Digital Preservation, Archival Science
and
Methodological Foundations for Digital Libraries

Professor Seamus Ross
HATII at the University of Glasgow
s.ross@hatii.arts.gla.ac.uk

Abstract

Digital libraries, whether commercial, public or personal, lie at the heart of the information society. Yet research into their long term viability and the meaningful accessibility of their contents remains in its infancy. In general, as we have pointed out elsewhere, 'after more than twenty years of research in digital curation and preservation the actual theories, methods, and technologies that can either foster or ensure digital longevity remain startlingly limited.' Research led by DigitalPreservationEurope and the Digital Preservation Cluster of DELOS has allowed us to refine the key research challenges—theoretical, methodological, and technological—that need attention by researchers in digital libraries during the coming five to ten years if we are to ensure that the materials held in our emerging digital libraries are to remain are sustainable, authentic, accessible, and understandable over time. Building on this work and taking the theoretical framework of archival science as a foundation this paper investigates digital preservation and its foundation role if digital libraries are to have long-term viability at the centre of the global information society..

1 Introduction

Good morning. It is a pleasure to return to another ECDL Conference and in particular to Budapest, which is one of my favourite cities.

Libraries have long played a critical role in the transmission of scientific knowledge and culture. As they undergo a metamorphosis from the physical to the virtual they continue to serve this role, although their nature and reach may be very different. Increasingly, though, as institutions invest in developing digital libraries they come to recognise that the digital assets on which their library depends—their capital assets, so to speak—are fragile and require substantial investment of finance and effort if the holdings themselves are to remain accessible over the longer term. In fact there is a rising buzz within the information management communities about the preservation of digital objects. In the next forty-five minutes I am going to talk briefly about the digital preservation challenge, about some of the concepts of archival science that might add value to the design and delivery of digital libraries, and about the research agenda for digital preservation. By my conclusion I hope that I will have stimulated in your mind thoughts for debate, and engaged more digital library researchers to contribute to delivering the digital preservation research agenda.

Digital objects break. Digital materials occur in a rich array of types and representations, are bound to varying degrees to the specific application packages (or hardware) that were used to create or initially manage them, are prone to corruption, are easily misidentified, and normally poorly described or annotated (i.e. they generally have insufficient metadata attached to them and where they do it is time constrained). Beyond maintaining the intactness of the bit stream (which is fairly straight forward), the long-term curation and preservation of digital materials can best be described as a labour intensive artisan or craft activity. While this approach may work well when the numbers and types of objects are small, their complexity narrow, and the scale of digital libraries limited, there is widespread agreement the approach will not scale to support the longevity of digital content in the diverse and large digital libraries that are emerging.

Please note that this is the author's presentation draft and is not final for publication.

Digital preservation is about more than keeping the bits—those streams of 1s and 0s that we use to represent information.[1] It is about maintaining the semantic meaning of the digital object and its content, about maintaining its provenance and authenticity, about retaining its 'interrelatedness', and about securing the context of its creation and use. At the heart of preservation initiatives lies planning and the recognition that 'digital curation and preservation is a risk management activity at all stages of the longevity pathway'.[2] In undertaking preservation individuals and organisations must 'right size' their risk. Our approach to preservation must be variable and 'digital object responsive' because for some materials held in digital libraries retaining the content is a sufficient outcome, for other material we must also retain the environment and context of creation and use, and for still other materials we must reproduce the experience of use if we are to ensure that the right information is passed to the future. Consider a library of literary texts, one of scientific reports linked to data sets, and a finally a digital library of computer games. In all these cases each rendition of a digital object must carry the same force as the initial instantiation, sometimes labelled as the original. As each instantiation is a 'performance' representing different functions and behaviours, we need ways to assess the verisimilitude of each subsequent performance to the initial one. So while digital libraries should hold a 'unique' exemplar, they will not hold originals.[3]

The likelihood that digital materials will be properly curated over time is closely tied to their recurring value or to their continued active usage. Recurring value arises from the use of digital objects for their evidentiary value, say to limit corporate liability, to demonstrate primary rights to an idea, invention or property, to meet compliance or regulatory requirements, to achieve competitive advantage, for facilitate education and learning, or to support new scholarship. Recurring value can arise when a resource can be re-exploited whether through repackaging, or release in some new and unexpected way. Certain data sets that are regularly exploited for commercial or research purposes, such as metrological, medical, or biological data sets (e.g. protein databases) are likely to benefit from a level of care that will ensure their longer term accessibility. One problem is that recurring value has variable time-depth and in some instances digital objects, like their analogue counterparts go out of fashion or use and must survive very long time periods of benign neglect before they become the subject of scholarly or commercial interest again. Digital objects do not respond well to benign neglect.

