Content_Retriever
- - - - E466_Content_Skills
- - - - E467_Contingency_And_Reserve_Fund
- - - - E468_Contingency_Fund
- - - - E469_Contingency_Non-Monetary_Resources
- - - - E470_Copyright_Trigger
- - - - E471_Copyrighting_Mechanism
- - - - E472_Custodial_History_Record
- - - - E473_Custodial_History_Records
- - - - E474_Data_Documentation
- - - - E475_Data_Management_Skills
- - - - E476_Data_Security_Enforcement
- - - - E477_Data_Transformation_Plans
- - - - E478_Dedicated_Budget

- - - - E479_Dedicated_Human_Resources
- - - - E480_Delineated_Roles_And_Responsibilities
- - - - E481_Deposit_Agreement
- - - - E482_Depositor_Fixity_Values
- - - - E483_Disaster_Plan
- - - - E484_Discovery_Metadata
- - - - E485_Documentation_Discovery_System
- - - - E486_Employment_Flexibility
- - - - E487_Employment_Incentives
- - - - E488_Expenditure_Projections
- - - - E489_External_Evaluators
- - - - E490_External_Policy_Influences
- - - - E491_External_Skills_Pools
- - - - E492_Feedback_Mechanism
- - - - E493_Fire_Detection_And_Suppression
- - - - E494_Focus_Group
- - - - E495_Formal_Contracts_And_Terms
- - - - E496_Format_Documentation
- - - - E497_Format_Support
- - - - E498_Forum_For_Technical_Exchange
- - - - E499_General_Hardware
- - - - E500_General_Software
- - - - E501_Generated_Fixity_Values
- - - - E502_Glossary_Of_Preservation_Terminology
- - - - E503_Historical_Policy_Records
- - - - E504_Identifier_Resolver
- - - - E505_Income_Generation_Skills
- - - - E506_Income_Streams
- - - - E507_Ingest_Platform
- - - - E508_Justification_Of_Resources
- - - - E509_Legal_Advice
- - - - E510_Legal_Expertise
- - - - E511_Legislation
- - - - E512_Logger
- - - - E513_Logical_Security_Monitoring_System
- - - - E514_Management_Board
- - - - E515_Management_Skills
- - - - E516_Mandate_Definition

- - - - E517_Means_For_Format_And_Media_Representation
- - - - E518_Media_Degradation_Diagnosis_Tools
- - - - E519_Media_Support
- - - - E520_Membership_Of_Partner_Network
- - - - E521_Metadata_Creation_Guidelines
- - - - E522_Metadata_Extraction_Software
- - - - E523_Metadata_Management_System
- - - - E524_Metadata_Records
- - - - E525_Metadata_Schema
- - - - E526_Metadata_Standards
- - - - E527_Moisture_Detection_And_Mitigation
- - - - E528_Network
- - - - E529_Obsolescence_Metric
- - - - E530_Package_Relationship_Documentation
- - - - E531_Package_Specification_Documentation
- - - - E532_Peer_Evaluator
- - - - E533_Physical_Security_Monitoring
- - - - E534_Policy_Documentation
- - - - E535_Policy_Makers
- - - - E536_Policy_Review_Manager
- - - - E537_Policy_Stakeholders
- - - - E538_Preservation_Capacity
- - - - E539_Preservation_Documentary_Resource
- - - - E540_Preservation_Management_System
- - - - E541_Preservation_Plan
- - - - E542_Preservation_Platform
- - - - E543_Preservation_Policy
- - - - E544_Preservation_Validation_System
- - - - E545_Processing_Record
- - - - E546_Quality_Assurance_Infrastructure
- - - - E547_Recruitment_Network
- - - - E548_Redistribution_Rights
- - - - E549_Redundant_Data_And_System_Site
- - - - E550_Redundant_Resources
- - - - E551_Redundant_Storage
- - - - E552_Redundant_Utilities
- - - - E553_Relationship_With_Partner_Associations
- - - - E554_Representation_Information_Registry

- - - - E555_Rights_Database
- - - - E556_Risk_Intelligence_Data
- - - - E557_Risk_Register
- - - - E558_Room_Access_System
- - - - E559_Secure_Network_Infrastructure
- - - - E560_Secure_Safe
- - - - E561_Secure_Storage_Location
- - - - E562_Security_Platform
- - - - E563_Skills_Monitoring_System
- - - - E564_Societies_And_Professional_Organisations_Membership
- - - - E565_Stakeholder_Liaison_Forum
- - - - E566_Stakeholder_Relationships
- - - - E567_Storage_Platform
- - - - E568_Succession_Partner_Agreement
- - - - E569_System_Documentation
- - - - E570_System_Maintenance_And_Support_Agreement
- - - - E571_Technical_Capacity
- - - - E572_Technical_Community_And_Literature
- - - - E573_Technical_Skills
- - - - E574_Terms_Of_Access_And_Use
- - - - E575_Terms_Of_Use
- - - - E576_Training_Budget
- - - - E577_Training_Materials_And_Infrastructure
- - - - E578_Transaction_Documentation
- - - - E579_Understandability_Definition
- - - - E580_Update_And_Upgrade_Prompts
- - - - E581_User_Database
- - - - E582_Validation_System
- - - - E583_Weather_Protection_System
- E584_Preservation_Risk
- - E585_Accidental_System_Disruptions
- - E586_Activity_Is_Overlooked_Or_Allocated_Insufficient_Resources
- - E587_Ambiguity_Of_Understandability_Definition
- - E588_Archival_Information_Cannot_Be_Traced_To_A_Received_Package
- - E589_Authentication_Subsystem_Fails
- - E590_Authorisation_Subsystem_Fails
- - E591_Budgetary_Reduction
- - E592_Business_Fails_To_Preserve_Essential_Characteristics_Of_Digital_Information

- - E593_Business_Objectives_Not_Met
- - E594_Business_Policies_And_Procedures_Are_Inconsistent_Or_Contradictory
- - E595_Business_Policies_And_Procedures_Are_Inefficient
- - E596_Business_Policies_And_Procedures_Are_Unknown
- - E597_Change_Of_Terms_Within_Third-Party_Service_Contracts
- - E598_Community_Feedback_Not_Acted_Upon
- - E599_Community_Feedback_Not_Received
- - E600_Community_Requirements_Change_Substantially
- - E601_Community_Requirements_Misunderstood_Or_Miscommunicated
- - E602_Deliberate_System_Sabotage
- - E603_Destruction_Of_Primary_Documentation
- - E604_Destruction_Or_Non-Availability_Of_Repository_Site
- - E605_Documented_Change_History_Incomplete_Or_Incorrect
- - E606_Enforced_Cessation_Of_Repository_Operations
- - E607_Exploitation_Of_Security_Vulnerability
- - E608_Extent_Of_What_Is_Within_The_Archival_Object_Is_Unclear
- - E609_Externally_Motivated_Changes_Or_Maintenance_To_Information_During_Ingest
- - E610_False_Perception_Of_The_Extent_Of_Repository_Success
- - E611_Finances_Insufficient_To_Meet_Repository_Commitments
- - E612_Financial_Shortfalls_Or_Income_Restrictions
- - E613_Hardware_Failure_Or_Incompatibility
- - E614_Hardware_Or_Software_Incapable_Of_Supporting_Emerging_Repository_Aims
- - E615_Identifier_To_Information_Referential_Integrity_Is_Compromised
- - E616_Inability_To_Evaluate_Effectiveness_Of_Technical_Infrastructure_And_Security
- - E617_Inability_To_Evaluate_Repository_Successfulness
- - E618_Inability_To_Evaluate_Staff_Effectiveness_Or_Suitability
- - E619_Inability_To_Validate_Effectiveness_Of_Dissemination_Mechanism
- - E620_Inability_To_Validate_Effectiveness_Of_Ingest_Process
- - E621_Inability_To_Validate_Effectiveness_Of_Preservation
- - E622_Incompleteness_Of_Submitted_Packages
- - E623_Inconsistency_Between_Redundant_Copies
- - E624_Ingest_Subsystem_Fails
- - E625_Legal_Liability_For_Breach_Of_Contractual_Responsibilities
- - E626_Legal_Liability_For_Breach_Of_Legislative_Requirements
- - E627_Legal_Liability_For_Ipr_Infringement
- - E628_Liability_For_Non-Adherence_To_Financial_Law_Or_Regulations
- - E629_Liability_For_Regulatory_Non-Compliance
- - E630_Local_Destructive_Or_Disruptive_Environmental_Phenomenon

- - E631_Loss_Of_Authenticity_Of_Information
- - E632_Loss_Of_Availability_Of_Information_Or_Service
- - E633_Loss_Of_Budgetary_Autonomy
- - E634_Loss_Of_Confidentiality_Of_Information
- - E635_Loss_Of_Information_Provenance
- - E636_Loss_Of_Information_Reliability
- - E637_Loss_Of_Integrity_Of_Information
- - E638_Loss_Of_Key_Member_Of_Staff
- - E639_Loss_Of_Mandate
- - E640_Loss_Of_Non-Repudiation_Of_Commitments
- - E641_Loss_Of_Other_Third-Party_Contracts_And_Services
- - E642_Loss_Of_Performance_Or_Service_Level
- - E643_Loss_Of_Trust_Or_Reputation
- - E644_Loss_Or_Non-Suitability_Of_Backups
- - E645_Management_Failure
- - E646_Media_Degradation_Or_Obsolescence
- - E647_Metadata_To_Information_Referential_Integrity_Is_Compromised
- - E648_Misallocation_Of_Finances
- - E649_Negative_Perception_Of_Curation_Capacity
- - E650_Non-Availability_Of_Core_Utilities
- - E651_Non-Availability_Of_Information_Delivery_Services
- - E652_Non-Discoverability_Of_Information_Objects
- - E653_Non-Traceability_Of_Received_Or_Archived_Or_Disseminated_Package
- - E654_Obsolescence_Of_Hardware_Or_Software
- - E655_Physical_Intrusion_Of_Hardware_Storage_Space
- - E656_Preservation_Plans_Cannot_Be_Implemented
- - E657_Preservation_Strategies_Result_In_Information_Loss
- - E658_Remote_Or_Local_Software_Intrusion
- - E659_Shortcomings_In_Semantic_Or_Technical_Understandability_Of_Information
- - E660_Software_Failure_Or_Incompatibility
- - E661_Staff_Skills_Become_Obsolete
- - E662_Staff_Suffer_Deterioration_Of_Skills
- - E663_Structural_Non-Validity_Or_Malformedness_Of_Received_Packages
- - E664_Unidentified_Information_Change
- - E665_Unidentified_Security_Compromise_Or_Vulnerability_Or_Information_Degradation
- E666_Preservation_Risk_Influence

Appendix B

Case Studies

B.1 Background to these Case Studies

The case studies which can be found below are structurally roughly equivalently to those sections presented in the draft RLG-NARA check-list document which represents the intellectual foundation for each of these assessments. This is also the adopted structure for the section that summarises findings across the institutions in Chapter 3. Each case study aims to offer a readable account of the areas of success, and perceived shortcomings within the approach and infrastructure adopted and demonstrated by the datacentre. Common sections are Organisational Infrastructure, Digital Object Management, and Technologies and Technological Infrastructure. Within each, further subsections vary between the case studies.

The National Library case study provides an exception to this model. Unlike the other evaluation case studies this assessment was not led by the author; instead he joined colleagues from the US Center for Research Libraries [CRL, 2012a] who were completing a number of assessments as part of their Trustworthy Digital Repositories project. This was the third in their series of four assessments. As a consequence the account is more brief and differs structurally. Nevertheless, outcomes from this activity were similarly useful in informing the ontology of preservation infrastructure described in subsequent chapters.

B.2 Letter of Invitation to Participate

Dear <Repository Coordinator>,

My name is Andrew McHugh, and I'm a colleague of Seamus Ross at the Humanities Advanced Technology and Information Institute (HATII) at the University of Glasgow. I'm writing to request your assistance in the development of audit and certification mechanisms

for digital repositories through participation in the Digital Curation Centre's pilot repository audit programme.

The JISC/EPSRC funded Digital Curation Centre coordinates work being undertaken in four UK institutions, namely the Universities of Glasgow, Edinburgh and Bath and the Council for the Central Laboratory of the Research Councils (CCLRC). The Centre's aim is to provide a national focus for research and development into curation issues and to promote expertise and good practice for the management of digital information. A key current work area is the development of processes, tools and services to support the development of digital repositories. A significant question mark that continues to surround repositories is the means by which we can determine which repositories can be trusted to maintain our digital assets so that they might be used and reused at an uncertain point in the future. Therefore, we are working in collaboration with various US and European based efforts to contribute towards the development of evaluation criteria for digital archives; we ultimately aspire to play a key role in the establishment of formal audit and certification services for digital archives within the UK. In parallel with this we are committed to the provision of training to institutions to prepare them for the challenges that audit will pose.

In order to determine the most effective methodology for digital repository evaluation and to assess the existing criteria that have been conceived we are currently planning a short series of UK based pilot audits to be conducted within the forthcoming months, intended to complement those already undertaken by our international collaborators. The goals closely mirror those pursued in existing efforts, with the conception of audit processes, the assessment of existing evaluation metrics and the identification of applicable costs among the highest priorities. When we first began to plan a short series of UK based pilot audits Seamus identified the work being done at <ArchiveName> and recognised the tremendous experience and expertise that has been accumulated. We would therefore be very grateful if you and the Archive would agree to become involved in our efforts by providing the organisational context for a pilot audit exercise. As well as contributing to the wider international understanding of the audit process, it is expected that participating organisations themselves will benefit from a range of insights into the effectiveness of their existing processes, organisation and methods and enable them to display leadership in best practice.

The attached paper, entitled "The DCC Approach to Audit and Certification" presents some more background information about the subject area as well as some more details of the team that the DCC has assembled, its proposed methodology for conducting audits and the anticipated outputs of the process. At this stage we'd be delighted if you'd offer a general indication of your willingness to participate in this activity. It's likely that the audit would take place sometime in the next two to three months, although needless to say, we'd be able to offer a great deal of flexibility in order to best suit yours and <ArchiveName>'s interests and requirements.

I look forward to receiving your response, and will be glad to supply any additional information that you might require.

thanks in advance,

Andrew McHugh

B.3 The National Library Repository

Introduction

Founded in 1798 the this National Library had been financed by its corresponding Ministry of Education, Culture and Science since 1993. The Library received an annual lump-sum grant from the Ministry; in 2004 this totalled 31.6m. Additional self-generated income (from, for example, library passes and document provision) amounted to less than 10 per cent of annual income. Budgetary autonomy enabled the reallocation of funding to support research and development activities. This facilitated the management of the library's digital repository facility. In addition, since 2003 the library received some 1.1m per year from the Ministry for system maintenance (outsourced to IBM) and part of the staff handling the operations. In 2004 a further 1m was added to this annual grant for preservation of both digital and paper content, which incorporated around 0.2m for research and development associated with long-term preservation. In 2005 a further 900k was contributed for such research and development.

The library's relationship with two internationally established publishers had been integral to the establishment of the repository. Negotiations with Elsevier began in 1996 with the aim of acquiring the content of Elsevier e-journals to incorporate within the library's repository. An agreement was signed in June of that year to permit the library to load Dutch language journal, followed shortly afterwards by the formation of a similar agreement with Kluwer. In 2002 the Elsevier arrangement was extended to cover the entire set of Elsevier (including future published journals and those digitised as part of Elsevier's retrospective digitisation programme). Consquently, the library became responsible for preserving approximately 1500 journals, covering all areas of science, technology and medicine. Following this agreement similar arrangements were established with Kluwer Academic Publishers (2003), BioMed Central (2003), Blackwell (2004), Oxford University Press (2004), Taylor and Francis (2004), Sage (2005), Springer (2005) and Brill Academic Publishers (2005). Each of these agreements required the library to preserve that which the publishers send to the library, intended to ensure that the preserved content reflects exactly the published content.

At the end of 2005 the repository accommodated around 3,500 e-journal titles, comprising some 5 million articles and totalling around 6.3 Tb.

Mission and Mandate The library's mission statement described a very broad community of end users:

"As a national library the library provides access to everyone in the Netherlands and beyond. Within this target group the library directs its attention especially to researchers and other people with a specific interest in [national] history, language and culture in a wide international context. In addition, the library wants actively to promote its collections among the general public. This aim to be there for everyone implies and anticipatory attitude and service focused on consumer orientation and reliability."

Designated Community Content of the repository was publisher-driven, with the obligations that exists within deposit regulations operating in an alternative fashion from most jurisdictions. Whereas the traditional approach was to formalise a system of 'legal deposit', the onus in this jurisdiction was on the library itself, which was compelled to accept any received published content. Under this system, around 95 per cent of regular publishers did deposit their materials with library; this could be attributed to the strong relationships that had been fostered with publishers and the way in which the benefits of electronic archival storage had been identified and promoted.

Two categories of designated community were identified. The first was primarily publishers, what the library described as their business to business profile. This was expected to be extended in the future with the addition of additional cultural and heritage depositors; the development of formal service level agreements would enable and facilitate these emerging relationships. The second category of relationship was with end users, described as the library's business to consumer profile. This relationship was less explicitly stated, and few formal guarantees were offered to those seeking content as to what was available and the infrastructure that was available to support its delivery. Nonetheless, library users could remotely access catalogue information about publications, access resources on-site, or access faxed or printed copies of articles in libraries elsewhere as inter-library loans. In turn the library was compelled by contract to provide a 'minimal level of functionality' which included bibliographic searches, publisher publication listings at the volume and issue level, listings of issue content, article views, copyright information views article or "smaller than article components" (e.g., metadata) downloads consistent with the terms of each contract.

System Functionality and Workflow The repository system consisted of functionality for processing, archiving and maintaining e-publications and for more typical digital library

functions. The Digital Information Archiving Service (DIAS) was the core deposit system, and represented a separate and dedicated entity within the library's digital infrastructure. It was therefore not necessary to duplicate the functions like cataloguing, authentication, and search and retrieval. The DIAS functional design was based on the CCSDS Reference Model for an Open Archival Information System (OAIS). It featured functionality to facilitate the receipt and loading of digital content, its preservation and its subsequent search, retrieval and delivery. The repository's primary function was preservation, and access was in contrast a low priority. Stored content was managed for preservation and access but except in limited circumstances remote access to content (i.e. outwith the library premises) was not available. The library's catalogue could be searched from elsewhere but National Bibliographic Number unique IDs were checked to determine whether requested content was one of the few open access documents or available only internally. Publishers were granted a limited number of user accounts to access protected resources from off-site locations; authentication was performed using LDAP. IP checks ensured that unprivileged users not *in situ* within the library were restricted from accessing the content and instead were prompted to visit the library itself. The technological means to deliver content outside the library (and ensure its continued integrity) were not established.

Prior to ingest, content tended to originate on installable CDs, in PDF format via the File Transfer Protocol or on locally received digital tapes. Installable content was installed along with the necessary helper applications on a reference workstation; the ingested content was a disk image snapshot of the reference machine. PDF documents (which represented the vast majority of received content) were validated via checksums and batched for processing. Both digital content and associated metadata were ingested, with bibliographic information standardised and a unique identifier (based on millisecond-level timestamps) associated with each object. Descriptive and structure metadata was provided by the publishers.

A technical questionnaire revealed a significant organisational investment in IBM software, with the Digital Information Archiving System software at the heart of the repository, providing the breadth of its functionality. Although this relied on additional off-the-shelf IBM products (such as Tivoli Access Manager for authentication and authorisation and Tivoli Storage Manager for object management and backup) the system was designed and built specifically for the repository application according to the OAIS reference model. IBM was chosen as the supplier following a tender process on the basis of mainly functional requirements. Questionnaire responses suggested good practice with measures in place to optimise performance and capacity, mitigate risks to system security, and deal with any environmental unpredictability (UPS and climate control). Some concerns were raised with regards to several technical responses. No off site backup facilities were employed and although the library had a disaster plan at the institutional level, there was nothing in place at the repository level, nor was there anything specifically addressing ICT concerns. This was thought to

be of particular concern given the library's low-lying nature, and its propensity for flooding. That backups were stored in facilities two floors below ground, in the same building within which the repository operates raised some concerns. Repository technical staff explained that moisture sensors were installed within these backup storage facilities.

Assessment Findings and Comments The library was also praised for its commitment to evaluating the extent of its achievements, as well as the areas in which it might improve. The most recent completed assessment was undertaken by KPMG, which highlighted the lack of off-site backup facilities among its chief concerns. It was of limited, but still notable concern that the KPMG staff responsible for performing the audit were not particularly expert in the area of digital preservation, and therefore repository staff were themselves responsible for identifying and documenting a significant proportion of the points detailed in the final report. Plans for a comprehensive risk analysis investigation to be undertaken by Zurich Insurance (encompassing every aspect of the library's operation, both technical and otherwise) were also discussed during this evaluation, again underlining the library's overall commitment to excellence.

The most critical shortcoming identified within the repository was the lack of off-site backup facilities, which (particularly given the low-lying Netherlands landscape) was of some concern. Repository staff assured the auditing team that this was currently being addressed, and had been highlighted in prior external investigations independently commissioned by the library.

A second criticism was associated with the identifiers allocated to objects within the repository. Consisting of a simple UNIX timestamp generated at the moment of ingest this was considered potentially problematic if multiple ingest machines were commissioned to operate simultaneously or the procedure was streamlined to facilitate the ingest of more than one object at a rate that exceeds the timestamp's lowest level of granularity. In the former case a solution would be to add a prefix to distinguish objects ingested by alternative machines. This would not address the latter concern however, and the repository's technical staff agreed that some kind of alternative means of conceiving identifiers would be preferable to mitigate potential future problems.

A further concern was associated with the fixity information that was collected and stored within the repository, specifically CRC32 checksums. These were currently stored within the archival repository alongside additional technical metadata. System compromises that threaten the integrity of metadata or objects could in theory also prejudice the integrity of these checksums, which were principally deployed to determine when and where unauthorised changes have taken place.

The issue of software escrow was subjected to similar scrutiny during the audit, and also

emerged as an area of some concern. The repository's technical infrastructure was essentially a proprietary system, consisting of both off-the-shelf and bespoke software developed by IBM. No escrow agreements were in place which might leave the library in a dangerous position in the event of the withdrawal of the Tivoli software suite or the discontinuation of its support. Dismissing such concerns, staff explained that the issue was given significant consideration, but that IBM were deemed the only adequate supplier given their technological requirements. They perceive the likelihood of vendor collapse as extremely minimal, and irrespective of this, since the data and system software were separable, such an eventuality could be survived until appropriate alternative software became available. Additionally, it was argued that since a number of large global banks use and rely upon the same IBM software, there was sufficient international weight to ensure that IBM will continue to maintain the software in its current, or similar form. That this was deemed acceptable by the auditors suggests that the check-list was subject to varying levels of compulsion. It indicates that even where a course of action has identifiable risks, the important factor in determining a successful repository was its willingness and capability to undertake the appropriate risk/benefit assessment exercises.

In a sense related to the software escrow concerns were fears that without formal succession plans the library was exposing its content to future risk. Once again these were to an extent mitigated; the library staff argued that its legally defined mandate and obligation renders such plans unnecessary, ensuring the library's permanent existence. Notwithstanding this, some doubts continue to persist, particularly associated with those collections not subject to these legal considerations, such as international, non Dutch materials.

A final issue of concern identified was an example of system bottlenecking that was being experienced within the system during the visit, preventing the ingest of objects. This issue concerned a small script responsible for the allocation of identifiers. Since a system restart this script was no longer operational and consequently no objects could be added to the system for the duration of the problem.

B.4 The National Archives Data Centre

Organisational Infrastructure

The service itself had over 25 years of experience of managing and preserving large quantities of digital materials. Originally serving a UK University, the service subsequently grew into a regional and then national computing centre, diversifying its services to include information hosting and management, web site development and e-learning advice and training. Digital Archives was a department within the service, which remained legally part of its

parent University, but also operated as a limited company, which was owned in its entirety by the University. Most services had traditionally been provided on a commercial basis to outside parties, supported by contracts that made explicit the University's responsibilities. However, increasingly, services were being provided to the University itself. A key service was the National Digital Repository, which provides a range of digital preservation services for various customers. The contract with The National Archives to run the datacentre was the biggest by some considerable distance.

Mandate and Mission A legislative mandate was covered by the Public Records Act 1958 (as amended by the Freedom of Information Act 2000). A contract made explicit the business aims with respect to the datacentre. It stated as follows:

"The AUTHORITY's aim, as set out in its Corporate Plan 1997 - 1998, was 'to assist and promote the study of the past through the public records in order to inform the present and the future'. It's supporting aims are:

1. selection: to safeguard records covered by the Public Records Acts and ensure the selection of those worthy of permanent preservation;

2. preservation: to acquire and preserve the records that ought to be kept;

3. access: to provide access to, and encourage and promote the use of, the records"

This represented the mission statement of the the datacentre service, but since it remained inaccessible to stakeholders other than TNA and the service fell somewhat short of the expectations of the audit check-list which demands not only the existence of a mission that expresses a preservation commitment, but also its availability to depositors and other stakeholders. The datacentre website's 'About' page contained an expression of its mandate and objectives, but failed to explicitly define the legislative relationships that justified its existence. Further background was freely accessible from the web pages within data transfer overview documentation, which described in more detail the applicability of legislation, the obligations arising from it and the particular data that the datacentre was responsible for preserving. One would perhaps like to have seen this information presented in a more prominent location, encapsulated within a succinct and clearly defined mission statement. The service did have its own mission statement ("the service aims to be the preferred provider of information, communication and learning technology service across the public sector") but, while far from incompatible with the datacentre's commitment to long term data management and access provision, was hardly synonymous.

Succession Arrangements The datacentre contract described an exit and transfer strategy, and a corresponding Service Transfer Plan covered the transfer of the service to TNA or another contractor. This remained necessarily vague in places, but its existence reflects a pragmatic acknowledgement of the likelihood that the value of at least some of these data were likely to survive the the datacentre contract, or the the service itself. Return of digital objects to depositors was not considered to be particularly applicable, since TNA's archival responsibilities were permanent. Copies of transferred datasets could be requested by the department from which they originated, but this was a separate issue, and the service's responsibilities to provide this service would cease at the conclusion of the the datacentre contract. As outlined within the check-list self responses, the provision of succession or contingency plans to address the issue of wider funding cessation or legislative amendment that threatens the existence of an the datacentre contract more generally was really the responsibility of TNA, and was to some extent beyond the scope of this assessment.

Staffing and Staff Development The Digital Archives department existed within the the service's Application Services group. As well as archiving, this group incorporated the JISC Regional Support Centre for London. The Digital Archives Department's intrinsic objectives, responsibilities and service levels were spelled out quite comprehensively within the TNA contract, and given practical reality within a range of procedure manuals. A team of archivists, content specialists and software specialists, led by a department manager collectively ran the datacentre. According to job summaries, archivists were responsible for deputising for and assisting the senior archivist in all aspects of work connected with the datacentre and other the service projects and services, including transfer, accessioning, cataloguing and dissemination of electronic data and related documentation. Each one was a fully qualified archivist, supported by archival assistants who were generally from cultural heritage backgrounds, such as museums or galleries, often with experience of digitisation projects or digital imagery more generally. Content and software specialists were responsible for working within the the datacentre team, collaborating with archivists and IT specialists, liaising with TNA and government departments, providing specialist support to users and contributing to system development. Content specialists had varied skillsets, reflecting the position's invented origins. Individuals within these posts exhibited a range of experience, including database development and administration, public service in government, professional IT development, administration, auditing and systems assessment and work flow analysis and validation. Software specialists were primarily programmers. The divisions of responsibility appeared to be well expressed, and logical, and therefore the staff members available appeared competent to undertake the identified duties. The availability of both archival science and more domain specific data expertise was laudable, and to some extent enabled many of the challenges to be mitigated. Dovetailing was evident in internal

interactions between these groups. As part of a recent ISO 9000 certification process the datacentre's job descriptions were revised to correspond more closely with duties. Similarly, as part of the parent university's own job appraisal scheme, job descriptions and personal development plans were developed and reviewed on an ongoing basis.

There appeared to be sufficient procedures in place to ensure that staffing numbers were adequate. The heads of Application Services and Digital Archives, in association with senior staff, were responsible for allocating staff resources, and monitoring appeared to be undertaken to detect staffing shortfalls based on the requirements made explicit within the the datacentre contract. To this end, there appeared to be adequate staffing provisions at the time of the audit.

Professional staff development within the service was covered by both in house training policies and procedures, and by the wider infrastructure provided by the parent university. With respect to the latter, job appraisal schemes provided opportunities for staff and line managers to jointly mould personal and professional development, to identify their own training needs and to arrange for them to be formally addressed. A training budget existed for the Digital Archiving Department and given the involvement of the service in a variety of other training activities there were ample opportunities for specific training. Furthermore, the datacentre's Inhouse and Training Procedure Manual described fairly comprehensive processes associated with staff induction, training needs assessment and review and training delivery. New staff were presented with a staff handbook, which described benefits, facilities and responsibilities in fairly wide terms. They then received a tailored induction process, which was generally delivered verbally and incorporated discussions with other the datacentre staff about aspects of work and procedures implicit within the repository. Following appointment, new staff were interviewed as part of their probation review; steps were taken at this stage to identify skills shortcomings, which were made evident by comparing existing skills with requirements defined within a relevant job specification. Training requirements were then documented and planning was undertaken for meeting these as part of probation. Training needs were also assessed on an ongoing basis (at least annually) to determine further emerging training requirements and staff members can suggest training that would help with their job at any time. Also, if new procedures or responsibilities were introduced for any job an opportunity was taken to assess whether additional training was required.

Training was delivered in a variety of forms, including, but not necessarily limited to:

- Inhouse training by fellow staff members
- Inhouse training by an outside training provider
- Outside training

- Attendance at regular courses organised by professional bodies
- Membership of relevant professional committees and working parties
- Production of reports for inhouse circulation by staff who have attended training courses and professional meetings
- Training of staff who have not attended courses by those who have

The range of training types and the monitoring infrastructure in place appeared more than adequate, and interviewed staff spoke highly of the procedures that were in place. The fact that training requirements were being identified by both staff and their line managers ensured that few opportunities were missed.

Designated Community The designated community served by the datacentre was defined both legislatively, and in extremely general terms in the 'Help' section of the the datacentre website. The Public Records Act of 1958, as amended by the Freedom of Information Act 2000 described the following:

It shall be the duty of the Keeper of Public Records to arrange that reasonable facilities were available to the public for inspecting and obtaining copies of those public records in the Public Record Office which fall to be disclosed in accordance with the Freedom of Information Act 2000.

"The Lord Chancellor shall, as respects all public records in places of deposit appointed by him under this Act outside the Public Record Office, require arrangements to be made for their inspection by the public comparable to those made for public records in the Public Record Office"

A broad designated community was therefore established, encompassing the public in its widest sense. The web definition alluded to the reasonable facilities that were in place to comply with this legislation, stating that users must have a compatible web browser and at least a rudimentary knowledge of how to use it. There were also accessibility provisions made explicit within the website, although one can contrast usability in a web interface from the issue of understandability, as expressed within the Reference Model for an Open Archival Information System. In a pre-audit correspondence, the department manager suggested that this web definition also explicitly mentions adults with a reading knowledge of English, but this could not be found. During interview it was suggested that data were made available in an identifiable form, although no assurances were offered as to its usability or understandability. There was little evidence within any of the specific the datacentre documentation of a designated community definition; the service level agreement between the service and TNA described various responsibilities implicit within the the datacentre contract but none really

associated with the specific communities that must be able to use data within the archive. It was suggested that notwithstanding the breadth of its designated community it would have been prudent to publish, probably within a mission statement, the fact that the datacentre was legislatively bound to serve the public as a whole, and that therefore represents its designated community. It would have been similarly worthwhile to make explicit the data producers from whom public records originate, which was outlined in the amended legislation. Formal feedback mechanisms (including regular appraisals of server logs) indicated that the datacentre's users were using both the data and the catalogue information which indicated a more historical interest in the fact that data existed at all, as opposed to the specifics of that data.

An interesting stance related to the designated community was presented by one staff member, who described the fact that the datacentre was ultimately bound to reflect TNA's own policies on data accessibility, and restricted in terms of the strategies they may implement to facilitate discovery. The datacentre's own catalogues for instance were required to be broadly compatible with the Catalogue. It was the staff member's contention that defining and monitoring the designated community was the sole responsibility of TNA. This was worrying because irrespective of where the parameters of the community were determined, archival decisions within the datacentre should have been based at least partially on their expectations, capabilities and knowledge.

Policy and Procedures The datacentre had a range of policy and procedural documentation available. Formally documented procedures included:

- Security Procedures
- Transfer Procedures
- Site Exchange Procedures
- Digital Preservation Procedures
- Paper and Paper Preservation Procedures
- Helpdesk Procedures
- Inhouse and Training Procedures
- Closed Data Access Procedures
- Data Protection Act Procedures
- Finding Aids Procedures

- Contingency Planning Procedures
- Style Guides (issued by TNA)

Each set of procedures was realised in one or more associated documents; this is an extensive list.

The datacentre also maintained policies that described in detail necessary steps to introduce, review and retire procedures within the archive. As outlined within the *Inhouse and Training Procedures Manual*, all the datacentre procedures were documented in appropriate procedures manuals, supplemented where necessary by detailed working instructions. These could be changed in three ways. The first was where the datacentre service manager had identified the need to revise a procedure manual, and this duty could be delegated to a member of staff. A draft revision was then presented and discussed in a physical meeting, by email or in the datacentre usenet discussion. Amendments were then actioned, prior to the creation of a final draft, which was then approved by the the datacentre service manager, and linked to from a central HTML index page. Staff were informed that all prior versions should be immediately disregarded. Alternatively, any staff member could suggest changes at any time. Suggested changes were circulated via email or usenet, comments were aggregated and a brief report conceived, for discussion at a subsequent meeting. Final revisions, and the replacement of earlier versions were actioned as above, subject to the service manager's approval. Finally, all procedures manuals were subject to ongoing review, on an at least annual basis. The introduction of new procedures was conducted on a similar basis. Once more, these could be prompted by the service manager, an independent member of staff or during a regular procedures review meeting.

The quality assurance and consultation procedures associated with the conception of new procedures were broadly equivalent to those associated with amending existing procedures. A new procedure could be justified by the introduction of a new procedure that extended the range of work; a major change in the way a procedure was undertaken; sufficient numbers of small changes to an existing procedure to necessitate a wholesale review or its granularisation into multiple procedures; or, finally, the insistence by a staff member that a new procedure was otherwise necessary. In most cases of procedural change, the existing manuals would simply be updated. The benefits of this approach to procedures management were clear. The availability of each of these three methods for introducing and modifying procedures were intended to ensure that they remain relevant, representative and comprehensive in their coverage. Procedures were allowed to both dictate the work undertaken within the datacentre, and reflect emerging working practices that may reveal themselves and optimise the repository's efforts.

Further policy describes the mechanism for retiring redundant procedures within the datacentre. Again, this could be prompted by any staff member, or within discussions within

a procedures review meeting. In the event of a procedure being nominated for retirement, a staff member was delegated the responsibility for collecting comments to support or oppose the motion (or suggest merely revision of the procedure). These were amalgamated into a report which provided the basis for subsequent discussion and a final decision by the the datacentre service manager. Retired procedures were moved to a special section of the staff intranet, with a note describing the fact that the procedure did once exist and had been superseded. Details of any procedures manuals that did supersede retired procedures were also recorded. Procedures could be retired if they were classified as redundant; this could incorporate situations where working practices had changed to the extent that the procedure was no longer relevant; staff hade suggested that the procedure was no longer relevant; procedures corresponded to work areas that were no longer active; sufficient smaller-scale changes necessitated reformulation of policies or the creation of more granular policies; or where procedures were subsumed within an existing procedures manual.

There could be little doubt about the datacentre's commitment to ongoing periodic review, assessment and self measurement. As well as participating in this pilot assessment, perhaps the most convincing evidence of this was the service's completion of the ISO 9000 series quality assurance certification.

Costs and Financial Information On occasions, costs associated with the accessioning of datasets led to the datacentre's resistance to archive. Data had been turned down in the past due to unnecessary expense associated with it a prominent example was a dataset encoded in a proprietary format associated with an unnamed document management system. Mechanisms were offered by the system's developers to export these documents to PDF, but the datacentre was unwilling to pay the charge. Had TNA insisted, it was acknowledged that the datacentre would have had to go ahead, such were the terms of the contract. This was to some extent worrying, although it appears that the relationship that existed between the service and TNA would make it unlikely for TNA to insist that transfer should take place where it might significantly undermine the financial position of the service and the datacentre.

The service's turnover associated with digital preservation amounted to approximately 1m, about 20 per cent of its overall turnover. The datacentre contract was negotiated at a fixed price, which was adequate to meet most costs, although in some circumstances where the service exceeded expectations there were additional costs that had to be met. History suggested that increased costs would be met with favourable terms in subsequent contract renegotiation. The original datacentre contract was priced too low; the service was losing money and forced to rely on its own additional funding reserves. A subsequent contract compensated this loss and acknowledged the increased cost of providing the service.

The datacentre was required to meet performance targets in order to ensure its financial remu-

neration. A system of service credits was made explicit in Schedule 11 of the the datacentre contract. TNA could waive its right to reduce the fees payable in the event of such circumstances, and there was no liability for degradation of service caused directly by a failure of TNA or transferring departments, assuming the failed responsibilities were previously agreed in writing. Potentially costly failings could include less than satisfactory service availability, failure to meet accession timetable agreements, inefficiency in satisfying access requests and network non-availability. These provisions were generally fair and thought unlikely to prejudice the datacentre's budget. A maximum of 10 per cent of fixed charges for any three month period could be deducted according to this agreement.

Legal Issues The contract that the service had with TNA afforded the former a degree of protection, since the liabilities were expressly indicated within that contract, and many implicit issues were TNA's responsibility. Similarly, questions about the the datacentre mission being at odds with the service or the parent university were largely moot since the University had legitimised the contract, with the vice chancellor signing the contract and formally expressing his satisfaction of the datacentre's alignment with the University mission.

At the end of the the datacentre contract the service was required to return the content to TNA and destroy any copies that might continue to exist on their own systems. Either TNA or the service could choose to back out of the contract giving a minimum of six months notice; there were directions to follow in the event of this happening, but specific details would be negotiated at that time. Generally speaking, conflicting contracts with the commercial sector would be avoided by the service.

The datacentre maintained ongoing relationships with both government departments and data owners (as well as client and contract managers at TNA) to ensure that its procedures were endorsed where appropriate and that it was ultimately able to adequately fulfil its mandate. Dataset transfer forms changed hands during the initial stages of transfer, following the notification by TNA of data that was to be preserved. Signed by data owners and departmental records officers these provided the means to issue formal authorisation to transform source data, detail parts of data sets that must remain closed or be redacted, and describe conditions for managing transport media. Subsequent receipts issued by the datacentre further formalised the agreement that preservation would take place, and confirmed the instructions issued by owners and departments. An accessioning tracking system monitored and maintained a record of every interaction between parties and interaction with data.

Intellectual property rights were unlikely to concern the datacentre too much since it was dealing with public records with an explicitly expressed legislative mandate. Nevertheless, evidence highlighted a concerning shortfall in policy in the event of an intellectual property rights challenge. One staff member recounted in a checklist self-response that the datacentre

had been challenged in the past and had redacted data, in the absence of a suitable policy saying otherwise. She expressed some (albeit tentative) concerns that perhaps everything might be redacted if challenged. This should have been addressed by both the development of formal documentation describing a policy for this situation and internal awareness raising to communicate more clearly the legal status of the datacentre records. It seemed unlikely that the datacentre would be conforming to its contractual requirements if it acted unilaterally on this occasion, and therefore an expression of TNA policy in this area was probably quite adequate. In fact, in this case the approach went via TNA who oversaw negotiations prior to making a decision. Further ambiguity surrounded challenges to non-availability; for example, a database containing fifty year old information about beer duty was deemed by the Department of Trade and Industry to be too commercially sensitive to release but may be covered by FOI. It appeared that there has been no retrospective assessment of previously closed datasets in light of FOI, and this was something that could have been considered. FOI requests remain primarily the responsibility of TNA, to whom the request should be issued. Following a representation by the relevant government department, TNA makes a decision as to whether content should be released.

Further legal complications arose as a result of some records within the datacentre being exempt from the Public Records Act, due to the fact that they were not in fact Crown Copyright. For example, materials originating from the Coal Authority must be preserved only under explicit license. Furthermore, on some occasions parts of certain accessions would be subject to intellectual property law. For example, software user manuals have been submitted in the past as part of a dataset's accompanying documentation, and this introduced some ambiguities that could have been addressed within a formal policy document.

Digital Object Management

The datacentre responsibilities were summarised during the audit meetings as follows:

- Arranging Transfers from Government Departments
- Confirming the Receipt of Datasets
- Aggregating Contextual Material
- Preserving and Describing Datasets
- Providing Access to Datasets (where they were to be made available)

Acquisition and Ingest The datacentre's duties began when datasets had been identified for transfer by TNA working in association with Government Departments. Dataset transfer prompted a variety of datacentre processing activities, while preservation characterised most data interactions. Archival activities included acquiring documentation, dealing with government departments, cataloguing datasets, and specifying access conditions and culminate with the upload of catalogue information. Concurrently, data specialists assumed responsibility for ingesting datasets, performing analysis, transforming and converting data, documenting data and exposing data to validation and checking procedures, before the data were uploaded. Meanwhile, digitisation activities were also ongoing, including the scanning of paper documents, their conversion, processing and checking, and eventual upload. Once catalogue information, data and associated documents were uploaded, they were subject to final checking prior to being made available as live items on the the datacentre website. Administration also played a key role, as the issuing of transfer receipts, establishment of dataset-specific targets and maintenance of liaisons with TNA legitimised the process. The final group of responsibilities were developmental, with software support, tool design and process development aimed at facilitating and improving every other aspect of the datacentre's activities.

Every stage of dataset acquisition was recorded within the datacentre's Accession Tracking System (ATS). Its role was to document all events that related to individual accessions. These included communications that took place surrounding the dataset (whether internal or with data owners, government departmental records officers or Client Managers or Contract Managers at the National Archives); suspensions on the accessioning process, for instance where the government departments' inactions result in the stalling of the process; and the final public release of the dataset. This provided provenance and traceability up to the point of the dataset's dissemination.

The procedure to be followed for dataset transfers was documented mainly within the *Transfer Procedures Manual*, which made explicit procedures for initiating dataset transfer, appropriate communications that must be undertaken, physical transfer procedures, checks, documentation and receipts that must be issued. Templates for various documentation that were required to be exchanged throughout this process were available from the the datacentre website, and also within its internal electronic filing system. As detailed above, The National Archives was responsible for the appraisal and selection of datasets, which was undertaken at the level of individual datasets, and the transfer process was prompted by the TNA's written notification to the service that a dataset has been identified for transfer. This notification contained information gathered during TNA's appraisal activities, and was accompanied by a notification form. Usually, transfer would be expected within the current contractual year. Occasionally, TNA would request the transfer of content that did not actually exist. The datacentre would then communicate this error. When severe problems arose with datasets

the datacentre could petition to TNA for the transfer to be abandoned. Ultimately, TNA had the discretion to compel transfer irrespective of concerns. Datasets that contained unreliable or unuseful information were likely to be accepted, and documented nevertheless. The only absolute condition that the datacentre imposed with respect to the transferred datasets was that they should be accompanied by the appropriate, signed transfer forms.

The service's initial responsibility was to document on transfer forms any information already supplied by TNA about the dataset, including appraisal information. A Departmental Records Officer (DRO) within the transferring government department was contacted and issued with a copy of the DRO transfer form and transfer list, generally within two weeks of the issue of a transfer notification. An accompanying letter stated that completion and return of the transfer form indicated a readiness to transfer a dataset to the datacentre and incorporated guidance notes or a reference to a location online where these were available. Concurrently, the service sent copies of both a Data Owner Transfer form and the transfer list to the relevant data owner, along with a covering letter, unless explicitly instructed not to in TNA's originally submitted notification form. In the event of such an instruction, both DRO and Data Owner forms were sent to the Departmental Records Officer. Government departments were not required to provide finding aids for electronic records, and therefore the service was required to prepare these based on the information provided within transfer forms.

Government DROs and data owners could complete transfer forms themselves, or with the assistance of relevant individuals such as Information Systems or IT staff. Documentation was aggregated by the relevant department and if available only in physical form packaged in containers to be sent to the datacentre. Upon receipt of transfer forms the service could request further information. Under normal circumstances, both DRO and Data Owner transfer forms were required to commence transfer; the receipt of the former indicated an authorisation for transfer to take place, and was required to have been received, except in exceptional cases. No formal documentation described such circumstances however. Data Owner forms made explicit more technical details, which was of particular value for new Series, and essential in such cases. However, it was conceded that for repeat transfers it could sometimes be acceptable for just the DRO form to have been received. The exceptions that permit this should have been made more explicit, or if down to a human judgement call, then this should be formally stated.

A number of media types were suggested as appropriate for transfer, and these were made explicit within the explanatory notes document that accompanied the Data Owner transfer form. 'Approved' media were:

- 9-track reel tape (6250 BPI GCR encoded and 1600 BPI phase-encoded)
- CD (High Sierra format or ISO 9660 format)

- 3.5" floppy disk and 5.25" floppy disk, in MS-DOS, MacOS, VMS or multi-volume TAR format (as produced by GNU tar)
- Exabyte 8200 and 8500 format 8mm tape cartridge (2.2 GB and 5 GB versions only)
- DDS2/3 DAT (4mm) tape cartridge, 4GB and 12GB types, but not including older DAT tapes that use proprietary methods to exceed 2 GB of storage capacity
- 3480, 3490 and 3490E 0.5" cartridge tapes which hold 200 MB, 400 MB and 800 MB respectively

Despite these strictly defined parameters, the transfer procedure did maintain that alternative media could be suggested by the owning department when the transfer form was returned, and the service would subsequently deem this acceptable or otherwise. The overwhelming majority of data arrived on CD or some kind of magnetic tape, with some low volume and non-confidential material also appearing by email. It was not clear whether integrity checks were undertaken prior to delivery to ensure that what arrives at the datacentre corresponds to what left the government department. In one staff member's checklist reponse it was noted that objects were manually or electronically checked to ensure that the contents match the information in the "Datasets Transfer Form", but no detail of how this was done was offered.

Following these stages the transfer would begin in earnest; departments packaged data and associated documentation along with accompanying transfer lists. Dataset documentation and associated documentation were packaged separately. Each box was marked with reference codes and titles of the datasets contained within, and numbered, starting at 1, and describing the total number of boxes (e.g., 1 of 5, 2 of 5, etc.). Although numbering was not applied on a per item basis, the transfer list documented each item being transferred. Help and advice about packaging and transporting datasets was provided to departments by the service.

Upon receipt of the datasets, documentation and accompanying transfer paper work, the receipt was logged in the Accession Tracking System and items placed within the the service's paper store. This was be the permanent environment for paper dataset and associated documentation. For electronic materials, it represented a temporary home, prior to their accession into the hierarchical storage management system (HSM). Security was maintained in this interim period by requiring the recording of any media movements within a loans register. Items were verified against the transfer list to ensure completeness and an initial receipt was sent to the transferring department. This was a confirmation that content has been both transferred and received, that documentation appeared to be adequate and that preservation would be undertaken.

Following the issue of this initial receipt a target date was set by which time the dataset would be completed in accordance with Schedule 11 of the the datacentre contract. The

datacentre would then continue with the creation of finding aids; these were based upon a combination of appraisal information supplied by TNA, the information on transfer forms and other information arising from the dataset and its documentation. A second receipt was issued to transferring departments to confirm that processing has been completed, all rights had been agreed and that the datasets were to be released.

Every transfer undertaken by the datacentre was allocated a processing record, and following the initial notification by TNA, responsibility for each datasets was allocated to an individual, named data specialist. This individual need not have been the sole contributor to this dataset's transfer, but was responsible for ensuring that resources were allocated, procedures were followed appropriately and that the processing record was completed. Generally, the same individual would be responsible for subsequent dataset processing too.

Each accessioned dataset was allocated with a reference code to uniquely identify it within the repository. This took the form CRDA/n/aa/m where:

- 'n' was a number representing all occurrences of a particular datasets (irrespective of whether these were received in one batch or incrementally), allocated in the order in which datasets were notified
- 'aa' was an alphabetic code indicating the format of the records listed, including DS for a dataset, and parts thereof, DD for dataset documentation, and AD for accompanying documentation
- 'm' was a numeric value indicating that the dataset was one of a series, which may be distinct as an annual dataset, or by its type, by snapshot, by other modification, or by sub-series.

Following copying to the HSM transfer media was disposed of. As outlined in the *Digital Preservation Procedures Manual*, departments would have outlined at the time of transfer whether they wanted source media to be erased and returned or securely destroyed. Irrespective, except in the case of non-rewritable media (e.g., CD-ROM), erasure was the first step. Procedures demanded that staff check, and recheck with the assistance of a colleague to ensure they were dealing with the correct volume. Once satisfied, media specific erasure steps were made explicit within the documented procedures. Media were then either returned in their original or similar packaging, or alternatively, irreperably destroyed using the the service's approved destruction kit (strong sacking, gloves, a hammer and goggles). Both returns and media destruction were subsequently recorded within the locations register.

Procedures for ingesting bitstreams into the the datacentre repository were outlined within the *Digital Preservation Procedures Manual*. Like each of the datacentre's digital preservation procedures these were generally carried out by the data and software specialists within

the the datacentre team, occasionally with the support of systems staff (who for instance were always responsible for media checking). Security was a key consideration throughout amid concerns that sensitive data within even open datasets might required anonymisation. Amendments or deletions were not made without the explicit prior permission of TNA. All material was required to be preserved, although its form was permitted to change as part of the transformations that were implicit within the preservation process. As a general security measure, policy demanded that media write protection mechanisms were enabled at all times.

The creation of 'bit-wise' copies of source data was the first step that immediately followed the successful physical transfer of datasets. These were stored on a server in a filesystem that was designated for the storage of incoming content. Thereafter, it was these copies that were subjected to subsequent transformations or further processing, rather than the source originals. Numerous devices were available to read media although only those within the secure network and equipped with necessary software and validation tools were deemed suitable. Policy demanded that sys-admin staff oversee the use of appropriate drives and systems for this purpose. The fact that only those devices deemed suitable were actually capable of writing to the server provides an additional technological control. Formalised procedures also acknowledged that the creation of copies could be time consuming and staff were therefore required to follow documented security steps when leaving their computer during the process (using the xlock or similar screen locking programs was compulsory while terminals were unattended). Basic characteristics of the copied data were required to be checked following their accession (e.g., file size and content type) and any discrepancies noted in the dataset's processing record. Transfers over FTP were performed in binary mode to avoid undesirable modifications. Virus checks were implemented with procedures in place that describe the appropriate steps to take in the event of the discovery of a virus. In the event of a virus discovery its existence and any steps to remove it were recorded. The documentation suggested that TNA approval may have to be acquired prior to virus removal if normal transformation steps would result in virus removal as a convenient side effect then such authorisation was unnecessary, although it was acknowledged that it was unlikely to be withheld.

Transformation and Preservation The point of transforming data within the datacentre was to regularise its form to facilitate access and usability. Target formats were chosen based upon their amenability to subsequent conversion as part of ongoing preservation, and their ability to preserve the content and intellectual ordering of the original dataset. Datasets were organised and documented to facilitate the representation of their implicit information, and not necessarily to reflect their form when they arrived. Nevertheless, documentation made it possible to trace back to see the structure of content upon accession. All steps to transform the data were recorded, with the procedures demanding that staff were satisfied

that not only could they repeat the process exactly with only the original data, their description of the process and the metadata that accompanies the dataset, but that a different staff member could do the same.

The initial stages of data transformation required staff to document the source dataset's structure, as well as the content and format of each field within. Following their initial assessment, data specialists were required to formally document their anticipated actions within an 'Approach' document, to be evaluated by other members of the team.