2 An Appreciation of the Problem

How widespread is the appreciation of the digital preservation problem? Just before ERPANET, a preservation activity supported under European Commissions Fifth Framework Programme, ended in November 2004 it completed one hundred case studies, some seventy-eight of which are publicly available on the ERPANET Website to understand just this issue.[4] Our studies provide insights into current preservation practices in different European institutional, juridical and business contexts as well as across both the public and private sectors. The case studies and results are complemented by research conducted elsewhere such as InterPARES,[5] the recent survey of fifteen National Libraries,[6] the DPE survey of archives and libraries in the EU Member States, the AIIM surveys in 2004 and 2005, the 2006 Digital Preservation Coalition UK survey 'Mind the

[1] S Ross, 2006, 'Approaching Digital Preservation Holistically', in A. Tough and M. Moss (eds.), *Information Management and Preservation*, (Oxford: Chandos Press), 115-153. Ross, S., (2000), *Changing Trains at Wigan: Digital Preservation and the Future of Scholarship*, National Preservation Office (British Library), Occasional Publication, ISBN 0712347178.

[2] Ross, S., and McHugh, A., 'Audit and Certification: Creating a Mandate for the Digital Curation Centre', *Diginews*, 9 (5), 2005, http://www.rlg.org/en/page.php?Page_ID=20793#article1, accessed Feb. 2006.

[3] It was work led by the National Archives of Australia that first defined the performance aspects of digital objects.

[4] ERPANET conducted around 100 case studies between 2002 and the end of 2004, of which 78 are published on the ERPANET website and are forthcoming in Ross, S. et al., *ERPANET Case Studies in Digital Preservation* (Glasgow, forthcoming 2006).

[5] http://www.interpares.org

[6] Verheul, I., *Networking for Digital Preservation. Current Practice in 15 National Libraries*. IFLA Publication Series, (München: KG Saur, 2006). An online version is at http://www.ifla.org/VI/7/pub/IFLAPublication-No119.pdf

Gap',[7] and surveys of national and local archives which Hans Hofman reported on in Enabling Persistent And Sustainable Digital Cultural Heritage in Europe[8](2004). Basically we found that:

- Awareness of the issues surrounding digital preservation varied markedly across organisations.
- The lack of preservation policies and procedures 'represent an issue that still needs a lot of attention.'[9]
- Retention policies were not often noted but where they were, they too were not necessarily implemented across the entire organisation.[10]
- There was a general recognition that preservation and storage problems were aggravated by the complexity, diversity of types or formats, and size of the digital entities.
- Few organisations took a long-term perspective and those that did were either national information curating institutions (e.g. archives) or institutions from telecommunications, pharmaceuticals and transportation sectors that felt regulatory risk exposure.
- Costs were poorly understood.[11]
- An organisational strategic approach to preservation was rare.[12]
- The value placed on the digital materials by organisations depended on how dependent the organisation was on the material for business activity; with the highest value placed on information by organisations that either saw or depended on exploiting the potential re-use of information or identified the risks associated with its not being available.
- Benefits to be derived from long-term preservation have proved elusive.[13]
- Organisations were waiting for solutions from technology developers and researchers.

Preservation of digital materials is a dynamic and evolving process. It is hard, and the hype surrounding it has made it harder. We might wonder what twenty years of digital preservation research can offer to digital libraries; I fear precious little of any real value. As I have argued elsewhere during the period members of the archives, library, records management, and research communities have worked hard to create 'an agitating buzz' about 'things digital'.[14] We have successfully socially amplified the perception of risks associated with digital materials,[15] but mainly within our community. Perhaps we have done this for such very good reasons as some of us want to ensure that our cultural and scientific memory is passed to future generations, many of us want to ensure accountability of individuals and public and private institutions in the digital age, others of us seek to create opportunities for creative and knowledge economies to emerge as a result of effective curation of digital materials, and still others of us see that the continuity and long

[7] http://www.dpconline.org/graphics/reports/mindthegap.html

[8] Hofman, H. and Lunghi, M., 'Enabling persistent and sustainable digital cultural heritage in Europe: The Netherlands questionnaire responses summary and Position Paper', 2004, *http://www.minervaeurope.org/publications/globalreport/globalrepdf04/enabling.pdf* (accessed February 2006); XLIV, presented at the *Dutch Presidency on Towards A Continuum of Digital Heritage – Strategies for a European Area of Digital Cultural Resources*.

[9] ERPANET, 2003, 'Policies for Digital Preservation', ERPANET Training Seminar, Paris, 29–30 January 2003, *http://www.erpanet.org/events/2003/paris/ERPAtraining-Paris_Report.pdf* ., p. 16.