Where content arrived in popular formats such as *Microsoft Access* .mdb files documentation was fairly straightforward to automate. In other cases, where more proprietary or bespoke formats or data structures were employed it could be necessary to use more labour intensive techniques, such as text analysis of data documentation or alternatively manual keying. In even more complex cases, it could be necessary to reverse engineer software to retrieve a description of data structures. It was not clear whether the potential legal implications of such techniques had been formally explored, but it was suggested that this should be done and documented with some priority. The end user license agreements of software vary, but a policy statement encouraging staff to investigate their rights with respect to such procedures was considered appropriate for inclusion in the *Digital Preservation Procedures Manual*. If reverse engineering failed to yield a clear definition then raw data analysis was undertaken, using tools such as od, or the datacentre's own flook. Concerns surrounded such procedures which amounted to little more than (highly educated) guesswork, but given the shortcomings implicit within the received data, and the pressures placed upon the datacentre to archive whatever they were given, this was probably unavoidable from time to time. It appeared that communications with departments and data owners were suitably extensive to limit the likelihood of these circumstances in almost every case. Irrespective of which methods weere employed for documentation, the required elements remained consistent. Characteristics that were documented for each accessioned dataset included:

- File layouts
- Record Structures
- Field formats (including field widths, repeat counts and relationships between fields (e.g., whether they were keys or indexes)
- Field descriptions usually one line descriptions, explaining what a field was for, and also documenting any ambiguity that might surround the data specialists's interpretation

Data were also anonymised at this stage if there had been indications from the transferring department that this was necessary. The transferring department would indicate in their

initial correspondence how this anonymisation should take place (summarisation of data or suppression of certain fields). An *Anonymisation Procedures Manual* contained detailed descriptions of the process that should be followed.

Following analysis, description and the completion of any required anonymisation, data specialists decided upon the form within which the dataset would be preserved. It was not necessary to maintain the original table structure, and it in some cases could be desirable to normalise if this has not already taken place. Any temporary, or redundant tables could be discarded, but their prior existence was recorded so that it was possible to maintain an understanding of the form of the data at the point of its transfer. Any proposed conversion or disposal was expressed within the 'Approach' document detailed above, a discussion group or in direct conversation with the service manager. These modifications were approved by TNA prior to their execution.

Departments were empowered by the Public Records Act to request copies of their transferred data. Therefore, the original bitstream was copied onto new media and also preserved. Indistinguishable from the packages that represent the datacentre's submission information packages, these were probably classifiable as separate Archival Information Packages. The right of departments to recall their content under the Public Records Act was unlikely to preclude the archive (or the custodians in this case), from retaining their copy, although the issues of possession associated with digital content were quite different from those associated with physical materials.

Of some concern were situations where coded values could not be transformed, or indeed translated, since information (such as lookup tables, or references linking data to existing lookup tables) had been omitted from the documentation supplied by data owners or departments. In such cases the datacentre staff were required to simply deduce the meaning of these codes.

Content Validation Validation was a mechanism to ensure that transformations were true and accurate, and sufficiently representative of the original source dataset. The identification of inaccuracies or deficiencies in the original data could also be found at this stage and brought to the attention of researchers or data users. Two terms, 'transformation validation', and 'content validation' were internally coined by the datacentre to describe the two types. Software was available to perform content validation on various types of data, performing a similar role to the US National Archives and Records Administration's ERIC tool. It validated against metadata descriptors to ensure that content within database fields corresponded to the documented schema, and if so, that metadata was retained with the preserved dataset. For instance, it was capable of checking that columns that ought to be dates were dates, and those that ought to be integers were integers. The tool was also used to

automate the creation of data description metadata based on the characteristics of the data where none previously existed. Measurement checks ensured that content corresponded to that described in transfer forms, accompanying documentation or in any other referenced publication. This typically compared averages, counts, or other quantitative characteristics of the dataset with this evidential information. Results of each of these checks were recorded within the dataset's processing record. Irrespective of how poor the results of these checks were, the datacentre had a policy not to change data, even if errors were obvious and straightforwardly correctable. All such problems and inconsistencies were documented. Only once exception to this had been documented, and related to corruption prior to accession by the datacentre that prevented data processing. The dataset in question had relied upon fixed-width fields to distinguish individual content fields and the corruption had misaligned the data, with significant effects, whereby some closed data (i.e. not for general viewing) had been shifted into open field positions. This was therefore repaired. It would have been useful to have a more formally expressed policy to document the circumstances within which such interventions would be permissible. Documentation described an example occasion where intervention was legitimate, but it was suggested that the datacentre extrapolate this into a more generally applicable policy statement.

Immediately prior to committing a dataset to permanent preservation storage, responsible staff submitted it for review by a fellow the datacentre staff member. An individual was appointed with responsibility for data checking, although in the event of his/her non availability other staff might have been required to provide this final quality assurance input. Typically, the review comprised checks for consistency between the transformed dataset, the original source and associated documentation, completeness and accuracy. Documented procedures supported random checking by senior datacentre staff, and mechanisms were in place to involve more than one individual in this final review process (up to the entire the datacentre team) for datasets identified as presenting particularly challenging problems or layers of complexity. The datacentre *Checking procedure checklist* was completed immediately prior to the dataset being committed.

Media replacement took place in one of four circumstances. The first was the result of ongoing activities, with the latter three reactions to atypical circumstances that might arise.

- A tape has reached its maximum usage count (set at 10,000 mounts) or age (set at 7 years);
- A tape has been damaged due to hardware failure;
- Unanticipated readability problems;
- Other failures or the procurement of information that suggests a tape or batch might be suspect.

Automated checks within the system monitored usage and tape age the current values were subject to review on an occasional basis, and systems staff were granted suitable discretion to retire tapes prior them reaching the maximum age or usage if more convenient.

When faced with errors or discrepancies administrator made a judgement as to whether it was the media upon which data reside that was to blame. Suspect media was disabled from interacting with the wider system until this judgement was made. When the media itself was found to be at fault a procedure existed for media replacement. In the event of above average media failures administrators were expected to pursue with manufacturers the possibility that a batch was affected with a common fault. Where hardware was identified as being at fault staff would liaise with engineers from vendor who would perform appropriate maintenance and corrections. Consultation would follow to determine whether any media on a failing hardware drive might have been affected.

Initially, administrators would determine from the system which files were stored on the tape that was set to be replaced. Each of these files was recalled to online storage, and verified to ensure their integrity has been maintained. These files were then copied to a new tape, and the system was instructed to disregard the previous tape. Media retirement was recorded within the media register, and the media was then erased and destroyed (presumably according to the guidelines expressed within the *Digital Preservation Procedures Manual, although this was not made explicit*).

Preservation Technologies According to the *Digital Preservation Procedures Manual*, the datacentre's software handled binary or mixed format best when dealing with fixed-width character fields and/or numeric or date fields. An alternative approach was the use of CSV (comma separated values) files, with data represented entirely by text. The latter approach was used in circumstances where data was unclean, and content did not conform to data types that correspond to particular fields, something that most RDBMs will not tolerate. Character data was always converted to extended (8-bit) ASCII, but this could introduce some problems depending upon the encoding of source data, with some glyphs inconsistent across different platforms, and some not available. Staff were required to use all available evidence (e.g., any printouts of original data that were available, other contextual material that exists) to ensure an appropriate representation was used. Similarly, mappings were isomorphic to ensure that even if 8-bit ASCII was incapable of rendering a particular glyph, it maintained its distinct status and could be reverted to its original encoding, with its meaning retained.

For binary files, integer data was stored in twos-complement integers of 8, 16, 32 or 64 bits. Unsigned integers in source data were changed to positive integers using a larger storage width if necessary. Floating point numbers were stored using in the IEEE format using dou-

ble precision (64-bit) unless the original used IEEE single precision, or a less precise IEEE variant. For precise dates, the ACM jday algorithm was used to create 32-bit Julian day numbers. Two digit year representations were avoided. Imprecise dates which exhibit inconsistency of expression (e.g., some tuples contain a year, others a month and year) were preserved as just character data, as described above. Binary object data implicit within datasets was extracted and packaged on a per table basis using tar, with a reference to each item within stored in the database. Each binary file was named using a fixed width integer sufficient to name all the files corresponding to a single table (e.g, up to a thousand images would use the names 000.tif, 001,tif and so on). TIFF was suggested as the appropriate format for images (unless the original was JPEG encoded, and could be left as it was) and NeXT .au format appropriate for audio. No other formats were suggested for alternative multimedia types. The *Procedures Manual* did make explicit the immaturity of the policy in this area, and that the light of experience would inform the development of more specialist treatment and guidelines. Nevertheless, the procedure did highlight the possibility of ingesting word processor documents, spreadsheets or Powerpoint-style presentations as part of submitted databases. It was suggested that a policy about the appropriate file format for maintaining these types of resources should also be made explicit, even if it might be subject to further refinement. Irrespective of whether government departments embed binary content directly into databases or store links to content in the form of file-system references, it was the responsibility of the datacentre to preserve such assets, and therefore a policy for managing potentially diverse file formats was necessary. Policies for the transformation of digital documentation were quite explicit, and correspond closely to those applicable to datasets.

All digital documentation was retained in a plain text version. In addition, rich, word-processed files were also retained in pdf or image versions. There was no explicitly preferred choice for particular circumstances, or documentation describing what image format(s) were considered suitable. Postcript encoded documentation was stored in both its native format and a TIFF version, to facilitate text extraction, using optical character recognition. All available metadata was extracted from files prior to their regularisation. Metadata was generally encoded within a MySQL database or in XML, in conformance with the Encoded Archival Description standard's Document Type Definition.

As a member of the Digital Preservation Coalition, the datacentre was exposed to emerging preservation trends and performed an active role within the community more generally. On a six monthly basis, according to the contract with TNA, the datacentre was required to flag any concerns of a technical nature, which provided further assurances that current awareness in the area of preservation was maintained.

The datacentre's concept of understandability appeared to be mainly related to accessibility, with the only cited examples concerning format conversion based on end user feedback. Less emphasis appeared to have been placed on more semantics-oriented issues of understandabil-

ity that might rely upon specialist knowledge and levels of understanding that may evolve over time. Representation information stored by the datacentre amounted to just finding aids and preservation or descriptive metadata.

Access and Dissemination The web site provided a range of functionality including extensive browse and search (including advanced search) features. As described above however, there appeared to be some restrictions imposed on the datacentre in this context, since their access mechanisms were required to conform to a TNA specification. Various metadata were created to facilitate discovery, although omissions in technical metadata (which must be inferred from initial deposit communications) were noted.

Referential integrity between archived objects and descriptive information was established with the creation of links between the ISAD(G) cataloguing and CRDA dataset identifers, with the access mechanisms also creating unique HTTP URIs that corresponded to individual APIs.

Access to the datacentre's archived content was almost exclusively via its website. Until the introduction of the Freedom of Information Act in 2005 the distinction between closed and open content within the archive was quite clear cut. Since then there is more confusion. For instance, statutory bar, the mechanism that enables government departments to collect sensitive data with the proviso that it may not be used for purposes other than its original stated one might no longer apply to the same extent, or at all in some circumstances. The *Closed Data Access Procedures Manual* describes the procedures for:

- Datacentre staff access to closed data
- TNA staff access to closed data
- Departmental staff access to closed data
- Privileged access by members of the public to closed data
- Reasons for which access might be granted to any of the above
- Means for establishing authorisation for access to closed data
- Permitted mechanisms for accessing closed data
- How must such accesses be recorded

It was of some concern that the most current version of this procedures document predated the introduction of FOI legislation. Enough time ought now to have passed to issue a revised version that corresponds more closely to the requirements set out in this legislation. There

was an awareness of the implications of FOI expressed within this documentation, and of the possibilities that policy revisions might be necessary ("further updates to this manual covering the implications of the Freedom of Information Act may be required prior to the coming into effect of this Act in January 2005"), but at the time of the assessment updates had not taken place.

'Closed data' covered any data, documentation or associated material held within the Archive that was to be withheld from public access. Generally, this would only apply in situations where a Freedom of Information exemption applied. The thirty year access rule for public records no longer applies (although the *Closed Data Procedures manual* erroneously included this as a possible justification). Where FOI exemptions did exist these were communicated to the datacentre at the time of transfer by the data owner or departmental records officer. Various options were available, including closure of an entire dataset, closure of one or more of its tables, closure of selected fields, aggregation of data to a higher level or any combination of these. Data was nevertheless transferred in its entirety, with the datacentre taking the responsibility to close the relevant sections as instructed. A cron task executed a script that opened closed datasets when their closure period expires. It was unclear following FOI whether this script was still active or necessary.

The datacentre did operate a user registration service, formerly akin to an archive application process and a required precursor to accessing content. With the advent of FOI however, users were required to register only in order to take advantage of additional website features.

At the time of the evaluation, various alternative access provisions had been considered, following feedback from the datacentre's designated community. For example, the format that databases were disseminated in had been subject to some discussion, with suggestions including the adoption of an XML Document Type Definition to support database markup, or the introduction of an SQL export function so that users can more straightforwardly use the data within their own local RDMBS.

Within the the datacentre system, DIPs were created dynamically, constructed from the corresponding database or flat CSV file to provide an interrogable, web accessible dataset. Catalogue data was available alongside datasets, encoded within HTML pages. Individuals very rarely requested their own copy of archived datasets, and instead the web interface appears overwhelmingly the most popular means of accessing the datacentre's content. Nevertheless, if tables were required to be delivered in an incomplete form (either due to legal restrictions or to conform with a specific sub-set request) this could be be done. A checksum was created at the point of the DIP's request, which was intended to ensure that it was both complete and correct with respect to the request issued. Similarly, the original SIP bitstream could be requested when it was of value. An example offered during discussions was of certain Geographical Information Systems (GIS) datasets that could not really be understood when

converted from their source form and separated from the application that created them.

In technical terms, the system that stores the archived materials had no direct contact with the outside world. The datacentre web server operated as a client to the archival servers, with limited, read only access. Therefore if the web server was compromised the extent to which a malicious individual could damage the archival storage component of the system was limited.

The service's expected performance levels were made explicit within Schedule 11 of the the datacentre contract. For datasets not exceeding 2.5 GB delivery was required to follow within 10 minutes of requests in 90 per cent of cases and within 30 minutes of the remainder. 5 minutes was added to these targets for every gigabyte of data over the 2.5 GB threshold. This appeared to be comfortably achieved, and was reflected in the system design; after a request had lasted 10 minutes without response it was cancelled and the user was referred to the ordering page. A further requirement compels the service to provide access to paper documentation within five days of a request. Granular targets based on media type were specified but since in practice most requests were satisifed with online delivery these were not largely relevant. Service availability was also required to meet levels of satisfaction. Availability for 98 per cent of working hours was the minimum threshold and periods of scheduled non-availability were communicated explicitly to the datacentre users via the web. In practice, the datacentre maintained its dialogue with customers by presenting detailed accounts of all service disruption on their website. Extensive failures associated with any aspect of the the datacentre service level agreement could be penalised with the imposition of service credits leading to a reduction in funding, as described above.

Technology, Technical Infrastructure and Security

All the datacentre systems aimed to use community-supported software and hardware, including open source systems where possible. There was a variety of bespoke code that was used during various data processing and validation stages but this was generally written in mainstream scripting languages. Some data that arrived within the datacentre was encoded in closed formats and therefore during the ingest and accessioning stages it becomes necessary to rely on both proprietary software and hardware. Generally speaking however the choice of system infrastructure raised no substantial risks of itself being irreplaceable, irrepairable or subject to unanticipated and unavoidable licensing changes that will prejudice the ability of the datacentre to continue at the same level.

Backups and Synchronisation Although backup strategies were well known throughout the datacentre they differed from many aspects of policy in terms of their limited corresponding documentation. One software specialist within the datacentre was unsure about the

specifics of the procedures, and suggested that logs were "probably available", and unable to say for sure whether backup validation and firedrill recovery procedures were in place. Furthermore, he suggested that only a single copy of data was retained, which contrasted from the apparent reality, where four copies of each the datacentre dataset were maintained. PCs within the service were purchased from Dell. There were also Sun servers in use, with storage technologies provided by StorageTek. Other vendors documented as key equipment maintainers (for the service as a whole) were Silicon Graphics, PDQ Computers and IBM.

Backup policy was alluded to within the *Site Exchange Procedures Manual*, which outlined the procedures for moving media between onsite and offsite storage, and recording these procedures. Mainly performed by systems administrators working with archival assistants, other than during normal repository operation, these processes were undertaken when both onsite copies of data were lost at the same time, in which case a third (off-site) copy was recovered and if a particular media batch was determined to be faulty. At no point were all copies of a particular dataset permitted to be in one place simultaneously; four copies were maintained in total. All movements were recorded, identification of transit staff was required to be checked and confirmations of receipt acknowledged and logged.

A Cron task ran a script that checked and compared checksum information (MD5), and highlighted any discrepancies to be subsequently manually checked and repaired. This was sufficiently regular to ensure that within the space of a week all the datasets (at the time of assessment around 1 TB) had been checked. Some questions surroundedd closed files that resided on unmounted filesystems, and whether or not these were subject to the same checks.

In order to ensure the ongoing appropriateness of the datacentre's technological provisions, the service's newsgroups provided a forum to maintain current technological awareness and discuss emerging trends, and any potential problems that might threaten the ongoing viability of current mechanisms. Furthermore, the service maintained an active role within the wider digital preservation and technology world and this enabled it to absorb a great deal of up to date information. Finally, it has been responsible for providing training materials to a diverse selection of communities and was therefore implicitly well versed in contemporary technologies and trends. Omitted from the documentation provided to auditors was an up to date hardware and software asset register which according to the service's continuity plans does exist.

Information Security System security provisions were mainly outlined within the datacentre's seemingly quite comprehensive *Security Procedures Manual*. This outlines procedures for:

- obtaining security clearance for staff;

- the method of recording such clearance;
- notifications to TNA that were required under the contract;
- other declarations required from staff, including dataset-specific declarations;
- physical access to archive areas
- procedures associated with closed or anonymised materials (these were made even more explicit within the dedicated Closed Data Procedures Manual)

Security clearance was provided only with the consent of TNA, after they had received a submission from the datacentre containing details about the individual seeking access. Anyone that would have regular access to archived data as part of their work required such clearance, as did those that while not working explicitly with the data, could access by virtue of their work (such as systems administrators). Forms were available (with criminal record checks performed as standard) and following the granting of clearance this paperwork was retained by the datacentre. A declaration was required to be made by all staff in accordance with the Official Secrets Act, in common with all employees that might have access to government information. Access to specific datasets required additional declarations on an individual basis. Staff awaiting or refused clearance could work in other areas within the service but not where exposure to data was possible.

Each of the three locations where archives were stored were governed by procedural restrictions. The *Security Procedures Manual* described only physical controls; according to this document, logical security mechanisms were described elsewhere, but these were not disclosed during the evaluation. The datacentre *Procedures Manual* index did not appear to include a separate manual dedicated to measures to control electronic access.

Key-only controls limited physical access to the paper storage area within the datacentre, and a register of key holders was maintained. It was prohibited to loan keys to other staff and any loss of keys was required to be reported to Infrastructure Services Management (the parent university service dedicated to maintaining security), the key issuer or the head of building services at the service. No one other than datacentre staff was ever permitted to be left alone in any archive storage area. The silo storage facility had similar controls, with physical interactions with data, as well as logical accesses recorded and logged. If third party personnel required access this was to be provided with a member of the datacentre staff present. At the offsite facilities administered by Recall Ltd., access was only available to the datacentre staff following prior arrangement, under the terms of the service agreement.

The service building was protected by extensive physical security. A front desk was manned at all working times by staff, with an on-site security company covering reception and building security outside the hours of 0700-1800, Monday to Friday. All members of staff were

required to display a parent university ID card and visitors received and were required to display temporary passes, after having signed in. External doors were alarmed, although during working hours these were disabled and CCTV cameras were strategically placed around the outside of the building.

Documented procedures describe further good practice for the datacentre staff; these included not leaving workstations unattended and unlocked, not copying data to workstation disks, or to personal computers or media and not to send closed data via e-mail. It was forbidden to print out material on shared printers, other than in the specific designated staff area, and these copies were not to be left unattended. Source copies marked for destruction could be destroyed only by the individual responsible for processing that particular dataset. One apparent anomaly that arose relating to these restrictions was that several staff during the audit were believed to be working from home, but the security arrangements for remote working were not clarified.

The service business continuity plan made explicit some more aspects of security and contingency planning, presenting details of appropriate contacts and policy guidelines. Outlined within were emergency procedures, disaster recovery and service continuity arrangements and an organisational risk register. This document applies to the service as a whole as opposed to just the datacentre. Detailed documentation described procedures that would follow the loss of the the service computer suite, although these were mainly focussed on JANET, the UK Academic network that was partially based within the service.

A range of contingencies were considered, including personnel accidents, break ins, flooding, electricity or gas problems and fire. For each a description of avoidance and treatment mechanisms was provided, along with appropriate steps to take and details of both external individuals and staff that should be informed.

Computing systems within the datacentre provided some contingency with independent failure of most components unlikely to prove fatal. Disk redundancy was built in with RAID 5. The tape robot represented the archive's only single point of potential failure; since it had only one arm its non-availability would impact on service. A complete robot replacement could be provided in only six hours, so this did not represent too severe a risk, and it was unlikely that such extreme measures would be required. StorageTek, who supplied this hardware, were responsible for undertaking repairs should a fault occur.

Finally, the continuity guide also contained a list of temporary office space providers for use in the event of the non-availability of the Guilford Street Offices.

Three backup copies of all the datacentre data were created, ensuring that there was always a set at the service, one at the offsite store and another at either of the locations or in transit between them. According to documentation, the entire service could be recreated from one set of backup tapes, which also contained cataloguing information and metadata.

Conclusions

One area of concern was the lack of a true mission statement for the datacentre, that was sufficiently succinct and widely available. Similarly, a more cohesive and well expressed definition of its designated community would be useful, if only to further legitimise the numerous policies and procedures that was has formally documented. Also in organisational terms some ambiguities exist within the legal context that surrounds the datacentre. The closed data procedures should be updated to reflect FOI and greater consideration given to the data that accompanies (and was necessary for the interpretation of) Crown Copyright materials, but was itself licensed under different terms.

With respect to digital object management, more explicit and granular planning for specific formats would also be welcomed. The likelihood was that binary data formats will be ever more present within government data sets and the approach to dealing with these appears somewhat ad-hoc in places. Related to the shortcomings defining a designated community, it would be good to see the datacentre explore in somewhat more detail the understandability requirements of its users, and the anticipated requirements over time, in more expansive terms than just accessibility. the datacentre should also take every effort to avoid having to deduce information from datasets that arrive at from government departments. While these departments continue to exist the bandwidth of communication should be extended wherever possible to involve the departments more explicitly with the accessioning process. Some questions also surround those materials that were maintained in their original bitstream form to support contemporary usability (such as the GIS datasets described during discussions). Whether these were being actively preserved was open to debate.

B.5 The UK Research Council Data Centre

Organisational Infrastructure

Mission and Mandate The datacentre's mission statement was most formally expressed within the 'About' section of its website. It read, "The role of the [datacentre] was to assist UK researchers to locate, access and interpret atmospheric data and to ensure the long-term integrity of atmospheric data produced by Natural Environment Research Council (NERC) projects". This clearly reflected a degree of commitment to persistent retention, management and access. NCAS itself, which was the National Centre for Atmospheric Science did not include "long term" within its mission statement, but its self-described role did incorporate issues of stability and sustainability, and reflected a commitment to such issues. However, the datacentre core services included only "physical storage and adequate backup for data

collections", which was not quite the 'long-term retention and management' that the checklist demanded. The fundamental objective of the Centre appeared to be data provision, which despite the implicit temporal issues associated with that, was not quite long-term preservation.

The NERC Data Policy Handbook described the detailed responsibilities of all NERC data centres. NERC grant holders within academia were required to lodge research data with an appropriate subject data centre. In turn, the appropriate data centre was required to:

- ensure adequate physical custody, validation, dissemination, review and purging of that data;
- maintain standards of data stewardship;
- proactively seek out data within its subject area that would merit stewardship;
- promote the case for investment where necessary to facilitate the above;
- promote the use of data;
- formally arrange licenses to control the release of datasets to non-NERC recipients, and the uses to which it can be put, and to protect NERC from legal liability;
- advise on the licensing or purchase of non-NERC data that was required by researchers;
- handle all requests for data within its subject area made to NERC;
- maintain up-to-date details of holdings available via the world-wide web;
- act as a gateway to other NERC data custodians;
- represent its discipline within NERC on matters concerning data

There can be little question that these generic requirements placed the burden of archival responsibility upon the datacentre. Furthermore, the datacentre was expected to not only respond to the submission of content, but also seek out relevant data, secure further resources to support its activities, assume responsibility for the legality of its operations and work to develop a network of NERC data centres. These requirements spell out the fundamental mandate of the datacentre, and provide an operational context within which assessment was meaningfully carried out.

Further data protocols and policies described other specific responsibilities that related to data originating from only particular NERC programmes. For instance, the QUEST Data Policy builds upon the NERC generic policy, extending it for data originating from QUEST

member research. In some ways, these data policies were analogous to deposit agreements, and defined more specific responsibilities for both depositors and the data centre. Some points from this example, which appeared to be quite illustrative, include:

- The datacentre has primary responsibility for all QUEST data sets;
- The datacentre can refuse data that was of insufficient quality (e.g., lacks appropriate documentation) or was of little long term value;
- Data must be lodged with datacentre as soon as they were validated and no later than three months after acquisition (except where it cannot be determined whether data was suitable for long term post-project curation);
- Data must be made available to all QUEST community following submission to datacentre. Within the first year after submission uses by members of QUEST that were not producers must get the permission of producers for use in order to enable principle investigators to get the first chance to exploit its value. All QUEST creators were encouraged to share, and required to keep this embargo period as short as possible, up to a maximum of one year;
- All data must be made publicly available after one year. However, any data users within two years from the end of the originating project will be required to give the name originators of data the option of co-authorship on any resulting papers.

Immediately apparent from these requirements was the onus placed upon the datacentre to determine the quality and value associated with particular datasets. This would seem to indicate that there was considerable emphasis on archival appraisal for submitted data, although there seemed little to suggest the availability of sufficiently robust policies and procedures to support this. Policies for data review and retention were stated within the centre's *Operations Manual*, although the implementation of these appeared somewhat weak.

Organisation and Steering A NERC Steering Committee was responsible for advising on programme development, and ensuring the implementation of associated data management plans. A data management sub group, including representation from the datacentre (or other appropriate data centre(s)) was convened to support the Science Coordinator in these activities. The Steering Committee was responsible for ensuring that data management was carried out effectively (by providing adequate support and resources during the programme); an appropriate data management plan was created; a realistic proportion of the overall programme budget was devoted to support data management; and all holders of programme awards comply with the data management policy of the programme, as outlined in the Data

Management Plan. On this last point in particular the curation manager revealed some scepticism, indicating that there were issues with enforcement. Consequently, circumstances have arisen where NERC grant holders have failed to deposit content, or have provided incomplete datasets. The datacentre was not involved in the grant process, and therefore could not compel deposit where it would be worthwhile. It was suggested that the datacentre would have benefited from access to a more detailed online register of grant awards, to support their pursuit of worthwhile data. In practical terms, discussions suggested that the datacentre was simply not offered a great deal of data; however, it was also suggested that this situation was changing, and therefore it was imperative that the datacentre was equipped to cope with a consequent upsurge of deposited data.

Organisationally, the datacentre existed within a research institution setting under a Service Level Agreement which describes core services, infrastructural developments, support for other NERC Centres for Atmospheric Sciences and research. The SLA was renegotiated on an annual basis. Core services as stated in the 2005-2006 SLA include the following, which correspond closely to the NERC data policy requirements:

- Acquisition and distribution of observational data from the MET Office;
- Acquisition of numerical weather prediction data from the Met Office and European Centre for Medium-range Weather Forecasts;
- Physical storage and adequate backup for data collections;
- Computing system to support data storage and limited user processing;
- An online catalogue of all data collections to help users to find the data they require;
- A distribution service to allow users to access data (including maintenance of both FTP and web interfaces to the data.

Similarly, and again under the SLA, the datacentre was required to provide technical and management advice for existing data activities in other NCAS centres, promote the datacentre and NCAS by maintaining a prominent presence in meetings and providing appropriate publicity and liaising with other data centres from other disciplines within the UK. There was an additional research requirement, whereby in order to maintain the validity and relevance of datacentre activities, the Centre should carry out an active research programme in climate physics and data assimilation research.

The datacentre was one of two distinct NERC designated data centres sharing a parent institution. The two had resisted merger, favouring a 'two front doors' approach that seemed to permit greater influence over NERC.

Succession or contingency planning was not regarded as the responsibility of the datacentre. Instead, as noted in the curation manager's self-assessment responses, this was regarded as an issue that NERC was responsible for resolving. Little evidence was available to indicate whether NERC had implemented appropriate arrangements to deal with any cessation of repository operations.

Staffing The datacentre directly employs around twelve full time staff members, although the close relationship with its sibling repository mades it difficult to determine where some positions (specifically research oriented positions) predominantly lie. There was evidence to suggest that there were appropriate staff numbers to fulfil the functions of the repository. below the director and then curation manager levels the staff hierarchy incorporated a number of roles. The first was Environmental Data Scientists; these individuals were responsible for managing data ingest from data suppliers, with duties including the documentation of datasets, collation of metadata, negotiation with suppliers and responding to queries from users and suppliers. Generally, the staff members performing these roles were from atmospheric science backgrounds, with postgraduate qualifications at either Masters or PhD level, and capable of demonstrating general IT competence. The second role includes those responsible for Operations and Delivery, which was concerned with providing generic services for datasets. This includes hardware infrastructure, media handling, query management, access services, user management and other more generic curation issues. These staff were required to demonstrate expertise in one or more of a range of disciplines including computer science, systems administration and atmospheric science. The final role encompasses researchers and developers. The former were expected to have PhDs in atmospheric science with the latter expected to have accumulated experience of software development, most probably through the completion of a computer science degree.

Staff numbers were sufficient, and evidence suggested that the organisation had a broad understanding of the implications in terms of workload and personnel requirements of a move towards a more 'OAIS compliant' infrastructure. As indicated above, the organisation would have benefited from committing extra resources towards activities associated with preserving content. For instance, in physical bit-level terms, data storage, hardware, backups and checksum management could have been prioritised. Similarly, archival lifecycle activities such as ingest, data and metadata management could have been assigned explicit ownership by individuals or roles within the Centre. Furthermore, legal advise should have been at least solicited given the doubt that to some extent surrounds agreements with depositors and content creators. These recommendations were primarily focussed on enhancing the repositorys scalability.

Training was available for staff, although it was presented in a fairly ad-hoc fashion, based mainly on staff demand. At the time of the audit, recent training had been offered in scripting

languages such as Perl and Python whereby several copies of prominent learning texts were purchased and staff were encouraged to learn as a group. Atmospheric science courses were available for those with less experience of the datacentre's community's primary discipline. STFC forms were available to request other specific training opportunities, and management courses and health and safety training were encouraged for all staff. There were however few mechanisms in place to identify knowledge or skills gaps and therefore staff were expected to maintain a degree of consciousness of their own shortcomings, which was perhaps unrealistic. Better structured training programmes could have been developed and associated with particular roles within the Centre, in order to ensure the effective development of staff. Notwithstanding this, evidence and testimony within the Centre suggested that training was provided promptly when needs were identified, although this seemed to be most obviously true when a wide desire was expressed.

Turnover of staff within the Centre was low, which implied that those employed were both experienced and competent. An inevitable consequence of course was that longer term staff would cost more money to keep employed. Since CCLRC deals almost exclusively in permanent contracts with good terms this factor must be taken into consideration when accounting for the Centre's financial sustainability.

Designated Community The Centre's designated community was defined more narrowly than its overall potential user base. The definition included UK based atmospheric scientists or non-undergraduate researchers in atmospheric science who were English speakers and had access to Internet services. Of users registered within the system only 45 per cent fitted within these parameters, reflecting the diverse range of individuals interested in accessing content, but the Centre's commitment (in terms of preservation) was limited to making materials usable for that stated group. Although an internal understanding of the designated community of the datacentre was demonstrated during discussions with the curation manager, there was no evidence of a centralised, published definition. Similarly, little evidence was available to suggest that formal mechanisms were in place to monitor the evolution of this community, although the relationships that were maintained with depositors, who were in turn likely to be end users, were close, and provided insights into latest developments. However, of the 8000 or so registered users of the datacentre web access system, a small fraction (less than 5 per cent) were responsible for depositing. Further insights into community developments were available as a consequence of the research work that datacentre staff were personally involved in; staff were in many cases themselves part of the defined designated community, and their interests, knowledge base and expectations were to a greater or lesser extent representative of the wider world. A further means of monitoring was provided by conducting user surveys (usually comprising around ten questions, and part of regularly conducted NESC assessments). Finally, the Centre maintained a user

queries system whereby users can provide any feedback about data, services, or any other aspect of the datacentre. In reality, each of these mechanisms had community monitoring as something of a side effect, rather than representing its primary rationale. Nevertheless, they provide evidence of an ongoing dialogue with the designated community.

There was little to indicate the existence of formal mechanisms to react to accumulated evidence of community evolution. One anticipated an ad-hoc approach to dealing with changes in understandability or usability expectations. Frequently, the datacentre's approaches did appeared to reflect the capabilities of their defined communities. For instance, it was revealed that despite a management will to provide more digital materials encoded in the binary NetCDF format there was an internally held perception that this would represent a barrier to usability for up to 70 per cent of the designated community who would struggle with this choice in terms of usability (through lack of required skills or software). However at the time of this evaluation perceived changes in the designated community had not led to any notable changes to systems or processes. The sole recorded case where there was evidence of a reaction to changing community expectations was when the Centre sought additional resource for a period to cater for emerging communities associated with a particular surface dataset. The NERC Data Management Advisory Group financed this activity, enabling the datacentre to reflect the changing needs of the (expanding) designated community.

Policies and Procedures In terms of policies and procedures, the datacentre's most obvious omissions were a rigorously defined preservation policy and a complementary disaster plan. As well as lacking specific policies, it appeared that those policies currently in place were seldom subject to systematic review; instead updates seemed more likely to be applied reactively, in response to specific circumstances or problems that have arisen. The monitoring activity within the centre was very good, with considerable technological mechanisms to facilitate inter-team communication, and a weekly team meeting. However, evidence of structured policy assessment was lacking. Similarly, although research activities were extremely active and closely aligned with the management of the data service, and there was evidence of an awareness of emerging trends and technologies, this too seemed to be more ad-hoc than formalised. However, this may not represent a particular problem; it might be argued that as long as an awareness was maintained of the best contemporary exchange formats for the Centre's designated community, in association with sufficiently low risk storage and backup systems, the preservation aspects of the datacentre's mission were likely to be realised. Perhaps more important than monitoring technological changes, and therefore worthy of greater investment, was being equipped (both intellectually and in terms of resource availability) to react to changes when they became apparent. The conception of disaster plans, accompanied by the introduction of regular fire-drills and data recovery testing was also recommended.

The lack of an explicitly defined chain of custody within the Centre was of some concern; it was clear from discussions that the datacentre had no audit trail throughout the life-cycle of a data resource within its custody. This meant that trust in the data or data integrity could be compromised. Cases were described where users have requested proof that the data delivered to them was correct and complete, and the solution has simply been to verify this with the data creator. Such reliance creates obvious problems when data creators were no longer available or lack complete records of the content they've created. Although the tangible risks associated with such circumstances remain difficult to quantify, the damage to reputation and trust (two assets of considerable value to a service like the datacentre) could be considerable. In order to achieve trustworthy status greater investment was required into archive management practices and the creation and management of documentation.

Validation of Policy There was little doubt about the datacentre's willingness and enthusiasm to seek assessment and external review. Their very participation in this pilot audit, and overall enthusiasm about the process and the subsequent opportunities that may arise for formal certification indicated as such, and demonstrated an understanding of the potential benefits of such activities. Similarly, each of the NERC Centres for Atmospheric Sciences were required to undertake an external and independent Science and Management Audit, intended to demonstrate management effectiveness. In the report of the 2004 audit the datacentre was rated as excellent, delivering a significant national capability. Concurrently, during the same period the datacentre conducted a survey of its registered users (April/May 2004), and although only 7 per cent responded, the exercise yielded a sense of successes (range and quality of data; fast network; prompt human response) and shortcomings (access restrictions; lack of tools; update frequency) which have seemingly informed subsequent efforts.

A final concern with respect to policy and procedure was that the datacentre risk register, potentially one of the Centre's most useful documents, existed only in a prototype form. It was suggested that as a mechanism to support and facilitate contemporary organisational management (and of course more long term sustainability), this should be brought up to full production status at the earliest opportunity.

Business Planning The self-completed check-list response submitted by the curation manager prior to the onsite audit activity indicated the existence of an NCAS business plan, which contained details related to short and long term planning for the datacentre. However, these documents were not made available to auditors. Similarly, although processes exist at a parent level for the review and adjustment of these planning instruments, there was little evidence of an autonomous and more granular approach to business planning and ongoing management to ensure the ongoing strategic resource investment to facilitate infrastructural

sustainability. This reflects an earlier stated concern that NCAS was failing to engage adequately when negotiating issues that datacentre staff ought to be more closely involved with. Examples included the negotiation of contracts (or data protocols/policies) with NERC grant holders and the conception of succession plans or agreements.

Financial and Accounting Infrastructure Accounting procedures were consistent with those of the datacentre's institutional parenT, and evidence was made available to demonstrate compliance with standard accountancy best practice. Money was available to alleviate problems associated with short term funding gaps, but of some concern was the suggestion within the Centre's prototype risk register that there was a thirty per cent likelihood of funding being cut with a significant impact, meaning the irrecoverable loss of some medium priority and high priority data. It was suggested that the Centre may therefore wish to seek more formal assurances for financial allocations where these were available or alternatively investigate the possibilities that may be available for generating self-sustainable services. Nevertheless, the datacentre had contracts established with NERC to store and disseminate data originating from NERC funded research. There were three tiers of projects that contribute to datacentre budget, and therefore the risk that all three would disappear for a long period of time appeared to be quite low.

Much of datacentre funding came as a consequence of NERC directed mode programmes, where scientific researchers bid for a particular pot of money to undertake experimental research. Alternative NERC funding comes in the form of response mode grants, although these rarely generated data that the datacentre was interested in. Finally, NERC consortium grants, involving substantial funding and widespread participation, were a frequent source of income for datacentre, given the nature of data originating from such research. There was an intrinsic flexibility associated with the core NCAS funding. A written agreement allowed the datacentre to use this money to match funding in other activities, in order to bring in more money, assuming that core service needs were adequately satisfied. Similarly, money could be carried over from year to year, an important facility given the often unpredictable costs associated with maintaining a technical service. Internal datacentre research activities were financed partially by NCAS, although this amounted to little more than a single FTE researcher; there was an expectation that the Centre would pursue research grants to supplement these resources.

Of some concern was the revelation that the availability of content originating from the MET Office (which, it was suggested, the datacentre relied upon for its very survival) was not guaranteed due to the non-renewal of the agreement between the MET and the datacentre. An original archive and dissemination agreement ended in 1999 and its terms stated then that datacentre should destroy all data upon cessation of agreement. Notwithstanding an ongoing renegotiation of this contract between NERC and the MET Office, it was suggested

that this, and similar idiosyncrasies should be addressed with high priority, particular where unforeseen legal impediments might impact fatally on the datacentre's ability to continue to operate. At the time of the audit, a new contract between NERC and the MET Office was in the midst of being negotiated (agreement was regarded as a mere formality), with an appendix explicitly permitting the datacentre to redistribute MET Office data to bona fide researchers for the purposes of publishing papers. Although NERC incurred no direct charge from the MET Office for providing access, it costs a great deal in terms of time; at least 25 per cent of datacentre staff time was spent on MET office data requests. The extent to which this could be sustained might be questioned, and perhaps ought to have been more formally documented. There were suggestions that the availability of MET Office datasets were a necessary precursor to the datacentre's conception and necessary for its survival.

Ownership, Rights and Legal Issues The datacentre appeared to operate without formal deposit agreements, and where agreements exist there appeared to be a somewhat loose approach to contractual management. The relationship with the MET Office outlined above provides a good example. Data originating from NERC was subject to more formal terms, and these were outlined within the NERC data policy and associated program-specific data protocols and policies. These appeared to be negotiated at a level beyond the immediate datacentre management (although appeared to involve at least some degree of consultation with datacentre).

Further rights-related questions were associated with ambiguities over the datacentre's right to change data, which might be regarded as 'preservation rights'. The internally held perception was that the datacentre probably did not have the rights to change, or even reformat data. Instead the rights were to share data. In practical terms though, it was argued that irrespective of the existence of these rights, the datacentre would not change data because of the question of trust (the internally held perception being that users don't trust the archive to alter content), and the comparatively small number of widely acknowledged data formats: both of the principle formats utilised by the datacentre, NASA AMES and NetCDF, appeared stable and widely used. Ownership of data was similarly unclear; NERC was committed mainly to making data public, and therefore questions of ownership were given considerably less priority. Discussions and analysis of example NERC funding agreements suggested that NERC was unlikely to own data generated from funded research since this was not expressly stipulated in the grant award documentation. Notwithstanding such ambiguity, it was claimed that the datacentre might not to make data public if there was a notable associated revenue stream that might be exploitable. In the case of NERC data, contracts exist between NERC and the appropriate universities/researchers providing (and in many cases owning) the data. No direct agreements were formed between the datacentre itself and depositors, with mutual responsibilities mainly encompassed in a data protocol corresponding to each NERC pro-

gramme. There were questions about the extent to which the datacentre was legally entitled to make such management decisions based on their perceived value of archived data.

Monitoring of intellectual property rights existed in embryonic form in the *Operations Manual* on the datacentre wiki, but this fell short of the "comprehensive overview" demanded by the audit checklist. Nevertheless, it was acknowledged that this issue could be taken to extremes, and ISO 9000 series certification might be regarded as necessary in order to conform to the check list. However, within this context, it was felt that to make such demands of the Centre was not helpful and ultimately unnecessary. Of potentially greater relevance, conversations revealed that no formal policies or procedures exist relating to requirements arising from Data Protection and Freedom of Information legislation. the curation manager's (layperson's) view was that neither Act presents any problems or legal incompatibilities with existing datacentre practice. With respect to Freedom of Information, the rationale appears to be that existing Environmental Information Regulations supersede FoI, encompassing all of its requirements (and more). However, in order to ensure that everybody at the archive was aware of legal issues and the appropriate approaches to resolve any associated concerns, it was suggested that written guidance should be conceived. This was especially true for those members of staff responsible for answering user queries and providing user support. Legal doubts were apparently quite widespread and the current non-formalised approach, which seemed to be based on continuing as long as no legal challenges arise, might be unsustainable. Contracting a lawyer to advise for a month or two should not represent a vast investment, and was recommended. NERC has set some precedent with respect to legal issues, when it provided the datacentre with legal input for the purposes of drafting a limitation of liability statement. STFC has in house legal expertise available.

A final concern in this area relates to the current system workflow evident within the datacentre. Control of access to datasets was built into the access system/interface, with those responsible for information ingest responsible for determining the appropriate access level for particular datasets and content. There was therefore perhaps scope for concern when dealing with atypical access rules. Nevertheless, this was a small concern, as the system demonstrated sufficient fluidity to suggest that it could be altered to reflect emerging requirements.

Digital Object Management

Acquisition and Ingest Ingest processes at the datacentre were in an ongoing state of evolution, a fact that was acknowledged within the Centre's Online *Operations Manual*. The most urgent requirement appeared to be the development of increasingly formalised ingest methods to ensure both robustness and scalability amid an increasing quantity of deposited and retrieved data.

The datacentre's Environmental Data Scientists were responsible for engaging with data suppliers and providing overall management for the ingest process. At the beginning of each NERC project, staff liaised with the relevant community to determine expectations, and to get a sense of the form that data was likely to assume. In some circumstances data was relatively easy to obtain; the most notable examples were 'pulled data', retrieved automatically from remote, networked instruments. In other cases, a greater onus was on datacentre staff to actively pursue the acquisition of data.

In quantitative terms, the datacentre's five ingest staff each dealt with around twenty datasets, and each has a corresponding named contact. Following this initial contact, a number of mechanisms existed to practically support the ingest of content. The most common required scientists to simply upload their data following its capture via the File Transfer Protocol. Data transferred in this fashion was delivered to the datacentre 'incoming' directory, where it would reside until the ingest team manually transferred it into the archive. An alternative mechanism, the web based file uploader tool, was only marginally distinct from this approach, with additional file format validation and file integrity verification controls exercised prior to upload. Although such automated controls were not available for data arriving via FTP, some such content was subject to random validation and verification. 'Pulled data' was retrieved to a 'deliveries' directory, broadly equivalent to the 'incoming' directory mentioned above. Files ingested into the archive shared a common naming convention, as indicated below:

instrument_location_YYYYMMDD[hh][mm][ss][_extra].ext

Both instruments and locations were required to be registered with the the datacentre to ensure their validity. Some ambiguity existed though over the extent to which an instrument could change and still maintain the same identity. For instance, if components were replaced, was an instrument the same as it was before? Similarly, some confusion appeared to surround the location and time information of certain data, such as those mounted on aircraft. Generally speaking, such ambiguities, as well as processing that has been undertaken on particular data, will be documented within file-specific metadata. At the point of ingest, documents may have been associated with datasets in order to provide format descriptions, details of problems or further information about particular instruments. This documentation was associated with data in one of various ways. The first was to include it within a file named 00README that was located within the relevant data directory. An alternative approach was to create a file entitled 00instrument_location_date.txt, which seemed to be a more robust means of enforcing the association. Finally, staff explained that some file formats, most notably the NASA AMES format, supported the addition of in-line comments.

Following deposit or retrieval, ingest staff were required to undertake some subsequent processing prior to archiving. High level catalogue information was created at the datacentre;

a proactive process, this demanded engagement with creators and information owners. The CDML or CSML XML schema provided a transparent vehicle to describe model data, and this incorporated details of all permitted instruments and locations. Relational databases were used to store data metadata corresponding to non-modal/gridded data. Files were checked too for instance, the NetCDF file checker determined that mandatory constituent parts of such files were present and valid. Processing to split the files could also be undertaken. Scripts to handle this data processing and the ultimate deposit within the archive were created for each data stream being ingested into the datacentre, based on a generic deposit program. As outlined within the *Operations Manual*, this meant that there was a real need for rationalisation, given the wide number of only slightly different scripts that exist. Indeed, the *Operations Manual* outlined a number of requirements for an improved ingestion methodology, which closely reflect the findings of this assessment. Firstly, there have been historical examples of data files in the archive being incorrect or corrupt, caused by bad data from suppliers, or errors during data transfer (both from the outside Internet to the datacentre, and within the internal network of servers). Discussions suggested that in at least one case, content that arrived automatically could be ingested into the main archive even if lacking documentation, encoded in an incorrect format or non compliant with file naming conventions. This issue was one of trust; the datacentre's relationships with certain depositors were such that they would store any content that originates from them (even, it was suggested, if that content was rubbish). Such arrangements were not formally defined however. Questions of data inventory abound; during the audit staff appeared surprised at the existence of .avi movie files within the archive, with no obvious means to preserve such formats.

Several other areas for improvement had been already identified. These include more modularisation of the ingest process (so that for instance the method of ingest was not influenced by the physical manner with which data arrived at the datacentre); more automation, to limit errors and optimise staff time; logging to maintain an appropriate audit trail; better metadata; quality control mechanisms; and more formalised methods for data pre-processing. There was a requirement for substantial efforts at both an organisational and technological levels to introduce formal, internally enforceable policies on ingest.

Since no deposit agreements were formed between archive and depositors, it was difficult to determine whether all of the properties of digital objects were being preserved. The practice at the datacentre implied a commitment to preserve the received file format and its accompanying documentation. There did not seem to be any other explicit 'properties' defined that were being actively preserved. Similarly, the data policies and protocols associated with specific NERC programmes revealed little about acceptable file formats or levels of documentation that should accompany data. A specification did exist to describe the information that should accompany data submissions, and depositors were made aware of the deposit requirements. Unsurprisingly, SIP configurations may differ slightly for different deposits,

depending on how the data was collected and managed prior to deposit to datacentre.

Content Selection The NERC data policy did permit the datacentre some discretion in deciding whether or not to accept deposited content. "The sole reason for keeping data", described the datacentre *Operations Manual*, "is to distribute it for use". This did not presuppose contemporary use, instead acknowledging the fact that even long term curation was undertaken with a view to one day using the information that has been preserved. The two most influential considerations were the usability and usefulness of data. The former was influenced by such factors as the format within which the information was encoded (NASA AMES and NetCDF formats were the most common, and likely to be usable by most of the datacentre's registered user community) and any conditions of use associated with the dataset. The latter consideration related to the likelihood with which people would actually want to use the data its quality, including scale, coverage and number of gaps were relevant here, as were its uniqueness, its potential for strategic use by the datacentre and the breadth of its parameters, a key factor for determining its reuse potential. Expressed elsewhere, the datacentre described its efforts as a combination of facilitation and curation. The former was concerned with adding value, by storing and disseminating from within datacentre. The latter was about ensuring survival and ultimately, long term usefulness. These primary motivations were expressed with a greater degree of granularity within the Centre's *Operations Manual*, summarised below:

- Facilitation Arguments
  - Good coverage in terms of time and space, with few gaps and high resolution;
  - Data exhibits parameters of sufficient breadth to ensure their usefulness outside of the projects that collected them;
  - High contemporary usage;
  - Complementary to funding body objectives;
  - Shortcomings inherent in the current data source;
- Curation Arguments
  - Uniqueness;
  - Data lacks a primary archive;

Dataset Review Of note were the lack of efforts to assess the *ongoing* value, usability and performance of data stored within the datacentre. In fact, the Centre has considerable documentation describing a dataset review process, and it was therefore regrettable that to

date, this has not been particularly well implemented. Dataset reviews can be considered in similar terms to archival appraisal, and within the *Operations Manual* the procedure was described as a Retention Process. The process, as documented, was prompted by an automatic notification that data review was due, with a milestone in the project database conceived to correspond to each review. When this happens, a responsible individual was required to evaluate:

- the content of the data's corresponding catalogue entry (checking that links were working for example);
- the extent to which corresponding web pages were current, appropriate, informative and usable;
- the extent to which data was usable, accessible and adequately documented;
- whether any representation information (specifically software) was required to use the data;
- the effectiveness and security of corresponding ingest mechanisms; and
- the extent to which data was adequately documented, creating and aggregating documentation where appropriate.

Applying these criteria would result in evaluation marks between 1 ('Poor') and 5 ('Excellent') corresponding to both data usability and usefulness. Reviewers were then required to propose subsequent action, which could be to leave the dataset as is, keep the dataset but implement some changes or remove the dataset from the archive. The latter seemed to imply destruction, with no overt infrastructures in place to support transfer of stewardship to a more appropriate repository elsewhere. Despite the reasonably robust provisions for data review, discussions with the curation manager revealed that the process has seldom been undertaken. The number of datasets currently within the Centre, combined with the time consuming nature of the review process has been the most critical factor simply put, review notifications were arising more quickly than staff can undertake reviews. This was undoubtedly a problem, and threatens the viability of long term archiving within the Centre. The datacentre's emphasis was very much on ingest, with dissemination enjoying a comparable, albeit secondary level of prioritisation.

No explicit mechanisms existed to authenticate the source of data that arrived at the Centre, with depositors authenticated using the system relied upon for end user access. The expression of interest in depositing data with the datacentre was weighed by the ingest department using circumstantial evidence (e-mail address, name of institution named as employer, was

the depositor known in the scientific community of the subject area). No agreement or contract, legally binding or otherwise, was usually signed between the depositor and the archive. In many cases, the depositor could be compelled to deposit their data with the datacentre (e.g., through a funding contract with NERC), in which case authentication was by association with the funding agency. Submission Information Packages (SIPs) were verified, but usually as a statistical sample of the whole deposit. Since this was deemed as sufficient for the datacentre, there was no reason to expect that all SIPs should be inspected at full length. Especially given that some continuous data streams were uploaded automatically to the datacentre about every minute, it would be difficult to verify all data contained in these datasets in real time, before they become available to the users. Verification was performed by using automated scripts (python and perl languages) which were custom-developed for each dataset.

A transfer process was also performed by using automated scripts that enabled datacentre staff to verify that all files had been transmitted in a single submission session. However, checksums were not used at the data transfer stage nor during ingest processing (only AIPs receive checksums).

The introduction of tracking mechanisms was considered worthwhile to enable depositors to see at what stage of the ingest or preservation process their data had actually reached. During the audit, ingest staff spoke of some experiences where depositors had misunderstood the time-scales within which the ingest process would be undertaken and were expecting their data to become available immediately, following its upload to the datacentre. This indicates that the ingest workflow could be better explained and made more visible for the depositors.

Preservation Policy and Levels Within the archive, a distinction was drawn between three classes of data, tentatively known as A, B and C data, which correspond roughly to the extent to which their preservation was prioritised. The characteristics of each was explained in simple terms by the curation manager. Class A data was that for which the datacentre was the primary archive, and this amounted to approximately one third of all data holdings. Class B was that for which although the datacentre was not the primary or sole custodian, scepticism existed about the ability of the primary archive to provide adequate preservation services. Class C data was that which was adequately preserved elsewhere, but was sufficiently useful to retain. Only class A and some class B data for which the datacentre considers itself to be the primary steward was really relevant when considering issues of preservation. The fundamental differences between classes were not formally expressed anywhere at the time of this assessment - the internal classification was a realisation of an appraisal - but had tremendous influence over the preservation activities to which particular data sets will be subject.

Following ingest, data that arrived into the datacentre's archival storage were likely to have already been subject to some processing. Class A, B and C data were distributed across several disks; a single directory contained symbolic links that corresponded to each dataset, and pointed to the physical space where individual data streams were located. All access, including end user access was via this directory. The server that this directory resided upon probably represents the most vulnerable point of the system. If compromised this could limit the extent to which data can be retrieved or its completeness ensured. Worth noting here was the fact that by maintaining a single system for archive and delivery, the Centre was limited in terms of the extent to which system changes could be implemented while maintaining an optimal level of service.

The transition of a SIP into an AIP and subsequently a DIP was not regulated by a formal policy, nor documented anywhere other than the resulting directory structure on the archival storage media. The datacentre only accepts data that it has appraised as suitable for deposit. Therefore, SIPs were always transformed into an AIP/DIP. The choice of structure for archival and dissemination packages was made by the ingest staff, who make their decision based on how the dataset was most likely to be used. Therefore, the primary criteria for converting SIPs into AIPs was ease of use. But no strict policy existed, and the decision to convert a particular dataset in one way or another, was not documented in a separate log, audit trail or documentation. It was only visible from the consequent presentation of the data in the storage system.

No formal criteria were established to determine when preservation responsibility was accepted by the datacentre, and neither was the transfer of responsibility acknowledged in a deposit agreement exchanged with the depositor. In practice, the datacentre assumed preservation responsibility from the moment the data had been transferred (uploaded) to the datacentre and an e-mail had been received from the depositor containing the script of the completed transfer. The datacentre could have benefited from formalising this process, especially for the class A datasets. In terms of current practice, whereby the datacentre generally utilised a single file format for SIP, AIP and DIP, the conception of detailed depositor agreements may not seem to have been necessary. However, looking into the future when the datacentre may have had to consider AIP or even SIP migration as part of preservation processing, it would have to be clear about the rights and responsibilities it has with respect to data.

Since a finite number of AIP 'types' were generally accepted (measurement data, model data, satellite data, Met Office data), the *Operations Manual* specified an AIP configuration for each. However, the AIP configurations were not documented in a sufficiently structured way to permit the automatic verification or validation of an archival package. The AIP definitions maintained by datacentre were largely sufficient to meet long term preservation requirements, although, again, were rather poorly documented. The choice of file formats for each class of data stored was more based on data usage criteria than on issues of long-

term preservation. In fact though, they also happen to be suitable for preservation without requiring extensive processing at short intervals. Although the datacentre *Operations Manual* described the process of constructing archival packages from submission packages, no documentation was created in practice that enabled one to verify whether the instructions had been followed. The division of a SIP into constituent AIPs (i.e. files in a directory structure different from that of the original SIP) was not tracked no checks were performed to verify whether all files transferred to the datacentre were in the stored AIP.

Perhaps more relevant with respect to this point was the appraisal of datasets for classification into class A, B and C datasets. The appraisal criteria were fundamentally sound. The problem though was that the class or category assignment did not mean anything for the AIP configuration irrespective of whether a dataset was classified as A, B or C, it was kept in the same file format and supplemented with the same kind of documentation. The only difference was in storage and back-up practices whereby class C data was supported by fewer safe copies (if any). It could have been argued that the AIP configuration was sufficient for preserving the class A data, and that it could do no more harm to also apply it also to class B and C data. However, the in principle the content with high preservation priority should have been accompanied by richer supplementary information, in greater quantity. Documentation as such was a weak point at the datacentre, at least from the archival point of view. One could argue have argued that all AIPs of class A should have had their entire custodial history logged, all processing decisions documented, all usage occasions tracked, and all changes to documentation audited. This was not done at the time of this assessment. An example was offered of the Hierarchical Data Format raw data that arrived from the HIRDLS instrument aboard the Earth Observing System (EOS) AURA Mission Spacecraft. Given the nature of the data, which was tied very closely to the spacecraft's instruments, it was vital that the semantics of the data were appropriately documented with sufficient representation and provenance information.

Integrity Validation Identifiers were assigned at ingest as described above, and the file naming convention was publicly documented. The instrument name (itself unique), time and location information created sufficient heterogeneity for the file names to act as unique identifiers, at least internally within the archive itself. These were maintained throughout the archival process, and therefore ensured traceability between SIPs, AIPs and DIPs. Because each of these packages maintains a mutual one to one relationship (SIPs were never split into multiple AIPs for instance) then this maintained the integrity of the references.

Ingested files were not immediately protected against alteration (checksums were applied later in the archival process) and no audit trail was created from the processing done at the ingest stage. The ingested files were vulnerable to malicious or accidental alterations until checksums were calculated (upon creation of the AIP), but this process could take up to 40

days. A monthly automated script was responsible for calculating checksums and comparing them to the stored value that was taken at the point of AIP creation.

The only links between updates to an AIP exist in the form of directory structure of the AIPs within the storage system. The location of AIPs in the storage system was managed by a single machine which provided symbolic links from a single directory to the actual physical locations that data was stored. If datasets were moved then the symbolic links were updated to reflect that. Dissemination and archive management mechanisms use the directory when referencing content. Because only weekly backups were made of this machine there was a risk that links between access systems and archive management systems may have become fractured.

In terms of preservation metadata, some popular fields were present, but since the Centre had not formally compiled or published its metadata standard, there was no way to evaluate how the metadata creation taking place actually was. Some provenance information was included in the description of the dataset on the web, question marks remain as to how this was collected. It appears to be created by the Ingest department, but the extent to which it was based on information provided by the depositor, was unclear.

Metadata documenting preservation actions and processing performed at the datacentre was generally not created in a systematic way. The use of archive management software or a work flow tracking system may have facilitated metadata collection and management. Discussions touched on the possible use of datacentre's Trac WiKi system to record the tasks and stages that each dataset goes through within the repository.

There was no formal mechanisms to support representation information acquisition and management, or for determining approaching file format obsolescence. However, since content was mainly restricted to one of two file formats, and staff maintained close relationships with theire designated community, they should be reasonably well informed and forewarned should one of these file formats begin to appeared vulnerable. Questions remained though about the extent to which these relationships were formally and systematically explored. Understandability was ensured by consistently employing data scientists from within the communities that ultimately use archived content (and were therefore able to demonstrate comparable knowledge bases), and this seemed a sensible and pragmatic approach. Given the tradition of low staff turnover this might just be a problem over time if the communities in the outside world change unbeknownst to those operating within the staff of the datacentre, who have long completed their studies and other academic activities.

Since no preservation policy as such existed, it was difficult to remark on processes for changing it. There did not seem to be a steering board, council or working group that would meet regularly to discuss preservation issues and how to change the existing policies and practices. It was suggested that this might be rectified to give a clearer, more structured

insight into the ways in which the archive can evolve and develop. Similarly, there was no clear way to determine the effectiveness of the repository's preservation planning, although it was acknowledged that this was tremendously difficult to demonstrate; it was much more straightforward to identify where such activities have failed, and the datacentre appeared to have a good track record with respect to avoiding data loss, particularly of the most highly prized, class A assets.

Preservation was a somewhat simplistic task at datacentre effectively just storage of bits in an particular file format and no preservation activities were undertaken on a regular basis that could endanger the semantic properties of original deposited files. The rationale behind this was partially one of trust; as the curation manager explained, and was described above, scientists and other users do not trust the data centre to make changes to the data. A degree of scepticism surrounds science, the curation manager continued; historically, few scientists would use commercial software products due to fear of data distortion, instead favouring their own bespoke FORTRAN programs. The other factor was a general reliance on reasonable, well documented and open formats that were well supported by a range of widely accessible tools. NASA AMES and NetCDF were the primary atmospheric data formats employed. The former was simple, non-binary and requires a minimum of metadata, and although the latter was more complex, and a binary format, it was nevertheless an open standard. The decision to primarily support these two formats was cited as a preservation decision. Alternatives such as Gridded Binary (GRIB) and the MET Office's proprietary PP-format were also evident within the datacentre; the latter was supported by a range of contemporary software packages, and therefore appeals despite its limited features (one cannot encode comments into PP files for instance). GRIB, despite enjoying the status of World Meteorological Organisation (WMO) standard was a complex file format that seemed somewhat ill-suited to preservation each GRIB file consisted of six sections, each with their own header information, and required external look-up tables to be interpreted. More worrying was the range of other, undocumented file formats that existed within the archive, such as movie data encoded within the .avi format there was little evidence of appropriate preservation action to reflect such diversity. Little explicit time or effort appears to be allocated to monitoring emerging preservation strategies, with the datacentre service seemingly more focused on engaging with depositors to ensure that submitted data were appropriately packaged to limit risks of loss over time.

Dissemination and Access To facilitate the designated community's identification and discovery of content, the datacentre's website offered just a narrative description of each dataset. Search and browse functionality complemented this metadata. There was however no evidence of the use of formal description metadata standards for resource discovery such as ISADG, or EAD. Nevertheless, the requirements of their designated community were

probably met, with descriptive metadata available on the web sufficient to cater for the primary user group. A separate, but related, project called Claddier (funded by JISC) was developing further access methods, including better support for data citation. Metadata was both requested from the depositors and also created by datacentre staff. The metadata to be included in the SIP was stated for depositors, but covered only a description of the data and its implicit variables. Relationships between metadata and archival packages were maintained by storing metadata within a separate directory adjacent to the data directories of the corresponding data sets. No separate techniques, persistent links or identifiers were employed to make this association more explicit. Since staff interactions with data sets might feasibly result in disassociation of this metadata-to-dataset relationship this might be considered as something of an implicit risk.

Access to the datacentre's stored content was via two principle means, using the datacentre data browser (aka, the website, available at http://datacentre.rl.ac.uk/) and an FTP service. At the time of the audit, 8625 users were registered to access data. An Ingres relational database management system included the user database, the dataset catalogue and various metadata that supported information discovery. No formal policy was in place for informing the users about access conditions, and the datacentre staff interviewed seemed to admit that outreach to both depositors and users currently left a little to be desired. No formal policy existed, but access monitoring and statistics was built into the access system. Documentation was limited (predominantly as comments in the system scripts). Access attempts were logged as FTP login instances and accessing directories where datasets were stored. Access control was maintained at the storage level using standard UNIX user/group based security, with access rights managed on a directory-by-directory basis corresponding to where particular datasets were stored. Access cannot be controlled directly on the file level (except where directories contain only single files). The ProFTP server software used for dissemination permits access based only on the conditions specified for individual directories. Restrictions were changed manually when access requests for specific datasets were made; the user was added to the group with the appropriate permissions to read information from a specific directory.

Formal deposit agreements did not really exist for NERC data, but access conditions were set based on pre-ingest negotiations with depositors, and on the basis of data policies that related to specific NERC programmes. Datasets fell into publicly available and restricted access categories. In order to gain access to the restricted datasets users had to register and qualify for access. Registration may include sending in a signed user agreement as a paper document. In some circumstances (for instance, within one year after the data was created), principle data creators were required to endorse requested access, in order to ensure that they had the first opportunity to exploit research results. Access was not granted immediately, even for the completely automated forms, since datacentre staff were required to provide final authorisa-

tions for all registered users. This represented good practice from the perspectives of both security and user management, in the absence of more sophisticated user validation functionality. The access and user authentication systems appeared to implement the requirements that have been set by depositors. Auditors did not get a chance to witness the user registration process from the repository's perspective, and therefore it was difficult to comment on the extent to which the human decision-making process corresponded to the check-list criteria. The logging of unsuccessful access attempts was currently not really done, although any problems were formally monitored via the FootPrints helpdesk software system, which facilitated tracking and escalation of user queries. The FootPrints software and its usage policy were documented as part of the *Operations Manual*.

Because submission, archival and dissemination packages within the datacentre were generally synonymous, the need to demonstrate that the DIP (or AIP) creation process was complete and correct was perhaps less pressing. Interviews revealed that the last time someone ordered a dataset to be delivered on transfer media was a long time ago. At the time of the assessment, DIPs were delivered exclusively online. Since the data was accessed via an online interface, DIP creation was virtually a one stage process, with archival packages simply delivered via web or FTP protocols. The onus of ensuring that packages correspond to requests was placed upon the user. Subsets of web accessible data were not really supported as such.

Technologies, Technical Infrastructure and Security

Technical Platform The datacentre appeared to operate upon a combination of standard infrastructural hardware and software systems with a number of scripts conceived within the organisation aimed at achieving specific workflow goals. Standard open source software in use included (but was not limited to) GNU/Linux, ProFTP, MySQL, PostgreSQL, and Apache. Staff were computer literate above an average level, so there was no perceived problem with the management of the software side of the repository. System administration expertise was available within the core datacentre team. There were some concerns over documentation that existed most notably with respect to the bespoke scripts created for uploading, verifying, storing and downloading datasets. Over time, these scripts had been written in many different computer languages (at the time of the evaluation the most common examples were written in Python or Perl) and had not been sufficiently documented to enable re-use after longer periods of time. The solution to this was that the scripts were re-written in a new language and the old ones discarded.

Backups and Synchronisation Three principle methods were available for data backup within the Centre. These were to a local tape archive stored within a datacentre fire safe;

via rsync to separate disk storage and to a bespoke petabyte datastore, also maintained at the same facilities but around five hundred metres away. The UNIX df command revealed around 62 terabytes of content within the archive. Class A data were backed up to local tape on a mainly ad-hoc basis, and to the petabyte store as part of a regular backup job. Smaller class A data were also subject to daily backup via rsync. Class B data was subject to similar backup processes, although it was rarely if ever backed up to local tapes. Little documented backup policy surrounded Class C data; large data within this category were unlikely to be backed up at all smaller datasets may be backed up to the petabyte store for convenience. The datacentre had at least one recorded instance of data loss, when a large Class B satellite dataset was lost following a catastrophic filesystem failure, leading to the loss of everything stored on the RAID array. Reacquisition of the data was possible, albeit complicated, and its size was considerable; other data was consequently afforded higher priority, and this particular dataset was not backed up at datacentre. It was unclear what explanation was provided to the depositor and users about the non-availability of this dataset. There were also daily dumps of Ingres, MySQL and PostgreSQL databases that supported the catalogue, website and various ancillary systems within the datacentre. These were stored on tape within a fire-safe, rsynced to another disk and backed up to the petabyte datastore. With respect to Atlas storage, at least two copies of class A datasets were maintained, with at least one copy of class B examples. The datacentre should have therefore had at least 4 copies of each class A dataset. Documentation was included in the directory structure of an AIP, so was also backed up. There was one issue identified with respect to the backing up of large datasets; the physical capacity of storage systems would sometimes limit the extent to which policies could be adhered to with larger files manually split between storage volumes and partitions. The petabyte system was a Storatek tape library (produced by Sun Microsystems). The datacentre's own storage system relied upon Cyberview servers configured for RAID 5. Redundancy was therefore maintained on all disks clusters; however, although a single disk's failure could be tolerated, the failure of two or more within a single cluster would result in loss (albeit in many cases recoverable).

All of the copies of data were maintained in a common geographic area (distributed up to 500 metres). None of the redundant storage provisions could really be classified as off-site. This was of particular concern given the safety notification that one receives upon entering the site - a sounding bell denoted a firm alarm, whereas a klaxon would sound in the event of a nuclear incident! This did not appeared on any of the datacentre risk assessment documents, and although no doubt capable of destroying much of the local environment, little has apparently been done to assess the threat or to conceive of contingencies (most obviously to store redundant copies of data in a more remote location).

Tapes were checked randomly, but not systematically and there seemed to be little evidence of documentation of test results. Tapes and servers were decommissioned at regular inter-

vals, and it appeared that resource availability was not a premium concern with respect to the replacement of faulty media or infrastructural hardware. However, although procedures for replacing a tape were in place, there were no formal mechanisms for identifying faults. Similarly, although a disaster recovery plan did exist as part of the datacentre *Operations Manual*, it could have been more detailed and ought to have been tested in fire drill procedures and the test results documented.

Security Arrangements The data centre maintained fairly stringent physical access requirements around its main data store, where backup copies of the highest priority datasets were maintained. Similarly, archival/access storage facilities were subject to physical security systems. Pass card authentication was enforced within the store facility, and although visitors could be signed in, they were required to be accompanied by authorised individuals. More generally, the main site facility boasted considerable security. All employees were required to display identification badges at all times and visitors to liaise on arrival and departure with gatehouse security staff in order to be issued with a visitor's pass.

System security was perhaps less well enforced. Many of the software scripts utilised by the BADC staff ran as a shared UNIX user, which immediately let any staff member access the full range of functionality that all of the repository's collective scripts offer. Essentially a root account, this could be exploited to tremendously destructive effect if compromised, and the responsible individual would be very difficult, if not impossible, to trace.

Conclusions

The pilot assessment revealed that policies ought to have been more rigorously conceived, documented and formalised, and circulated widely among repository staff to create a culture that understands exactly what the repository is engaged in, and how it is ultimately operating. In terms of staffing, greater resources should be invested in staff skills development, most notably archival skills which were lacking in comparison with scientific expertise.

It was concluded that technical approaches to information integrity maintenance and verification could be considerably refined. Then-current checksum provisions and other information integrity measurements were largely insufficient to create an audit trail for the data and processing in the archive. Monitoring and checking schedules were well described but in practice rarely applied. Likewise, data management rarely extended beyond the association of simple web-page type information with digital datasets. More sophisticated data management provisions were required for the archive to consider itself to be OAIS compliant.

Preservation planning at the data centre was undertaken in an extremely ad hoc fashion. Dealing with preservation issues and considering pitfalls and potential solutions was not

explicitly part of anyones job, nor were there any (collective) reports written on this to influence or guide decision-making. Related to this, the stringent requirements on file formats dissuaded depositors from submitting content to BADC. A potential solution would have been to be more open minded about acceptable ingest formats, but employ people or acquire software capable of performing appropriate SIP to AIP to DIP conversion. A lack of funding was an obvious a barrier here, although on the other hand, opening up the range of ingest formats might have been a means to solicit greater funding. Fundamentally, curation activities seemed to take something of a backseat to facilitating access. To be considered a trustworthy repository demanded that the organisation embrace challenges of preservation, and determine the issues associated with mandate, legal status, services and functions that ought to be amended in support of that goal.

B.6 The US State Digital Archive

Organisational Infrastructure

The eleven state universities in Florida fell within the remit of a state-wide central board of governors. This group fulfilled various central duties and delegated a variety of other responsibilities to the individual University boards of trustees. Universities consequently enjoyed a degree of autonomy with regard to the way that they conducted their affairs. However, there were some services and programs that while of great value (and necessity) to individual institutions could not be provided independently at each. For instance, although many Florida Universities were involved in marine research, it was not feasible for each to acquire their own vessel for conducting field studies. Instead, such resources were provided at a centralised level; there were about twelve to fifteen system wide resources that were essentially shared among the eleven Universities. The datacentre was similarly structured, and with its budget generated at a central level it was accountable to each of the other University libraries and ultimately the state board of governors. Like each of the shared system-wide resources the datacentre was based at a single institution. The rationale for this was largely political, with the intention to demonstrate that money was being allocated more directly to students. Although the budget was maintained at the local level it was generated centrally, and the host institution was not permitted to revise or restructure financial allocation prior to its subsequent delivery to datacentre. The datacentre was required to report to both the host institution's provost and also the Council of State University Libraries (CSUL) made up of library directors at each state University. That board's role was officially an advisory one, but it derives a degree of power from the fact that dissatisfaction among other state libraries would have negative implications for the datacentre.

Within the datacentre a number of services were provided for University libraries. These

included the operation of a shared integrated library management system and licensing of electronic resources. The archive, upon which this study focuses, represented an additional datacentre service, and consisted of a repository infrastructure providing support for the preservation of digital collections and the digital outcomes from state university studies and research.

Mission and Mandate Eight key responsibilities were outlined for the archive within the archive *Policy Guide*, describing the major constituent parts of their activities, and their relationships with their depositors. These were to implement bit-level or full preservation of submitted content (determined according to preservation agreement); to restrict those who were authorised to deposit or sanction the withdrawal or dissemination of content; to provide detailed ingest and error-related feedback for every submitted package; to preserve 'original' files as submitted, maintaining integrity, viability and authenticity; to employ appropriate preservation strategies to persistently maintain a usable version of each file for which full preservation was sought; to provide dissemination information packages (DIPS) on request; to provide appropriate reports to affiliates for management purposes; and to ultimately achieve and maintain certification as a trusted digital repository, when the infrastructure to support this becomes available.

Affiliates were those eligible groups that had signed agreements to use the archive's services. Eligibility was limited to state university libraries within the state of Florida and their PALMM (Publication of Archival Library and Museum Materials) partners. The latter included any institution that had a formal partnership agreement with a Florida state university library to participate in one or more of these projects. Non library units within the state university infrastructure were permitted to deposit content, but this was required to be done indirectly via the responsible library at that particular institution.

The archive adopted an attitude of shared responsibility for preservation, between both the archive and its affiliates. It was perhaps for this reason that institutions were describe not passively as 'depositors', but instead as 'affiliates', suggesting a degree of mutual cooperation. To this end, as documented in the archive *Policy Guide*, affiliates were responsible for negotiating an agreement (which must be counter-signed by representatives of both their institution and the datacentre), incorporating details of authorised individuals for deposit, withdrawal and dissemination and details of projects and sub-accounts; selecting content for archiving and maintaining adequate local descriptive metadata; ensuring legal permissions were obtained and transferred to the archive (assuming liability for breach of intellectual property rights occasioned by the deposit); submitting content to the archive in the format specified in the archive Submission Information Package (SIP) specification; maintaining records of what was archived within the archive (including at minimum the entityID of the SIP and links to locally stored metadata); verifying the success of the submission process

via the generated error and ingest reports; requesting withdrawals where preservation was no longer required; and requesting dissemination when access to information was necessary.

The archive's mission statement was "to provide a cost-effective, long-term preservation repository for digital materials in support of teaching and learning, scholarship, and research in the state of Florida". This was endorsed by the datacentre board (which consisted of the CSUL group plus representatives of the Florida's Division of Colleges and Universities and Division of Community Colleges and the Florida State Librarian) lending the commitment a weight of legitimacy.

Succession Arrangements Two 'options' existed in the event of the archive's cessation of operations, as described in the archive *Policy Guide*. The first was to simply return content to the appropriate affiliate, in the form of a Dissemination Information Package (DIP). This was practically viable, although one may question whether returning content was a compelling succession arrangement. In addition, the success of this approach presupposed that the archive operations would be maintained for a period that was sufficient to permit a comprehensive dissemination. The second option, which was at the time only planned and not practically implementable, was to send content to an alternative preservation repository in a DIP exchange format (the precise format of which was yet to be conceived). While in principle much closer to a true succession plan, the practical barriers of no format and no repository greatly impeded its viability. However, discussions were already ongoing with the California Digital Library with whom research monies were being sought to collaboratively define an appropriate exchange format and it was hoped that this joint endeavour might be extended to represent a reciprocity agreement capable of facilitating succession and the remote accommodation of content.

In addition, given the anticipated public release of the DAITSS software, it was quite feasible to suggest that if widely adopted, the barriers to information exchange across common systems would be considerably lessened, and that many of the practical or technical difficulties associated with succession planning would be lessened. Nonetheless, it was acknowledged that several of the barriers associated with succession planning and feasibility are not technical, and instead are based in inter-organisational, political and legal concerns. It was suggested that the archive continues to pursue a formally expressed collaboration with CDL or another equivalent preservation repository to seek formal assurances for succession and service continuity, and to define means for effective inter-organisational digital object exchange. Although it was suggested that funding was reasonably assured for the foreseeable future there was no evidence of a legal or regulatory compulsion upon the state to continue to support the archive. It was suggested that if such assurances could not be obtained then this should be considered and documented within an overall risk mitigation strategy.

Staffing Twelve individuals had direct formal responsibility with respect to the archive, including five dedicated IT staff, but most contributions were less than full time. The datacentre's director assumed administrative responsibility for the efforts of the archive (proportionately 0.04 of an FTE), although currently many of the more hands on aspects of management were undertaken by another individual (0.3 of an FTE), whose time was distributed between the archive and other datacentre library systems. Five IT staff were employed full time in continuing to develop the DAITSS system and maintain relationships with affiliates and process the content they deposit and respond to dissemination requests. Finally, five systems administration staff dedicated a proportion of their time (from a fiftieth to a fifth of an FTE) to maintain the repository's hardware and software infrastructure and its associated security systems. Additional input came from the datacentre board. As noted above, this board convened in mainly an advisory capacity, but the nature of the organisation conferred upon it a degree of leverage, and it was generally the first stop in seeking endorsement or policy approval.

Each staff member was subject to an annual review, which provided an opportunity to reflect on the work of the previous year and offer projections for the future. In addition, any training requirements were identified by individual staff members based on their expectations of changing or emerging roles and responsibilities. The travel and training allocation was documented in the archive's prototype budget as being five thousand, five hundred US Dollars, although it was suggested during discussions that this figure was modest and that the real allocation was significantly in excess of this sum. Given the number of staff and the nature of training that was undertaken (which included attendance at international conferences and workshops), the stated figure was low. It was suggested that in addition to staff identifying their own skills shortages and training requirements that a top-down approach to training could be adopted in parallel (beginning with a more representative and reliable budgetary allocation).

The assembled team was well suited to the current activities within the archive: a significant amount of development work was still being undertaken and therefore the availability of software development expertise was essential. However, as the archive's technical infrastructure approached completion and emphases were increasingly placed on the operation of the archive it was suggested that the organisational structure may be less than optimal. Discussions revealed aspirations within the archive to introduce a charging model for archival services based on quotas, to provide better statistical information to those groups that the archive were accountable to and to build and maintain closer relationships with affiliates. Each of these would require a level of managerial and administrative input that was not feasible with existing staff allocations.

Within the self-assessment document completed by the archive staff it was suggested that the creation of an the archive manager position might be desirable; this action was strongly

endorsed, particularly with an eye to the viability of the archive's future activities. This step would also facilitate a more proactive training approach a manager could more effectively provide a top-down identification of skills gaps and learning opportunities within the archive. A managerial appointment could also coincide with a more general shift in focus from development to operations. It was suggested that this may require the acquisition of at least one more staff member to undertake a more operational role. Alternatively one or more of the current IT developers could be redeployed to more operational duties.

A final staffing concern related to an absence of documentation. At the time of the evaluation, even within the sample job descriptions surveyed, there was little evidence of granularly defined roles and responsibilities. This had obvious advantages in terms of the flexibility it provided in terms of the activities of staff members, but to some extent threatened the comprehensive fulfilment of archival responsibilities and limited the extent to which trustworthiness may be attained. It was therefore suggested that the archive should aim to describe and document the tasks being performed within the context of each position.

Designated Community The archive defined its designated community based upon what was realistic for it to achieve, and its contextual spacing, with regard to depositors and end users. Eschewing the need to cater for diverse types of access the archive described a shared responsibility with the affiliates that deposit content for long term preservation. From the *Policy Guide*, "[f]or the archive, the Designated Community is the set of professional staff of the archive affiliates. Staff members interact with the archive and serve as proxies for the constituencies they serve in the academic and research communities. They must be able to render materials disseminated to them by the archive and present these materials to users in understandable form. This may require them to write or acquire rendering software, for example METS-based page turners or media players, but it will not require extraordinary efforts, such as digital archaeology or the acquisition of obsolete software or hardware". The archive operated as a dark archive, with strictly limited access, and as described above, information packages requested from the archive may not have been immediately usable, possibly requiring transformation which should be undertaken by the appropriate affiliate.

Policy and Procedures Various aspects of repository policy were outlined in a range of publicly available documents. These included the archive *Policy Guide* itself and a range of documents within the Digital Archive Information page of the archive's website, including documentation describing the specification for valid content submissions, background reports and action plans for preservation of specific formats and practical recommendations for affiliates. Less transparent were aspects of policy that were embedded in software code an example that emerged during the audit was the procedure for dealing with inconsistencies between copies of archived materials (multiple copies were retained for redundancy, as

explained below). Discussions suggested that the software was capable of determining the authentic version and taking appropriate steps, but its method of doing so and the supporting algorithm remained unclear. Such embedded policies should have been extracted and more clearly documented.

Rights and Legal Responsibility Given its partnership-based approach, the responsibilities of the archive itself were limited, and the extent to which end user needs might vary became less relevant. Nevertheless, numerous mechanisms were in place to monitor feedback from affiliate organisations. For example, a mailing list was available for technical contacts at each affiliate university and a queue had been established within the datacentre's problem reporting system exclusively for the archive. There was little evidence of policy or designated community refinements based on the feedback received, although this may have simply indicated a satisfied community of affiliates. It was again suggested that a managerial role might incorporate responsibilities for monitoring feedback and redesigning policy to reflect the community's expectations and/or ongoing usability concerns.

Another useful consequence of the partnership model was associated with legal aspects of preservation. Liability for intellectual property rights infringements remained with the depositing organisation, and submitted content would only be accepted after standard Library agreements had been countersigned. The affiliate library was required to commit to being "responsible for compliance with all applicable copyright laws and other laws applicable to deposited materials, and that [they have] the authority to grant to datacentre non-exclusive rights to copy, display and create derivative versions of deposited files." In the event of legal challenge (which had at the time of the evaluation not occurred), the archive's policy was to disseminate the content to the owning affiliate and withdraw it from the archive. If a challenge subsequently faltered then they would replace the object without charge. The only concern associated with this approach was that it might be abused, with challenges possibly decimating the scale of the archived collection (admittedly a far-fetched scenario). Since the archive was operating as a dark archive it seemed unlikely that either hoax or legitimate challenges were likely, given that it remained impossible for non-affiliate parties to determine the nature of or access the archived content via any archive-provided mechanisms. Nevertheless, with the onus of proof of legality resting on the affiliate, this presented a potential risk.

A final point of legal concern related to materials that were associated with the digital objects, and also stored within the archive. The archive's documented 'localization' policy described a process that occurred when a submitted file contained links to other files (such as an XML file which references a DTD or Schema). In such cases, the remote, referenced file was retrieved and added to the archival content (AIP). There were obvious legal concerns, given that affiliates were required to only vouch for the legality of submitted content, not that to any referenced material. This had been earlier acknowledged, and the system

was modified (albeit without a corresponding change to the archive policy documentation) to download only a small format-specific subset of all linked files, most notably DTD and XML Schema files. The alternative would be that every online document cited within a PDF dissertation might be harvested and stored with no permission. A remaining doubt was whether legal permission was required to store these remaining file types. In all likelihood this would vary on a file-by-file basis, but it was recommended that archive staff explore, with some urgency, the legal implications of storing each of the linked schemas and DTDs within the digital collections. Liability in such cases may not be assumed to fall upon the affiliate within the current wording of the library agreement (given that it explicitly covered just 'deposited files'). Therefore, if potential legal consequences could be identified these should be addressed by either refining the text of the standard agreement or conceiving and documenting an appropriate policy that alleviates the remaining concerns.

Funding In addition to the 2002 IMLS grant which concluded in the latter part of 2005, the greatest proportion of repository funding originated from centralised, state channels. Although the budget came via the host institution, there was little to no direct independent budgetary interaction. Instead, budget plans were subject to review by the board of eleven University Library directors, who would offer their approval, assuming sufficient finances were available. Plans would then pass to the centralised, state board of governors, whereby each University's vice-president met and agreed before final ratification by the council of University presidents. Budgetary flexibility was evident, and had been exploited in the past. For instance, when systems were transferred from a costly mainframe system to cheaper UNIX systems monies were freed up, enabling the datacentre to acquire additional human resources. The base budget, allocated annually, continued automatically, although it was noted that in past periods of recession the datacentre budget has been reduced. For instance, in 1991 when the host institution was asked to reduce 3 per cent of its spending the datacentre was asked to cut that amount of their yearly budget, which was collected from the mid-year free balance. This accompanied a much wider commensurate series of public sector budget cuts within the state. Within the model, the datacentre was capable of maintaining a carry-over fund which was useful to meet costs that recur on a less than annual basis, such as replacement of expensive technological infrastructure equipment. These monies resided in a separate budget, which unlike the main datacentre budget could be accessed by the host institution. Such tampering would be likely to elicit a strong negative reaction from the other state university libraries, and was therefore considered unlikely. One consequence of the protection afforded to the datacentre budget (which further emphasised the value of the carry-over fund) was that no overspending could take place; budgetary separation meant that the University was unable to cover any deficits. At the time of the evaluation the datacentre budget had never been in the negative; rather their annual spending had consistently yielded

spare cash to carry over into the following financial year. There was also evidence of anticipatory budgeting for subsequent years, if not within the archive, certainly within the wider datacentre. For instance, when a significant proportion of the datacentre's library infrastructure moved from dumb terminals to PCs they were able to project across a five year period the anticipated budgetary requirements and spending.

There was no charging model in place for the archive services, but the library agreement countersigned by each affiliate included a caveat explaining that although no fee was currently payable, this may be introduced in the future. Discussions suggested that a quota-based system of billing might be adopted, with a view to both income generation and provoking a more thoughtful and selective approach to archiving from depositors. The administrative consequences of such a decision were thought likely to be considerable, and this again provided a clear justification for the appointment of a full time archive manager. There were many benefits associated with introducing a charging model, not least from a sustainability perspective. Perhaps the most profound was that the introduction of such a system would immediately reduce some of the concerns that surrounded the scalability of the archive; continuing without charge was quite conceivable if the level of content remained roughly the same. However, an increasingly widespread use of the archive services would introduce additional costs across the whole archive budget. If these could be mitigated by a self-sustainable, charging-based system then the archive could be less worried about attracting additional depositors. Income generation of this type provides a degree of insurance against possible future funding gaps, which as noted, cannot be met by the local host University. The datacentre director suggested in discussion that the primary operational budget was guaranteed, but given previous funding dips in periods of recession, it was realistic to think that funding may be less than expected.

There was at the time of the evaluation no formal, distinct budget for the archive, rather its allocations were consumed within an overall budget for the datacentre. This was a situation that staff were seeking to amend however, and recent efforts had been made to develop a prototype budget for just the archive, with the intention to make it increasingly independent from its organisational context. A spreadsheet detailed individual costs associated with staffing, software and hardware, and the third party hardware hosting services provided by the host institution's Computing and Networking Services and the Northwest Regional Data Centre in Tennessee. This overall budget amounted to around 550,000 USD of expenditure, a proportion of the total datacentre budget of just under 13 million USD. The most significant archive expenditure was staff salaries; given the strong suggestion above, it was likely that this would extend beyond the current 384,774 USD following the appointment of a full time manager.

Greater budgetary independence for the archive appeared appropriate. In order to not only manage but also actively demonstrate the sustainability of the archival operations it was

useful to isolate expenditure, incomes and assets (or proportions of each) that related to the archive. This would in turn facilitate business planning, and the allocation of monies for contingency. The archive could also consider maintaining a similar distinction with regard to the carry forward balance in order to ensure that cash saved in archival operations could be subsequently channelled back to cover those less frequent costs associated with archival preservation functionality. Preservation is unpredictable business and flexible assets are therefore extremely valuable, particularly in the absence of a parent organisation capable of providing support in times of financial strain.

It was suggested that greater physical separation of the archive might accompany a move towards greater financial independence. In contrast, this was not recommended, since the rich skills and other resources evident within the archive team's operational context provided scope for intellectual and resourcing economies of scale that would benefit both the archive and its associated services.

In terms of transparency, public law within the State of Florida ensured that all organisations funded by the state legislature were bound to full disclosure of financial record keeping, ensuring that transparency was maintained, and that shortcomings in accounting practice could be immediately identified and corrected.

Managing Risk There was a notable lack of evidence of appropriate risk management activity. The lack of risk-based strategy resonated throughout much of the organisational, technological and digital object management infrastructure at the archive. Emergent thinking regards digital preservation itself as a risk management exercise; by identifying contextual and object-centric uncertainties one can transform these into manageable risks, documenting their probability and potential impact, as well as any mitigation or contingency strategies one has in place to limit their likelihood or lessen the degree of harm that their occurrence (or non-occurrence) might cause. A repository with well managed (and well demonstrated) risk management was therefore one that was more likely to engender trust. Furthermore, the process of identifying risks was of value in and of itself, helping to identify areas where resource ought to be most effectively committed to overcome perceived barriers to success.

Every affiliate of the archive was required to countersign the datacentre Library Agreement, which as noted above obligates the affiliate to provide assurances about the legality of preserving the content, transfer the necessary rights to undertake preservation, and retain full liability for the illegality of any preservation activities. An appendix to this document, which could be amended and resubmitted by affiliates at any time, named the local individuals authorised to interact with the archive, and request withdrawals and disseminations and the preservation requirements. These agreement documents were maintained in paper form, with copies residing both in the datacentre offices and within the host institution's central

administration office. These had been digitised too, and it was anticipated that they would be deposited within the archive. The affiliate information that related to access and preservation policies within the system was also maintained in a MySQL database. The only 'access' to content that was available was via the archive's strictly maintained dissemination system, which would deliver content only to individuals named within the agreement appendix, and therefore issues of tracking the implementation of access rights and restrictions was less significant.

Digital Object Management

Acquisition and Ingest Digital materials selected for preservation by affiliates were submitted to the archive via the File Transfer Protocol. Upon upload they resided within a corresponding affiliate directory prior to their ingest. Affiliates were required to submit not only the content that requires preservation, but also an accompanying METS XML document that corresponded to a given schema and both referenced and described each incorporated file. Collectively, this amounted to the Submission Information Package; the archive SIP specification and METS SIP profile were documented on the datacentre web pages . Packages were generally represented using a single tier directory structure or zip, with the directory or zip-file's file-name corresponding to the name of the METS file within (the latter was suffixed .xml). Compressed or bundled files other than zip (e.g gzipped, rarred or tarred files) were not currently supported for packaging SIPs, although it was expected that this functionality would be introduced. Similarly, although the current FTP method of submission distinguished affiliates as different users with unique accounts and submission directories, it was anticipated that digital signature support would be imminently introduced to more effectively authenticate the source of materials. Part of the archive SIP Specification contradicted information provided in the DAITSS overview. Although the latter clearly indicated that .zip aggregated files were supported, the specification suggested that "[t]he the archive can not accept SIPs that were tarred or gzipped or otherwise bundled or compressed".

Identification and Naming Affiliates were not limited in terms of the file-names that they could allocate, which could result in unpredictable behaviour should multiple packages be submitted by a single affiliate with identical names. It was suggested that archive technical staff should explore the potential results of such action and if problems were evident implement more robust procedures. New arrivals to the archive were first processed with the prep module, part of DAITSS. This ensured the validity of the SIP, removing files that were not described within the corresponding METS file. In addition, support was maintained for the Metadata Exchange Format (MXF), an XML format previously defined for the SUS Digital Library and PALMM. When the prep module identified these files a conversion to METS

was performed, and the process continued as normal. If a submitted packaged contained no SIP descriptor then one could be created; this was contingent upon creation being specified as a term of the archiving agreement associated with the corresponding affiliate project or sub-account. Otherwise the package was rejected. When invalid or non-well formed SIP descriptors were identified packages were rejected and the process logged. Any files that existed within the submission package that were not documented in the associated package descriptor were rejected, although this step was not formally documented. Rather, the report that described the subsequent ingest listed each of the files that was successfully accessioned. It was therefore suggested that such disposal instances should be reported explicitly. The final action of the prep module was to place the processed package in the ingest directory for ingest processing.

Having determined the validity and completeness of submitted packages the operators proceed to the next stage of processing, characterised by the ingest module of DAITSS. After a series of simple checks to verify the completeness and correctness of packages and descriptors a selection of metadata was extracted from the SIP descriptor for subsequent use. Following this, data file objects were created to correspond to each file in the submitted package. Formats were identified (using mime-type, filename suffix and output from the UNIX file command); anti-virus checking took place (the presence of a virus resulting in rejection); agreed preservation level (full, bit-level or none) was determined by looking up the file format and project in the account's table of preservation requirements; and formats were validated (using combination of first and third party tools any format profile non-conformance was recorded, and in certain cases could result in an automatic downgrading of preservation level). Fixity checks were then implemented if the depositing affiliate supplied checksum values within their submission then these were compared to those evaluated at this stage. In the absence of affiliate-supplied values these initial checks were recorded for subsequent, ongoing comparison within the METS descriptor file. Files were then subdivided according to their implicit bit-streams (some files, such as .avi movie files contain multiple bit-streams), with each allocated a persistent identifier (consisting of the date of creation in numeric format with a daily, auto-incrementing alphabetic suffix). Individual identifiers were allocated to every intellectual entity (aka AIP), file and bit-stream. A persistent association was maintained with AIPs by encoding the identifier within the descriptor file-name, recording the identifiers within the XML content of that descriptor and logging the association in the management database. It was suggested that the archive may wish to consider the adoption of one or more identifier schemes that was capable of generating an ID that was unique within a global context. Handles and Digital Object Identifiers were two examples of possible approaches. One of the benefits of doing so could be realised if a reciprocal object transfer and storage agreement was reached with another organisation, where global uniqueness might be required. Technical metadata were automatically extracted wherever possible.

Localization was one of three preservation activities that then took place; any schema or DTDs referenced within packaged files and stored remotely were automatically retrieved and data object files created to accompany their referrer in the preservation environment. Where files were due to receive full preservation and suitable forward migration and/or normalisation methods existed these were performed, with the resulting file(s) added to the SIP, with a data file object created and allocated a persistent identifier. Metadata for intellectual entities and every data file object were created, with relationships and events documented.

Reporting was consistently undertaken throughout the ingest process, with affiliates notified via email of ingest activities that were being undertaken and of any errors that had been identified during the overall process. In addition to the delivery of XML encoded documents each event and outcome was recorded within the corresponding descriptor and in the MySQL database associated with DAITSS. Reporting enabled the archive to formally document the point at which preservation responsibility was accepted, and to describe the specific objects that were affected.

Generating Archival Packages The AIP creation procedure that exists within the archive and DAITSS more generally was well documented, although there were some notable omissions or shortcomings in terms of that which was fully described. Similarly, there were some concerns relating to a degree of bottlenecking within the ingest system that could be addressed by streamlining the physical process.

Following the initial stages of ingest associated primarily with management of the submitted package a number of steps were involved in the creation of a corresponding archival package. Each AIP corresponded to a single intellectual entity (some examples might include a volume, dissertation or home movie). Some files (perhaps those originating from digital collections at affiliate institutions) would have a preservation level of none, and in such cases these files would be excluded from the archived package. The next stage corresponded to a development decision to limit the duplication of consistently referenced files; a global directory existed to accommodate any files that may be linked to by several archived objects. There was actually little evidence of savings in terms of bandwidth or processing the remote, referenced file would still be checked in all cases in order to ensure that it remained unchanged from a previously retrieved globally stored example. The storage savings were likely to be negligible too for several reasons (including potential legal issues, as discussed above) only linked schema and DTD files were retrieved, and since these were text and comparatively small in terms of file-size the benefits appeared minimal. Conversely, the associated risks were potentially serious. Relying upon a system of shared files meant that no archived package was independently complete, and if one was acquired in isolation from the rest of the archive this may be problematic. Any value obtained from capturing remotely referenced content was at best threatened. In order to maintain the link it was necessary

to alter references to point to the global directory, not the remote resource, which could be argued to be in contrast to the archival goals. It was suggested that the archive should retire the global directory approach in favour of independently complete archival packages, despite the additional resulting storage overhead.

The AIP descriptor was created, corresponding to the original SIP descriptor but with additional documentation of all files, relationships and events that the object has been subject to within the archive. In comparison with the SIP, which was described thoroughly in the public SIP specification and associated METS SIP Profile, the archive's expectations with respect to the structure or content of AIPs were minimally documented. A short AIP definition existed within the DAITSS overview document, and the same document described the process within which SIPs were converted AIPs, but it was suggested that this should be extended, given that maintaining an understanding of the AIP was, in the longer term, a higher priority.

The final stages of ingest were to write the AIP to an output directory, write it to archival storage, which with DAITSS' redundancy support can be at any number of physical locations (the archive relies upon two local copies at host institution's CNS in Gainesville and a third at the Northwest Regional Data Center in Tallahassee) and commit the update to the management database. At the time, technological shortcomings within the DAITSS software presented some bottlenecking, as each ingest write process had to be completed prior to the software proceeding. Better software thread support would alleviate this issue and streamline the process. This was considered to be essential in order to facilitate scalability, and discussions suggested that the development was internally of high priority. The SIP was subsequently deleted from the ingest input directory, and finally confirmation information was formatted in XML and emailed to the appropriate affiliate.

Integrity Validation Although fixity checking was undertaken during the ingest process it was not performed on a periodic basis. The infrastructure to facilitate this was essentially in place, with Storage Maintenance functionality built into the system. The means to automate random or comprehensive monitoring were yet to be developed, but this was perceived as little more than adding some glue to bind the various aspects of functionality that already existed. DAITSS supported both MD5 and SHA1 message digest algorithms and was capable of recording both in association with a single object. Mechanisms and policies apparently existed for resolving a situation where a single archival package demonstrated corruption, although this was not demonstrated and remained apparently undocumented. To implement this functionality was a reasonably trivial challenge (opinions expressed by repository staff endorsed this view).

Understandability and Usability A notable shortcoming evident in the archive's checklist responses was that there was no process (documented or otherwise) for determining the

understandability and usability of archived content. It was suggested that this could be implemented in the short term by exploiting the existing communication channels that existed between the archive and its designated community. Without implementing a means for verifying ongoing understandability the archive could not confidently claim to be preserving content (other than at the bit-stream level). Given the finite breadth of its designated community it was thought to be quite feasible for the archive to establish a straightforward method. It was likely that such increased interaction with affiliates would require an additional administrative commitment; this would represent a further justification for the appointment of a full time archive manager.

Preservation Policy Within the archive the shape of preservation activities was based on the agreement between the archive and a given contracting affiliate. This agreement could describe preservation expectations as one of full, bit or none. Preservation levels were determined at the level of individual files at the ingest stage, based on the account (the identity of the particular affiliate, or the repository itself), the project code (enabling individual accounts to allocate alternative preservation levels for the same file format) and file format. Although sub accounts could also be defined within individual accounts, these were relevant only for billing and reporting purposes and were irrelevant from a strict preservation perspective. Full preservation meant that all applicable and available preservation techniques were employed, including migration, localization and normalisation. Bit preservation meant that files would be ingested and stored and subject to refreshing and integrity checks, but no further preservation methods would be employed. A preservation status of none was to accommodate content that arrived within a larger package of submitted content, which for some reason had not been isolated and removed prior to deposit.

Full preservation services were only practically applicable to a comparatively small subsection of all file formats. These had been identified based on a combination of their preservation viability and their popularity. For each format (the full range was listed on the archive information web page) a background report was prepared detailing a selection of technological characteristics, and documenting additional associated sociological or legal issues (e.g. adoption rate, licensing implications). These were internally ratified by the archive group as a whole to determine their completeness of coverage. No additional external registries (such as representation information registries) were automatically referenced although it was acknowledged that the research activity undertaken to understand each format could involve consultation with a variety of sources. Following the completion of an initial report a further document was conceived to detail the action plan that would be undertaken with respect to the corresponding format. This document represented the most critical aspect of preservation planning within the archive. In some respects, the format-centric approach has limitations in terms of the effectiveness with which one can preserve content, or more specifically, the sig-

nificant properties of individual items. The principal value of an item may relate to any one of its physical or semantic characteristics. There were implicit risks in adopting a preservation approach that dwells on formats and not objects. It was acknowledged that until relatively recently the archive had given very little consideration to the subject of significant properties at all. However an even more granular, affiliate-oriented approach should have been pursued; indeed, much of the overhead related to the identification of significant properties might have been allocated to affiliates as an additional responsibility. To date, no affiliate had explicitly notified the archive of the properties that ought to be preserved within any deposited content but it was suggested that once a suitable infrastructure was conceived to accommodate such varying degrees of preservation, it should be encouraged.

Three main preservation approaches were implemented within the archive, and reflecting the overall philosophy of the archive these were applied exclusively at ingest. In the case of full preservation the archived AIP would contain both an original bit-stream or bit-streams (that is, the originally deposited file or files) as well as the last-best migrated preservable example of that file or those files. In some cases the original and last-best preservable example would be synonymous. Normalization, Migration and Localization were all identified as means to manage format obsolescence, and based on format transformation. Normalization was intended to ensure that those files that were in formats that were less than optimal for preservation were created in a more preservation worthy format. For instance, PDF files would be normalised into a set of page-image TIFF files. Normalized files were not saved, rather the process itself was recorded as having been successful. Some question marks remained about the value of this process since the ongoing availability of a successful normalisation method relies upon the preservation of the corresponding tool or script. Migration was intended to alleviate the risk of obsolescence by creating a version of at-risk formats that was considered to be a reasonable successor to that format. This could be an equivalent but higher version example of the original format (e.g., PDF 1.4 files might be migrated to PDF 1.6) or a different format altogether. This then represented the 'last-best' preservation version, replacing any that might have existed within the AIP before. Localization, as discussed above, was intended to ensure that remotely referenced files were, wherever possible, harvested and stored locally to ensure the independent completeness of AIPs. For files subject to full preservation (as specified by affiliates in Appendix A of the datacentre Library Agreement), the appropriate preservation strategy was documented within each format-specific action plan. For those files formats that had no corresponding action plan the archive would commit to bit-level preservation until a suitable preservation strategy was identified. At that time the affected files would be disseminated and reingested, and during this process the appropriate preservation steps implemented. Discussions revealed that decisions to research and conceive background and action plans for new formats were prompted by the nature of content that had been received within the archive.

Since it took around three months to fully document a format and conceive an appropriate action plan it was suggested that the archive should seek to modularise the DAITSS code to encourage the development of format plugins from beyond the archive development team. By facilitating and motivating external development the work could be effectively shared and many more than the eighteen supported file formats (at the time of the evaluation) could be preserved. In addition, DAITSS adoption would be likely to increase and its status as a stable archiving solution increasingly consolidated. An action plan review schedule existed in order to identify when format information was approaching obsolescence, although as a result of intensive development commitments within DAITSS there had been evidence of failure to undertake some reviews in an appropriately timely fashion. A wider community of format specialists in a range of institutions would provide a considerably more effective, and ongoing means of policing to ensure that preservation planning remains both optimal and viable. Of considerable concern was the lack of regular integrity checking that was undertaken within the archive, an issue that was described above. It was hoped that the commitments made during discussions will be fulfilled, and an appropriate automated procedure will be conceived to execute fixity checking on a regular basis.

Removing Content In addition to continuing to preserve content the archive also supported withdrawal functionality to enable content to be removed from the archive. This would take place only upon the request of an authorised agent of the corresponding depositing affiliate. Although files belonging to a withdrawn AIP were deleted entirely from storage, the archive maintained a record of the object's ingest and subsequent withdrawal, with the affiliate notified of the withdrawal via an emailed Withdrawal Report. A common use of withdrawal functionality was to correct a previously submitted package. In such cases withdrawal would be followed by a subsequent ingest of the package, with any errors amended. The archive could unilaterally withdraw archived content if the preservation of specific material was subject to external legal challenge, in accordance with the policy described above.

Dissemination and Access Although the archive operated primarily as a dark archive there were examples of descriptive metadata maintained in association with archived content. The majority of descriptive metadata derived from SIP descriptors provided by affiliates. Information that would be captured when supplied by affiliates included a SIP package identifier (the only mandatory metadata), affiliate-assigned entity identifier, identifiers of external metadata records, title, serial volume and issue number. File names were also maintained for each file within the SIP. The archive would add further internal identifiers associated with each individual AIP, file and bitstream. All of this metadata was stored within the archive management database and within the corresponding AIP.

There were no end-user discovery functions incorporated within the archive or within DAITSS

software more generally. The archive had elected to place the onus for maintenance of descriptive metadata upon the affiliates, since only they were permitted to request dissemination of their own deposited materials. the archive *Policy Guide* demanded that affiliates "maintain records of what was archived within the archive [including] at minimum the entityID of the SIP and a link to any locally maintained metadata". Nevertheless there were residual concerns that the archive did not make it sufficiently clear to affiliates exactly what was required of them with respect to content description. It was suggested that the archive should define much more explicitly and specifically the metadata that must be supplied and recorded by affiliates for retrieval. It was likely that considerably closer interaction with the community would yield a greater understanding of the archive expectations; this additionally further justifies the appointment of an the archive manager.

The archive had committed most of its efforts to date to conceive effective procedures and infrastructures to support ingest and its three key preservation strategies of normalisation, migration and localisation. Only in the few months prior to this evaluation was the first dissemination functionality completed, and there had as yet been no instances of organisations seeking access to their archived materials. Nevertheless, the policies surrounding access were established, and communicated explicitly in a range of documentation, perhaps most explicitly within the archive *Policy Guide*. Archived content was disseminated via FTP when requested by an authorised agent of the depositing affiliate. A copy of the relevant AIP was placed within a special reingest directory, where it was treated as a SIP. Then, the package was subject to a reingest process, at which time files were once again identified using the latest identification techniques and subjected to the most recently defined preservation strategies. The resultant AIP was then transferred to archival storage, replacing the original AIP, and also written as an identical Dissemination Information Package to the appropriate affiliate's FTP output directory. This process ensured that the disseminated content was always as up to date as the system was capable of ensuring, and one could trace and verify that the DIP was complete in relation to the requested AIP. Upon a successful dissemination the affiliate received a comprehensive dissemination report that describes the content of the DIP.

During the evaluation some aspects of the dissemination procedure demonstrated unpredictable behaviours, and this was attributed to the newness of the software, and its comparative lack of testing. At one point, when dissemination of a recently added test object was attempted the process failed, with the system reporting that no such objected existed. Of greater concern were the means by which repository operators were required to process dissemination (and withdrawal) requests. The process was currently extremely reliant upon manual interactions, using UNIX shell interface. For disseminations a script was executed and it was up to the operator to provide details of the authorised individual requesting the content, which could be obtained by performing a manual look up of a relevant table within the management database. There was significant scope for human error. Additional automa-

tion would to some extent alleviate this problem and was actively encouraged.

The management of affiliate information (which was required to ensure the appropriate implementation of access policies) was rather crude. Database tables were updated manually and could conceivably be accidentally or maliciously altered to permit illegitimate dissemination or withdrawal requests. It was therefore suggested that more restrictive interfaces were developed to limit the opportunities for sidestepping or subverting repository policies at the database level.

Technologies, Technical Infrastructure and Security

Technical Platform The archive archive's technical infrastructure was part of a wider system associated with the datacentre as a whole. DAITSS itself runs within a Linux environment, and the chosen distribution at the archive was Red Hat Enterprise version 4. In addition to the core operating system software DAITSS relies upon Sun Java and the MySQL database server. Archival storage was managed by IBM's Tivoli software, a proprietary solution deployed in a wide and diverse range of storage environments.

Redundancy DAITSS supports multiple archival master copies, for the purposes of redundancy. Within the archive configuration three copies were maintained. A Tivoli client was installed upon the DAITSS Linux server. At ingest the system packaged and sent the archival files via Ethernet to an IBM AIX Tivoli server, based in Gainesville at the host institution's Computing and Networking Services (CNS). This in turn connected via a storage area network switch to the tape robot and library at the same location, committing two separately addressable copies of the AIP. The CNS Tivoli server also connected via the Internet to a further IBM AIX Tivoli machine located at North West Regional Data Center (NWRDC) in Tallahassee, where a third copy of the AIP was committed to tape. There were plans to simplify this model in the near future, installing a Tivoli for Linux server on the DAITSS machine, which, it was expected, would streamline the process. It was suggested that if AIP packages were stored as a series of independent files it could be difficult to maintain dissemination performance as content scales, and files becoming increasingly fragmented. It was possible that a single AIP could become distributed, requiring multiple passes over multiple tapes. This would also have implications for the expected lifetime of the media. The Tivoli storage manager could be configured to store files physically closer on archival media, although doing so could have merely transferred the processing overhead to ingest. An alternative solution would have been to package AIP content within a single, uncompressed format such as tar.

Tivoli's data management relied upon its own proprietary database, which made the system's ability to disseminate completely dependent on its availability. It was possible to export con-

tent from within the Tivoli-based environment, for instance using the tar command. This doesn't represent a significant concern; the software was in wide international use by a range of organisations, including many (such as national banks) for whom sustainable and persistent access to content were multi-million dollar concerns. IBM were very unlikely to simply withdraw Tivoli without warning, and similarly their existence seemed assured for at least the foreseeable future. While an openly accessible form would be desirable, the extent of management functionality offered by Tivoli probably outweighed the concerns associated with the proprietary barriers it presented for simple access.

Integrity Validation At the point of ingest the DAITSS system performed fixity checks to ensure that each master copy was identical. Since any AIP interactions were actioned by re-ingesting content these synchronous fixity checks would be undertaken at the point of dissemination or the execution of new preservation strategies. Staff offered a limited description of mechanisms and policies embedded in the DAITSS software code to resolve fixity inconsistencies, but as noted this was undocumented. There did not appear to be any explicit procedures or mechanisms to report bit loss or corruption to repository administration. The fact that no bit loss had yet been incurred was a weak justification for the absence of such mechanisms.

Storage media were refreshed annually, with a scheduled job within Tivoli to transfer all stored content to new tapes on the 9th of November. In addition, Tivoli supported a range of functionality to determine tape deterioration or increased error probability. A healthily paranoid level of administration was consistently maintained, and any tapes that prompted concerns would no longer be written to, and content immediately transferred to an alternative fresh tape.

System Updates System updates were undertaken based on a needs and risk based assessment. Numerous security mailing lists were subscribed to in order to determine potential problems associated with software that may need to be patched. New and update packages were installed using the Red Hat Package Manager (RPM) and updates made available via Red Hat's Update Agent. Upgrading was tested within the controlled environment in Tallahassee on legacy hardware that corresponded closely with the live configuration. This also demonstrated that the system operate adequately on even old hardware and offered a degree of assurances that its performance and functionality will be optimal on the production system. The archive staff met with system administrators on a biweekly basis providing an opportunity to plan software and hardware maintenance and customisation to suit any emerging user needs.

Backup Management Backups were performed regularly, with DAITSS system software and the MySQL management database included within the procedure. Three archival copies of AIPs were maintained at all times as discussed above. Although some datacentre software was also accommodated at the San Diego Supercomputer Center the archive content was not currently included. It was suggested that relationships should be continuously pursued with more geographically diverse organisations in order to conceive and build reciprocal agreements to mutually accommodate content. Although no complete system recovery tests had been undertaken, datacentre staff described a number of occasions where individual items had been recovered. It was noted that system administrator staff claimed to have undertaken simulations of data destruction and recovery from Tallahassee, which were apparently successful, but this was not documented in a prominent place, and appeared to be an ad-hoc test. Some aspects of this were covered in the archive's (at the time unfinished) *Continuity of Operations Plan* (COOP), conceived to meet state legislature requirements. This described the steps to overcome problems associated with disaster, although omitted to describe the specific steps required to reestablish the archive service or the location of key documentation.

There was little question of the suitability of hardware at either the CNS or NWRDC sites. Both featured dedicated machines for processing and storage, which were new and subject to appropriate renewal schedules. At Gainesville all but some networking facilities were exclusively deployed for the archive. Similarly, the tape robot at NWRDC was leased solely by the archive. At the time of the evaluation this arrangement was due to expire shortly and it was thought that for little additional financial outlay the archive would be able to buy their own tape robot, which would be housed at NWRDC.

Physical Security Infrastructure The level of security implemented at the physical facilities at Gainesville's CNS and within the associated DAITSS software was impressive. Secondary accounts of the security at Tallahassee's NWRDC suggest a similar high quality setup. Electronic locks protected the central machine room and each of the core network fiber huts. All doors opening to public spaces were configured to fail to a secure state. Alarms were immediately investigated by local staff or referred to campus police. Key fobs and proximity cards were required for access, with rights granted based on work requirements and staff integration needs. This meant that most technical staff would have access to the areas in which the archive machines were based. PIN codes were required in addition to physical fobs during non-working hours. A variety of environmental security measures were also implemented, with heat and water detection facilities subject to continuous monitoring. Uninterrupted power supply facilities provided power for all computer equipment in the event of grid failure, and a diesel generator offered a day's power for all systems before it needs to be refueled. The only perceivable shortcoming from a physical perspective were the

lack of hurricane proof windows within the central server room which were in any case due for imminent installation.

NWRDC was a bespoke secure data centre, and therefore physically optimised to ensure security, accessibility and connectivity. Non-stop security monitoring, video surveillance, air temperature and humidity control and monitoring, and redundant cooling were all available. Lightning protection, smoke detection and fire suppression and emergency power were also provided.

A notable concern was the lack of geographical diversity between the two sites, and one might conceive of a disaster (natural or otherwise) that might render both sites non-operational. The datacentre director described an informal hurricane threat assessment exercise that suggested that the chances of a single hurricane affecting both sites was very low; however, two hurricanes might occur simultaneously or in quick succession. The biggest continuity issues were largely organisational and the archive had already demonstrated a willingness to collaborate (notably with California Digital Library) to address these.

Logical Security Infrastructure Passwords were rotated every one hundred and eighty days, and were strictly enforced to include numbers, letters, punctuation, upper and lower case characters. A single database user permitted insert and delete rights, although these were applied to all tables (although according to established workflows only the affiliate user information should ever be changed by a human user). These rights could therefore be restricted to limit insert and delete privileges more strictly. DAITSS scripts were executable by the five IT staff within the DAITSS group, and config files were editable only by these individuals. Of more concern was the issue highlighted earlier was that there was potential for human error during the processing of scripts. It was suggested that applications and user interfaces should be refined to maximise automation, limit manual interactions and render the system less vulnerable to accidental or malicious misuse.

Managing Risk Another issued noted earlier concerns the creation of an organisational risk register. Its omission was of relevance throughout every aspect of archival operations. Some aspects of risk were covered in the archive's *Incident Event Threat Matrix*, but greater effort should have been invested into identifying the risks that threaten the business activities of the archive (i.e., the provision of preservation services) in economic, organisational, digital object management and information security terms. Each could be catalogued alongside details of their probability and impact, and descriptions of the repository's means to mitigate their likelihood or provide contingencies in the event of their occurrence or non-occurrence.

A further suggestion relates to certification; it was noted that system administration staff had been awarded various software certificates of competence and this was important in eliciting

trust. In addition to this, it was thought potentially valuable for the archive or datacentre more widely to welcome auditors within the organisation to certify information security provisions (according to international standards such as ISO 27001).

Conclusions

The archive provided an invaluable service to their state-wide affiliates and their efforts had been broadly successful. The infrastructure that had been established, the financial support that had been secured and the firm mandate upon which the archive was founded were all robust. As well as developing an infrastructure that corresponds favourably with much of the central work in this area, the archive staff demonstrated a keen willingness to determine the success of their efforts, and had already been quick to identify their weaknesses.

There were many suggestions incorporated in this report, but the most important were probably those that the archive had already identified themselves. The first was the appointment of additional staff, with the most high priority being a manager for the archive, who could engage with affiliates and plan and direct the future administrative and operational direction of the archive. The archive's primary goal within the near future was to increase affiliate numbers and the quantity and quality of content within it and to enhance its reputation.

There was a clearly identifiable need to engage with other organisations to facilitate secure storage that was sufficiently robust to meet the range of challenges that prejudice the integrity of our digital assets. Building relationships would enable the conception of succession or escrow arrangements, further remote storage of backed up materials, and ultimately, assuming the emergence of DAITSS as a widely adopted tool, collaboration in systems development and format description.

At the system level it was vital that the repository implemented a means for ongoing fixity checking, either conducted in a random or methodical fashion. Without maintaining assurances about information integrity until the point of dissemination there were implicit risks that even a well implemented backup strategy might fail to solve, if errors, accidents or malfeasance were noticed too late, and even backed up content demonstrates the emergent problem.

Another key point that emerged was that many of the problems being addressed in the archive's operations were dealt with on a somewhat ad hoc basis. There was little central coordination of risk or challenges, or of the operational means to overcome them. By composing its own catalogue of risks the archive could better equip itself to manage resource effectively to meet all of the challenges, at the points where the greatest threats were being faced.

On the whole though, the archive demonstrated its status as an effective and well managed

organisation. Its efforts stood up well to even considerable scrutiny according to the criteria within the RLG-NARA check-list, and also to those within comparable efforts such as the German nestor project's criteria catalogue.

B.7 The Cultural Heritage Archive

Organisational Infrastructure

Mission and Mandate The archive's primary mission as described by its staff (although not formally documented in a mission statement) was to provide access to its stored content, with preservation a notably lower priority. Discussions with repository staff confirmed that preservation was not a primary objective, and remained more of a by-product, or implicit part of providing ongoing access. However, agreements with funding organisations suggested that this view was not completely representative. At the time, and until the dissolution of the Arts and Humanities Data Servic, AHRC funding was provided with a general condition that electronic materials generated from funded activities should be deposited to this central resource for preservation. The archives funding contained a waiver to this deposit request; part of its justification in seeking this waiver was that alternative arrangements were in place for long-term preservation, and that the archives databases would "continue to be preserved and migrated".

Succession Arrangements Succession or contingency plans were vague if not non-existent. A perception existed that should the archive fail, the University which provided the operational context for the archive would assume custodial responsibility for the archival holdings and ensure their continued and ongoing availability, such was the extent to which their value was recognised. However notwithstanding this confident attitude, there were apparently no formal assurances that this would be the case.

Staffing The archive employed four academic staff including a principal archivist, two full time staff with content responsibilities and a further individual who although retired continued to contribute. A fifth, and final staff member, employed on what appeared to be a semi-consultancy basis, was responsible for the repository's technical infrastructure. This individual lived a considerable distance from the archive itself, and his presence required a car journey of several hours consequently he was rarely available on-site to deal with arising issues. Interviews revealed an internally held perception that there are insufficient staff, although this was a common contention in many working environments and was unlikely to definitively prove that numbers fall short of what's appropriate to support all functions

and services. Little evidence existed to demonstrate that duties have been formally identified, described and allocated within the archive; indeed, the suggestion was that all the staff contribute in a diverse selection of areas towards the archive's overall goals. There was little question of the competence of the archive's academic staff in terms of the content they were responsible for maintaining. Their qualifications and considerable expertise were self-evident. Similarly, the broad competence of the single technical staff member was clear. There was some suggestion of knowledge shortfalls with respect to some aspects of digital preservation, described in more detail in the Digital Object Management section below. Of some concern was the fact that there was little evidence of the availability of ongoing professional skills development, with the archive favouring an ad-hoc approach to training that implicitly required staff to learn whatever was needed during the course of business. Such a strategy may result in further knowledge shortfalls an independent mechanism for identifying training requirements where they are necessary, based on both internal needs or expectations and external developments tends to benefit staff and facilitate and legitimise their efforts.

Designated Community The archive's designated community remained determinedly broad; schools, universities, scholars or interested members of the public were identified as being part of what was a fairly wide and heterogeneous user group. Consequently, a similarly diverse knowledge base was assumed. Data was generally not annotated within the database; instead stored in a raw format and presented via the web alongside additional descriptive information. Few assumptions were made about the user community's abilities, service level expectations or available software or systems. End user software requirements were similarly undemanding; users could access the majority of the archive's materials using just a stock web browser. Notwithstanding this, in terms of the audit criteria there were recognisable shortcomings in the failure to formally document definitions or policies in a publicly accessible space. Similarly, the apparent absence of mechanisms to review or update policies over time represented a failure to comply with the strict criteria.

Technical Review and Development Discussions suggested that the archive's accommodating institution demonstrates a commitment to periodic technological review, but in reality this was performed reactively. New developments had been motivated mainly by the identification of problems, shortcomings, or loss of functionality. A good example was the introduction of the content management system to assist in the administration of the web pages that represented the database's main interface for end user access. This followed a wider institutional move to better satisfy the requirements of the Disability Discrimination Act, which cover various accessibility characteristics of web pages. Implementing the required changes to the previously static web content would have been an onerous undertaking,

and therefore an alternative technological solution was installed. Despite the lack of formal prior planning that characterised the archive's approach to technological developments the staff remained adamant that this has never resulted in a threat to the integrity or survival of digital assets within the repository. Changes that have taken place within the system were not formally documented, although an ad-hoc understanding of the system development had been maintained. Given the fact that the digital archive's lifetime has been observed from conception through each stage of development by the principal archivist it was argued that sufficient institutional understanding exists.

Stakeholder relationships Feedback from producers was rare, if not non-existent, and this related to the distant relationship between the archive and the creators of the original analogue content that was digitised to provide much of the digital archival collection. User feedback was similarly seldom referred to some eight thousand users had registered to access the advanced features of the archive, and therefore the archive was aware of their consumers' identities. However, little use was made of this information, or of further details originating from this source.

Transparency Undoubtedly the greatest concern with respect to policies and procedures was the lack of transparency, accountability and documentation that surrounded much of the archive's efforts. Notwithstanding the clear indicators of success in terms of funding consistency, user numbers and community reputation, there was little in place to facilitate understanding or sustainability, or to enable a newcomer to continue to build on the preceding efforts. In terms of accountability in particular the archive adopts a bullish approach, where the absence of a charge for their services translated to an apparent sense of non-accountability. The primary role embraced by staff appeared to be to add value to the materials, with preservation of lesser concern. This view did not appear to conform to terms of funding which compelled the archive to act as custodian and preserver of digital assets arising from funded activities; indeed the archive's own deposit waiver applications offered a commitment to undertake these activities as an alternative to the AHDS. Information integrity measures were defined, but not formally documented, although it seems that these were exclusively related to the creation, acquisition and ingest of new content. Once assets become resident within the database there was little evidence of ongoing integrity checks. No mechanisms existed to provide on demand measurements of information integrity.

Financial Infrastructure The financial platform upon which the archive was constructed appeared somewhat fragile. Funding was obtained almost exclusively from short term grants, awarded as a result of the self motivation, reputation and determination of the archive's principal administrator. Her role was such that she had become an integral part of the archive,

to the extent that its sustainability seemed to at least some extent dependent on her continued involvement. The community goodwill generated seemed linked to her, as much as the archive itself, to the extent that the two were almost indistinguishable. There was a clear perception that should she walk away it would have been extremely difficult, if not impossible for another individual to replace her. Part of the archive's funding reflects this - the principal archivist position was the only one within the archive financed centrally by the accommodating University; however, this funding was contingent on the individual then in place continuing to assume the role. No new appointment would be centrally funded. Generally speaking, the accumulation of funding was conducted in a fairly ad-hoc fashion. The European Commission, the Arts and Humanities Research Council and the British Academy were among the archive's funders. Anecdotal evidence suggested that around 30 per cent of grants applications had been successful. There had been suggestions that the accommodating University had considered providing more permanent funding, given the significant resource that the archive had evolved into, but this remained formally unsubstantiated. Despite the apparent financial instability, the archive had a degree of security in terms of its physical collection (which accompanied the digital resource), which, it was argued, the University would preserve indefinitely, although the less tangible digital assets may not enjoy such assurances. It was suggested a more significant risk was to staff positions, and not the electronic materials themselves. However, despite this concern, and the fact that employment contracts were provided on a six-monthly rolling basis, the length of service of staff was considerable, with three principal employees enjoying twenty, ten and eight years of service respectively. As described above, the principal archivist's position would continue even in the event of a cessation of funding. In such circumstances it was suggested that maintenance and delivery of the electronic resource would continue, although it would no longer grow.

Business planning Business planning appeared to be undertaken on a very short term basis in response to circumstances at any given time. The possibility for self-sustainability through paid-for-services had been explored to a limited extent; one potential revenue stream was from the provision of researcher-specific databases. The archive made it clear that while not averse to such developments, insufficient time or opportunities had so far been available for their realisation. It was argued that such changes would generate a degree of administrative workload that would stretch or perhaps exceed the capabilities of the existing infrastructure.

Risk Management Risk management was not formally documented, but seemingly well understood throughout the archive's staff. Rather than formalising risks in a risk-register or equivalent document, risks were explored and mitigated by planning for broad scenarios. It was assumed that any threatening technological consequences could be overcome by the

technological expertise available in-house, and that although depletion or cessation of funding would inhibit the data archive's growth, it would not be terminal to the continuation of delivery services. Sustainability in the event of a combination of both funding lapses and technological barriers were less well addressed. The consequences of key personnel leaving the archive were likely to be profound. It was suggested by staff that the technical director role could be assumed by another, and that the principal archivist role could be continued, given the momentum already established. However, there was a seriouse shortfall in documentation within the archive, which could exacerbate the implications of staff loss. The technological systems were documented from an end user perspective but little documentation was available for prospective developers to inherit and understand the system to the extent that it could be confidently administered. Similarly, almost every aspect of archival policies and procedures (although seemingly well established among repository staff, and reflected at least partially in the system's imposed workflow) remained undocumented.

Legal Issues The archive faced a number of potentially problematic legal concerns, which to date had been managed adequately, but, it was considered, may threaten the ongoing viability of the archive. The legal status of much of the material within the database remained quite unclear. No formal relationship was maintained with information publishers or producers; instead the archive's principal researchers operated quite independently, acquiring digital materials from analogue sources based on little more than their availability. Legal guidance had been sought in the past with regard to the dissemination of copyright controlled image materials; the suggestion then was that since the chance of rights holders seeking legal redress was negligible the archive needn't deviate from its existing practice. The repository administration offered three main justifications for continuing to distribute copyrighted materials these were the lack of charge levied by the archive for access to materials; the excellent track record that the archive had established as an authoratitve source; and the community interests that were being served it was argued that in the absence of the archive there would be no way for these demands to be met.

The lack of appropriate contracts or deposit agreements, and the legally questions surrounding data gathering procedures were a concern, and almost certainly represented a risk to the viability of the archive. The restrictions imposed on usage (content was free for personal and academic usage; copyright notices; digital watermarks and SPIFF technology used to encode copyright holders name intrinsically to images) would not necessarily satisfy content creators in the event of their legal objections. Reciprocal agreements were sporadically in place, enabling the archive to digitise content in exchange for appropriate credit on their web site and were worthwhile, but could be better formalised in order to limit the risk of legal liability. Even in those circumstances where producers or publishers directly interacted with the archive no formal written agreements existed. Irrespective of the fact that

such legal challenges might have been overcome by withdrawing content from the publicly accessible archive, the impact in terms of wasted staff time could have been considerable. The archive described only positive feedback from publishers, who according to anecdotal evidence regarded the archive's use of their materials as beneficial. Nonetheless, none would agree to waive copyright, and arrangements would have benefited from being more formally expressed.

The lack of legal controls was also problematic due to its impact on funding requirements, notably those imposed by the AHRC. Due to legal circumstances the archive had demonstrated an understandable reluctance to deposit content; it was thought that such behaviour might imply ownership. Instead the archive was required to commit to preservation activities which added to their core objectives to present content, and discussions suggested that preservation remained a very low priority for the archive, despite the fact that continued funding was contingent upon it.

Digital Object Management

The actual management of digital content within the archive fell some way short of the best practice espoused in such standards as ISO 14721 (Reference Model for an Open Archival Information System), but this was expected given that the overarching objective of the archive was the provision of access, and it was to this end that most resource was committed.

Information Properties Properties of material to be preserved were expressed as database fields within the archive's bespoke software. These consisted of a range of relevant kinds of information, and included descriptive and discovery metadata. There were no compulsory fields, and therefore no characteristics of the objects regarded as uniformly integral to preservation success. Little or no material was 'added' to the content ingested into the database. In fact, what might be identified as metadata information in almost every case represented part of the core digital object being preserved and made available to the user community. Issues surrounding the authenticity of the materials' source are largely moot, given that they arrived in physical, analogue form, and were digitised prior to ingest. Some images did arrive at the archive in digital form, mainly from contributing museums, but there was little evidence of a formally instantiated process for determining their authenticity. Completeness and correctness of accessioned content was to some extent verified within the software system; controlled terminology lists were imposed, with warnings prompted by the input of unfamiliar terms, an approval system contributed to the quality assurance process and the system offered the capability to merge records that were essentially the same. However, there was little evidence of appropriate policies for determining the extent to which content must be complete and correct, or what this precisely means within this organisational context.

Preservation Responsibility Once more, in apparent contrast to their funding requirements, archive staff argued that at no point was preservation responsibility accepted for the contents of materials that are ultimately accepted for archival storage.

Archival Storage There was little evidence of archival storage policies or procedures, and once more this mainly related to the low priority with which preservation was perceived within the archive. Notwithstanding this, some aspects of good preservation practice were in place. For instance MS Windows Globally Unique IDs were generated for each object, and these were stored within the database as part of the corresponding digital object's tuple. However, there were no visible mechanisms in place to ensure the ongoing completeness and correctness, or integrity of archived content, in addition to those that take place during the ingest process. For instance, no fixity checks were introduced; although database logs would reveal any manual or system interventions to stored content they would not record other changes or data corruptions. Such factors could otherwise have been identified using checksum data, like MD5 or SHA for instance. The discussions during the audit revealed that the web server had been compromised, although there was no evidence of database changes. This assessment of "no harm done" was questioned, especially given the shortfalls in protection described above. Irrespective, following such a compromise information security best practice was to rebuild, reverting to trusted back-ups to reassemble content. This was not done the reason given was that the archive would be imminently moving to a new machine anyway - and although University computer services were consulted for advice it appeared that a risky strategy was pursued. Some two hundred thousand records existed at the time of the assessment within the various databases, but since there was no accession log it was difficult to resolve to what extent the inventory was complete and correct.

Preservation Strategies Preservation strategies within the archive were undertaken in a fairly limited, ad-hoc manner, and were motivated or influenced by mainly non-object-centric factors. For instance, hardware refreshment was undertaken, but based on little more than resource availability, with new project funding usually provoking new hardware purchases. Insufficient resources were available to ensure the availability of other preservation or migration strategies. Other changes were introduced to facilitate delivery functionality, rather than the preservation of content; for instance, when it became clear that the system's Ingres database was incapable of accommodating image data the archive updated to a system with this functionality.

Metadata and Representation Information Representation Information was not formally relied upon, maintained or documented within the archive, although a variety of information that was associated with the digital materials might have been loosely termed as

such. For instance, a dictionary of terminology was maintained to assist in the interpretation of database entries, although neither this nor its entries were explicitly linked to relevant materials. Similarly, the SPIFF format used as a container format for images throughout the archive (itself an ISO standard) supported the recording of metadata in the form of a registrar-assigned 'license-plates', enabling various bibliographic and copyright information to be encapsulated within the images themselves.

The understandability of information content was to some extent measured on an ongoing basis; academics ran tutorials with students who actively used the database and in that respect acted as guinea-pigs; their insights were used as a gauge to assist in the determination of emerging expectations. Organisational fluidity had been demonstrated on numerous occasions by the archive, and a flexible approach to expectations had enabled systemic change on the basis of end user needs.

No minimum metadata requirements were in place within the archive in fact, since the metadata in almost every circumstance represented part of the content of the digital object, it varied greatly, and depended on little more than availability. Some records were no more than an un-captioned image there were no formal review mechanisms if metadata was short or incomplete and in system terms, no required fields when content was accessioned.

Preservation Validation In terms of preservation success, the archive was confident that it has lost no content throughout the full extent of its twenty-five year lifetime, despite a number of system migrations. This claim might be treated with a degree of scepticism given the lack of documentation about exactly what was expected to be within the collection.

Providing Access Access to the digital collection was provided exclusively via the web, and via this interface information was offered about the range of delivery options available. Access was restricted by the user managements system built into the archive's bespoke database, and for anonymous access interactions were logged. In the absence of formal access agreements, access terms were published prominently via the site, and digital watermarking and the SPIFF license plate represented measures that to some extent ensured users' adherence.

No formal means are available to demonstrate that the process that generated requested digital materials was complete (in relation to the request). However, this could have been implemented by reference to fixity information or the raw XML within the database.

All users were afforded read only access, although registration was necessary to access a comprehensive range of materials. Contribution rights were available by application, to either the principal archivist or technical director. The trustworthiness of authorised contribu-

tors, was beyond the scope of this assessment, although for completeness should be regarded as part of the repository, and therefore should be subjected to similar scrutiny.

Technologies and Technical Infrastructure

Technology Foundation The archive's system operated on the Microsoft Windows Server OS, an industry standard for web delivered materials and Microsoft's SQL server database provided the data back-end. The actual bespoke database system that provided the information environment for storage, was not well documented, but it was argued by the archive's technical director that the code, written in Microsoft's Active Server Pages was self explanatory, and could be straightforwardly inherited and understood, enabling development to continue. There were additional sub-systems with more obscure origins a Sun Java program for image zooming was available (including source code) and, although this required maintenance (specifically to introduce support for watermarks), there were few non-trivial barriers. More troublesome was the Minerva tool that facilitated the processing of SPIFF files, the container format for images within the archive. This existed only within binary form, and although it works fine on the current platform would be difficult to replace in the event of compatibility loss.

Backups and Synchronisation Backups procedures appeared adequate, with an onsite backup server providing a daily copy, backups to the institutional backup service collected every second day and off-site tapes recording weekly backups up to six months. Off-site copies, while not within the same physical building, were nonetheless stored within the University campus. No 'fire-drill' recovery had been undertaken, in a full sense, although individual files and databases had been successfully retrieved from backup storage. At the time of the audit there were plans to undertake a full experimental recovery.

Synchronisation was maintained by ensuring that any changes within the system created a brand new database record, with every tuple that was part of a single revision history sharing identifiers and associated XML content. Image information remained generally static after accession, but in the rare circumstances where this was not the case (for instance, if images were scanned poorly and required re-ingestion) a link to the original image was updated to point towards the replacement.

Technology Update and Replacement As noted above, hardware and storage media refreshment was pursued where sufficient money and opportunities were available. It was suggested that in the event of actual server failure money could be found, although in common with many aspects of the archive, no formal contingency mechanisms were in place, and this was not documented anywhere. Decisions during periods of change were generally

implemented in order to enhance the access platform. An example was the archive's move to SQL server, motivated by a high maintenance overhead of twenty separate databases and the fact that the MS Access database in use at the time was suffering from capacity issues. Updates to software systems were mainly administered centrally from within the accommodating University's network systems management services, and their testing and implementation policies were unavailable at the time of the audit. Critical changes to the system were evaluated using a trial database that ran in parallel with the live system, and could be undertaken within the live system if successful.

The archive claimed to maintain a security conscious approach had little supporting documentation. Only archive staff had physical access to the server itself (protected by card based door security systems), and administrative login rights were available only to the archive's technical director and the University network systems management services. General access was limited by user accounts, within a comprehensive and granular permissions system. System security was further facilitated using SSL to encrypt traffic between users' browsers and web servers, strong passwords were enforced, IP address-based authentication was supported and OS and database logs were consistently monitored.

Disaster planning was not documented at all, and although the archive's technical director suggested that the availability of backups guarantees that the worst outcome could be the loss of a single day's work, this should have been detailed in a significantly more considered fashion, outlining the risks, the ways in which they are mitigated and any contingencies in place.

Conclusions

Assessed according to the strict terms outlined within the audit check-list the approach adopted by the archive raised questions in a number of areas. However, given the historical success of the archive, and the esteem that it clearly enjoyed among its target communities it would be difficult to dismiss its efforts according to just these metrics. The most notable shortcoming was the lack of formal documentation that characterised much of the business activities of the archive. There can be little doubt that its current staff were competent in their positions, and that there was a shared sense of duty, responsibility and role. Similarly, the life cycle of content that was accessioned, archived and disseminated seemed well understood. Likewise, surrounding procedures were, although not formally communicated anywhere, well known. Without the discussions undertaken during this assessment there would have been little scope for forming any kind of organisational assessment. There was concern that a new staff member would face a similar struggle to understand the organisation's mechanisms, policies and scope without recourse to a resource within which they are formally, objectively and unambiguously expressed.

Associated with documentary shortcomings were issues concerning the archive's policy in a number of key areas. Legal questions abounded, and there seemed to be few formal assurances that the archive had legal authority to maintain much of its digital collections. Where agreements were in place they were generally informal or bore more similarity to 'understandings'. Relationships with organisations providing content should have been more formally established to provide the archive with the necessary protection to enable it to continue its business.

A similar problem followed with respect to the user communities closely related to internal documentation and transparency was the external issue of community trust. Based on its track record the archive had established a dedicated user base, and although one cannot dismiss the success with which this has been preserved for several years, there was a danger that without better external expression of their policies and procedures this might be threatened.

Preservation policy was perhaps even more of a widespread problem. Frequently described in this case study was an organisational uncertainty about the role of the archive with respect to preservation, and this ambiguity manifested itself in an approach to digital object management that fell short of that described in such best-practice benchmarks as OAIS for instance.

A further issue associated with sustainability that was of concern was the extent to which the staff, most notably the principal archivist were inextricably associated with the archive. Given the extent to which the latter's dedication, self-motivation and wide range of contacts had offered a degree of fiscal security to the archive during its lifetime there was a notable risk that her departure would be difficult to overcome. Partly this was an organisational concern her position existed (and was therefore centrally funded) only for as long as she continued to occupy it. The other factor was less quantifiable, but nonetheless persuasive, and was based on her unique personality and knowledge. That the archive had existed so successfully for over twenty-five years with little evidence of service disruption or data loss appeared to relate, to at least some extent, to the fact that the archive had enjoyed tremendous staff stability throughout its period of existence. Perhaps true sustainability could only be demonstrated, and this concern addressed, following a rotation of staff, where new individuals are expected to take over in key archive roles.

Overall the audit exercise identified a series of shortcomings that for the most part would probably manifest themselves only during a period of organisational disruption or change, or in the event of one or more unforeseen contingencies. However, it seemed that little was in place to mitigate such problems should they arise, and within the organisational model limited resource seemed available to do so. Many of the problems could be traced back to a lack of documentation and discussions highlighted concerns about the limited extent to which policies, procedures and legal relationships are formalised. Policies and workflows

were clearly well-ingrained into the management and archival activities needed to be more straightforwardly communicable to stakeholders in order to elicit trust.

Bibliography

[Abrams and Seaman, 2003] Abrams, S. L. and Seaman, D. (2003). Towards a global digital format registry. In *World Library and Information Congress: 69th IFLA General Conference and Council*.

[Ahmed et al., 2007] Ahmed, M., Anjomshoaa, A., Nguyen, T. M., and Tjoa, A. M. (2007). Towards an ontology-based risk assessment in collaborative environment using the semanticlife. In *Proceedings of the The Second International Conference on Availability, Reliability and Security*, ARES '07, pages 400–407, Washington, DC, USA. IEEE Computer Society.

[AIDA, 2010] AIDA (2010). Assessing institutional digital assets.

[Antunes et al., 2011] Antunes, G., Barateiro, J., Becker, C., Borbinha, J., Proena, D., and Vieira, R. (2011). Shaman reference architecture. Technical Report 57, INESC-ID.

[APA, 2011] APA (2011). Presentations from the alliance for permanent access 2011 conference, london.

[APA, 2012] APA (2012). Alliance for permanent access: Preparing for an iso 16363 audit.

[APARSEN, 2012a] APARSEN (2012a). Alliance for permanent access to the records of science network.

[APARSEN, 2012b] APARSEN (2012b). D33.1b report on peer review of digital repositories. Technical report, Alliance for Permanent Access to the Records of Science Network.

[Ball and Darlington, 2012] Ball, A. and Darlington, M. (2012). Review of dcc tools and guidance. ID number: redm6rep120202ab10.

[Barateiro et al., 2012] Barateiro, J., Antunes, G., and Borbinha, J. L. (2012). Manage risks through the enterprise architecture. In *HICSS*, pages 3297–3306.

[Barateiro et al., 2010] Barateiro, J., Antunes, G., Freitas, F., and Borbinha, J. L. (2010). Designing digital preservation solutions: A risk management-based approach. *International Journal of Digital Curation*, 5(1):4–17.

[Becker et al., 2011] Becker, C., Antunes, G., Barateiro, J., Vieira, R., and Borbinha, J. (2011). Modeling digital preservation capabilities in enterprise architecture. In *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*, dg.o '11, pages 84–93, New York, NY, USA. ACM.

[Becker et al., 2007] Becker, C., Kolar, G., Küng, J., and Rauber, A. (2007). Preserving interactive multimedia art: a case study in preservation planning. In *Proceedings of the 10th international conference on Asian digital libraries: looking back 10 years and forging new frontiers*, ICADL'07, pages 257–266, Berlin, Heidelberg. Springer-Verlag.

[Becker and Rauber, 2011] Becker, C. and Rauber, A. (2011). Decision criteria in digital preservation: What to measure and how. *J. Am. Soc. Inf. Sci. Technol.*, 62(6):1009–1028.

[Brewster et al., 2004] Brewster, C., Alani, H., Dasmahapatra, S., and Wilks, Y. (2004). Data driven ontology evaluation.

[BS 10008, 2008] BS 10008 (2008). Evidential weight and legal admissibility of electronic information. specification.

[BS 7799, 2006] BS 7799 (2006). Information security management systems. guidelines for information security risk management.

[BS 9300-003, 2012] BS 9300-003 (2012). Aerospace series. lotar. long term archiving and retrieval of digital technical product documentation such as 3d, cad and pdm data. fundamentals and concepts.

[Candela et al., 2008] Candela, L., Castelli, D., Ferro, N., Koutrika, G., Meghini, C., Pagano, P., Ross, S., Soergel, D., Agosti, M., and Dobreva, M., editors (2008). *The DELOS Digital Library Reference model. Foundations for digital Libraries*. ISTI-CNR at Gruppo ALI, Pisa.

[CCSDS, 2012] CCSDS (2012). Mission operations and information management area – digital repository audit and certification working group.

[CINES, 2012] CINES (2012). Centre Informatique National de l'Enseignement Supérieur.

[Coppens et al., 2010] Coppens, S., Mannens, E., and Van de Walle, R. (2010). Premis owl binding to workflow engine for digital long-term preservation. In Bradley, K., editor, *International Association of Sound and Audiovisual Archives, 41st International conference , Abstracts*. International Association of Sound and Audiovisual Archives (IASA) ; Association of Moving Image Archivists (AMIA).

[CRL, 2007] CRL (2007). Center for research libraries ratings schema.

[CRL, 2012a] CRL (2012a). Center for research libraries certification and assessment.

[CRL, 2012b] CRL (2012b). Center for research libraries metrics for repository assessment.

[CRL/RLG, 2007] CRL/RLG (2007). *Trustworthy Repositories Audit and Certification (TRAC): Criteria and Checklist Version 1.0*. CRL and RLG OCLC Programs.

[Dale et al., 2006] Dale, R., Choudhury, G. S., DiLauro, T., and Wal (2006). Center for research libraries - audit and certification of digital archives project - icpsr audit report. Technical report.

[Dale et al., 2007] Dale, R., Reilly, B., and Waltz, M. (2007). Center for research libraries - audit and certification of digital archives project - lockss audit report. Technical report.

[DANS, 2012] DANS (2012). About data archiving and networked services (dans).

[Dappert, 2011] Dappert, A. (2011). Risk management and digital preservation. TIMBUS Project Presentation.

[Dappert and Farquhar, 2009] Dappert, A. and Farquhar, A. (2009). Significance is in the eye of the stakeholder. In *Proceedings of the 13th European conference on Research and advanced technology for digital libraries*, ECDL'09, pages 297–308, Berlin, Heidelberg. Springer-Verlag.

[DCC, 2011] DCC (2011). Cardio online tool.

[DCC, 2012] DCC (2012). About the digital curation centre.

[DCC, 2015] DCC (2015). Dcc institutional survey 2015.

[DIN 31644, 2012] DIN 31644 (2012). Catalogue of Criteria for Trusted Digital Repositories.

[DNB, 2012] DNB (2012). About the deutsche national bibliothek.

[Doerr, 2003] Doerr, M. (2003). The cidoc crm - an ontological approach to semantic interoperability of metadata. *AI Magazine*, 24:2003.

[DPE, 2012] DPE (2012). About digital preservation europe.

[DSA, 2012] DSA (2012). List of repositories that have acquired the data seal of approval.

[European Commission,] European Commission. Community research and development information service - seventh framework programme (fp7).

[Gantz and Reinsel, 2011] Gantz, J. and Reinsel, D. (2011). Extracting value from chaos. Technical report, International Data Corporation.

[Giaretta et al., 2011] Giaretta, D., Conrad, M., Garrett, J., Longstreth, T., Lambert, S., Sierman, B., Hughes, S., and Tibbo, H. (2011). Audit and certification process for digital repositories. In *Proceedings of Ensuring Long Term Preservation and Adding Value to Scientific and Technical Data (PV) 2011*.

[Giaretta et al., 2010] Giaretta, D., Harmsen, H., and Keitel, C. (2010). Memorandum of understanding to create a european framework for audit and certification of digital repositories.

[Giaretta and Lambert, 2012] Giaretta, D. and Lambert, S. (2012). D33.1b report on peer review of digital repositories. Technical report, APARSEN.

[Greenberg and Marks, 2012] Greenberg, A. and Marks, S. (2012). Scholars portal trusted digital repository audit planning space.

[Harmsen and de Leeuw, 2010] Harmsen, J. and de Leeuw, L. (2010). *Data Seal of Approval: Quality Guidelines for Digital Research Data*. DANS.

[Hlomani and Stacey, 2014] Hlomani, H. and Stacey, D. (2014). Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey.

[Innocenti et al., 2008a] Innocenti, P., McHugh, A., and Ross, S. (2008a). Tackling the risk challenge: Drambora (digital repository audit method based on risk assessment). In *eChallenges 2008*.

[Innocenti et al., 2008b] Innocenti, P., McHugh, A., Ross, S., and Ruusalepp, R. (2008b). Assessing long term preservation of audiovisual digital contents with drambora. In Nesi, P., Ng, K., and Delgado, J., editors, *Proceedings of the 4th International Conference on Automated Solutions for Cross Media Content and Multi-channel Distribution, 17-19 November 2010, Florence, Italy*, pages 60–68. Firenze University Press, Firenze, Italy.

[ISO 14721, 2012] ISO 14721 (2012). Space data and information transfer systems. open archival information system (oais). reference model.

[ISO 15489-1, 2001] ISO 15489-1 (2001). Information and documentation. records management. general.

[ISO 15489-2, 2001] ISO 15489-2 (2001). Information and documentation. records management. guidelines.

[ISO 16363, 2012] ISO 16363 (2012). Space data and information transfer systems. audit and certification of trustworthy digital repositories.

[ISO 16919, 2011] ISO 16919 (2011). Space data and information transfer systems. requirements for bodies providing audit and certification of candidate trustworthy digital repositories (draft for public comment).

[ISO 17000, 2004] ISO 17000 (2004). Conformity assessment. Vocabulary and general principles.

[ISO 17021, 2012] ISO 17021 (2012). Conformity assessment. Requirements for bodies providing audit and certification of management systems. Competence requirements for auditing and certification of environmental management systems.

[ISO 20652, 2006] ISO 20652 (2006). Space data and information transfer systems. producer-archive interface. methodology abstract standard.

[ISO 21827, 2008] ISO 21827 (2008). Information technology. security techniques. systems security engineering. capability maturity model (sse- cmm).

[ISO 25010, 2011] ISO 25010 (2011). Systems and software engineering. systems and software quality requirements and evaluation (square). system and software quality models.

[ISO 27001, 2005] ISO 27001 (2005). Information technology. security techniques. information security management systems. requirements.

[ISO 28118, 2009] ISO 28118 (2009). Information and documentation. performance indicators for national libraries.

[ISO 31000, 2009] ISO 31000 (2009). Risk management. principles and guidelines.

[ISO 9000, 2005] ISO 9000 (2005). Quality management systems. Fundamentals and vocabulary.

[JISC, 2012] JISC (2012). About the joint information systems committee.

[KDLA, 2012] KDLA (2012). Kentucky department for libraries and archives.

[Kenney and McGovern, 2003] Kenney, A. R. and McGovern, N. Y. (2003). The five organizational stages of digital preservation. *University of Michigan Scholarly Monograph Series*.

[King et al., 2012] King, R., Schmidt, R., 0001, C. B., and Schlarb, S. (2012). Scape: Big data meets digital preservation. *ERCIM News*, 2012(89).

[Kovács and Micsik, 2005] Kovács, L. and Micsik, A. (2005). An ontology-based model of digital libraries. In *Proceedings of the 8th international conference on Asian Digital Libraries: implementing strategies and sharing experiences*, ICADL'05, pages 38–43, Berlin, Heidelberg. Springer-Verlag.

[Lawrence et al., 2000] Lawrence, G., on Library, C., and Resources, I. (2000). *Risk management of digital information: a file format investigation*. Council on Library and Information Resources.

[Library of Congress, 2008] Library of Congress (2008). Premis data dictionary for preservation metadata, version 2.0. Technical report.

[Lozano-Tello and Gómez-Pérez, 2004] Lozano-Tello, A. and Gómez-Pérez, A. (2004). Ontometric: A method to choose the appropriate ontology. *Journal of database management*, 2(15):1–18.

[Lyrasis, 2011] Lyrasis (2011). Digital preservation implications and solutions for cultural heritage institutions - 2011 workshop.

[Maedche and Staab, 2002] Maedche, A. and Staab, S. (2002). Measuring similarity between ontologies. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 251–263. Springer.

[Marketakis et al., 2009] Marketakis, Y., Tzanakis, M., and Tzitzikas, Y. (2009). Prescan: towards automating the preservation of digital objects. In *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, MEDES '09, pages 60:404–60:411, New York, NY, USA. ACM.

[Mchugh, 2009] Mchugh, A. (2009). Repositoryaudit.eu - drambora training programme.

[McHugh, 2011] McHugh, A. (2011). Collaborative assessment of research data infrastructure and objectives (cardio) - workflow planning and documentation. Technical report.

[McHugh, 2012] McHugh, A. (2012). A model for digital preservation repository risk relationships. In *World Library and Information Congress: 78th IFLA General Conference and Assembly*.

[McHugh et al., 2008] McHugh, A., Ross, S., Innocenti, P., Ruusalepp, R., and Hofman, H. (2008). Bringing self assessment home: repository profiling and key lines of enquiry within drambora. In *Archiving 2008: Program and Proceedings*. Society for Imaging Science and Technology. Reprinted with permission of IS&T: The Society for Imaging Science and Technology sole copyright owners of ?IS&T Archiving Conferences Proceedings.?

[McHugh et al., 2007] McHugh, A., Ruusalepp, R., Ross, S., and Hofman, H. (2007). *The Digital Repository Audit Method Based on Risk Assessment*. Digital Preservation Europe and Digital Curation Centre.

[Moore et al., 2005] Moore, R. W., JaJa, J. F., and Chadduck, R. (2005). Mitigating risk of data loss in preservation environments. In *Proceedings of the 22nd IEEE / 13th NASA Goddard Conference on Mass Storage Systems and Technologies*, MSST '05, pages 39–48, Washington, DC, USA. IEEE Computer Society.

[NARA, 2012] NARA (2012). What is the national archives?

[NERC, 2012] NERC (2012). Natural environment research council.

[Nestor, 2012] Nestor (2012). Network of expertise in long-term storage of digital resources - working group on digital repository certification.

[NIST, 2013] NIST (2013). Text retrieval conference (trec).

[NSSDC, 2012] NSSDC (2012). About the national space science data center.

[Ockerbloom, 2008] Ockerbloom, J. M. (2008). What repositories do: The oais model. via EverybodysLibraries blog.

[OCLC, 2012] OCLC (2012). Oclc research library partnership.

[Oltmans, 2003] Oltmans, E. (2003). Legal deposit of digital materials. *LIBER Quarterly*, 13(3/4).

[Open Objects Blog, 2008] Open Objects Blog (2008). News just in - no more funding for ahds from april 2008.

[Porzel and Malaka, 2004] Porzel, R. and Malaka, R. (2004). A task-based approach for ontology evaluation. In *ECAI Workshop on Ontology Learning and Population, Valencia, Spain*. Citeseer.

[PTAB, 2012a] PTAB (2012a). Frequently asked questions.

[PTAB, 2012b] PTAB (2012b). Preparing for an iso 16363 audit.

[Puhl, 2009] Puhl, J. (2009). Planets xcl owl ontology. Planets Project Deliverable.

[Research Information Network, 2008] Research Information Network (2008). *Stewardship of digital research data. A framework of principles and guidelines Responsibilities of research institutions and funders, data managers, learned societies and publishers.*

[RLG/OCLC, 2002] RLG/OCLC (2002). Trusted Digital Repositories: Attributes and Responsibilities.

[Rosenthal, 2014] Rosenthal, D. (2014). Trac certification of the clockss archive. Online.

[Ross and McHugh, 2006a] Ross, S. and McHugh, A. (2006a). Preservation pressure points: Evaluating diverse evidence for risk management. In *Proceedings of iPres 2006*.

[Ross and McHugh, 2006b] Ross, S. and McHugh, A. (2006b). The role of evidence in establishing trust in digital repositories. *D-Lib Magazine*.

[Ross et al., 2008] Ross, S., McHugh, A., Innocenti, P., and Ruusalepp, R. (2008). *Investigation of the potential application of the DRAMBORA toolkit: an assessment of the repository aspects in Digital Libraries*. DELOS Association. Fourth Italian Research Conference on Digital Library Systems, IRCDL 2008, Padova, 29-30 January 2008.; Fourth DELOS International Summer School on Preservation, Tirrenia, Pisa, Italy, June 2008.; Report on the use of DRAMBORA (Digital Repository Audit Method Based on Risk Assessment) for the assessment of repository aspects in digital libraries.

[SCAPE, 2014] SCAPE (2011 - 2014). Scaleable preservation environments.

[SEDAC, 2012] SEDAC (2012). About the socioeconomic data and applications center.

[Sinclair et al., 2009] Sinclair, S., Billeness, C., Duckworth, J., Farquhar, A., Humphreys, J., Jardine, L., Keen, A., and Sharpe, R. (2009). Are you ready? assessing whether organisations are prepared for digital preservation. In *Proceedings iPres 2009*.

[Society of American Archivists, 2009] Society of American Archivists (2009). Final report on accreditation project. Technical report.

[Stanescu, 2004] Stanescu, A. (2004). Assessing the durability of formats in a digital preservation environment: The inform methodology. *D-Lib Magazine*, 10(11).

[Steinhart et al., 2009] Steinhart, G., Dietrich, D., and Green, A. G. (2009). Establishing trust in a chain of preservation: The trac checklist applied to a data staging repository (datastar). *D-Lib Magazine*, 15(9/10).

[STFC, 2012] STFC (2012). About the science and technology facilities council - what we do.

[Stobo et al., 2013] Stobo, V., Deazley, R., and Anderson, I. (2013). Copyright & risk: Scoping the wellcome digital library project. *CREATe Working Papers Series*.

[Tarrant et al., 2009] Tarrant, D., Hitchcock, S., and Carr, L. (2009). Where the semantic web and web 2.0 meet format risk management: P2 registry. In *iPres2009: The Sixth International Conference on Preservation of Digital Objects*. Event Dates: October 5th and 6th, 2009.

[Thaller et al., 2008] Thaller, M., Heydegger, V., Schnasse, J., Beyl, S., and Chudobkaite, E. (2008). Significant characteristics to abstract content: Long term preservation of information. In *Proceedings of the 12th European conference on Research and Advanced Technology for Digital Libraries*, ECDL '08, pages 41–49, Berlin, Heidelberg. Springer-Verlag.

[Tzompanaki et al., 2011] Tzompanaki, K., Doerr, M., Theodoridou, M., and Havemann, S. (2011). 3d-coform: A large-scale digital production environment. *ERCIM News*, 2011(86).

[UKDA, 2012] UKDA (2012). About the uk data archive.

[Waller et al., 2006] Waller, M., Sharpe, R., and Coalition, D. P. (2006). *Mind the Gap: Assessing Digital Preservation Needs in the UK*. Digital Preservation Coalition.

[Waltz et al., 2010] Waltz, M., Reilly, B., and Jacobs, J. A. (2010). Center for research libraries - report on portico audit findings. Technical report.

[Waters and Garrett, 1996] Waters, D. and Garrett, J. (1996). Preserving Digital Information: Report of the Task Force on Archiving of Digital Information. Technical report.

[Wittenburg et al., 2006] Wittenburg, P., Broeder, D., Klein, W., Levinson, S., and Romary, L. (2006). Foundations of modern language resource archives.

[Zachman, 1987] Zachman, J. A. (1987). A framework for information systems architecture. *IBM Syst. J.*, 26(3):276–292.