[10] The findings of ERPANET in Europe are also borne out by evidence in the USA. In the recent case of ***In re Old Banc One Shareholders Securities Litigation***, 2005 US Dist. LEXIS 32154 (N.D. Ill., 8 December 2005), 'Bank employees testified they did not know missing documents should have been retained, and the bank did not inform employees of the need to retain documents for this litigation or have employees read and follow the electronic version of the policy that was established.'

[11] Ibid.

[12] ERPANET Case Studies, http://www.erpanet.org

[13] ERPANET, 2004. Business Models Related to Digital Preservation, http://www.erpanet.org/events/2004/amsterdam/Amsterdam_Report.pdf, 17, accessed Feb. 2006.

[14] Ross, S., 'Uncertainty, Risk, Trust and Digital Persistency. NHPRC Electronic Records Research Felloships' Symposium Lecture, University of North Carolina at Chapel Hill (2006)

[15] Kasperson, R. E., Renn,O., Slovic, P., Brown, H. S., Emel, J.,Goble, R., Kasperson, J. X., & Ratick, S. (1988). The social amplification of risk: A conceptual framework. *Risk Analysis*, *8*(2), 177–187. See also, Kasperson, R. E. (1992). The social amplification of risk: Progress in developing an integrative framework. In S. Krimsky & D. Golding (Eds.), *Social Theories of Risk (Ch. 6)*. Connecticut: Praeger.

term viability of ‘data-driven’ science depends upon how we handle the risks of digital materials. One might extend this to conclude from an analysis of the discussions which surround preservation risk that participants in these discussions could be classified within two core groups ‘risk amplifiers’ and ‘risk attenuators’, but as numerous studies have shown this is to put stress on the role of individuals in the process at the expense of the more complex social and cultural processes which result in the establishing of risk perceptions.[16] The difficulty is that what we are learning about digital preservation is being steered by the ‘agitated buzz makers’, those players who are socially constructing and manipulating our views of preservation risk. These are individuals who are setting the research agenda and are determining what will be the focus of our risk mitigation developments. Indeed as a result we might even mistakenly conclude that in creating ‘an agitating buzz about things digital’ that the preservation community has in a post-modern sense socially constructed preservation risk.

Nothing could be further from the truth. Preservation risk is real: it is technological. It is social. It is organisational. And it is cultural. In fact our heritage may now be at greater risk because many in our community believe that we are making progress towards resolving the preservation challenges. If you contrast two classic statements of the digital preservation issues Roberts 1994 with Tibbo 2003 it is obvious that though our understanding of the problems surrounding digital preservation has advanced the approaches to preservation remain limited.[17] So what we have done is to ‘socially construct a buzz about digital preservation’. There are lots of us talking about the problem now. We have not done sufficient underlying research necessary either to deliver the range of preservation methods and tools that are need or to allow us to reason effectively about risks or how to manage them in the same way as say an engineer might do so in the construction industry, or a transport safety expert might, or an epidemiologist in a hospital might. Some of the work that DigitalPreservationEurope, the digital preservation cluster of the DELOS NoE, and the Digital Curation Centre (UK) has done in risk management such as the DRAMBORA[18] toolkit that my colleagues have been developing will enable us to reason about risk at repository level, but we need similar tools to reason about risk at object level.

3 Digital Libraries and Archival Science

Scientific communication required a new mechanism for managing its scholarly production, dissemination, and preservation. Digital Libraries are proposed as a solution; there are lots of them —ACM, IEEE, Springer, or Elsevier Digital Libraries come to mind. But what exactly is a digital library? As I am certain that not all of us would agree on the same definition I am going use one that I prepared for the National Library of New Zealand as part of a review of their digital preservation initiatives and as a result it emphasises preservation. For my purposes here let us think of a digital library as ‘the infrastructure, policies and procedures, and organisational, political and economic mechanisms necessary to enable access to and preservation of digital content.’ [19] But if we are thoughtful about digital libraries we easily observe that they may be libraries by name, but they are archives by nature. The content they hold is essentially unique and does not really need to be held elsewhere because net-based services mean it could be provided where ever and when ever it was wanted from a single source. When users access the content from these domains they expect to be able to trust and verify its authenticity (although not necessarily its reliability), they require knowledge of its context of creation, and they demand evidence of its provenance. These are processes to which archives respond well because they have developed an appropriate theoretical

[16] Pidgeon, N., Kasperson, R. E., & Slovic, P. (2003). *The Social Amplification of Risk*. Cambridge: Cambridge University Press.

[17] Roberts, D., ‘Defining Electronic Records, Documents and Data,’ *Archives and Manuscripts* 22 (May), 1994, 14-26. Tibbo, H. R., ‘On the Nature and Importance of Archiving in the Digital Age.’ *Advances in Computers*. v. 57, 2003, 1-67.

[18] McHugh, A., Ruusalepp, R., Ross, S., Hofman H.: Digital Repository Audit Method Based on Risk Assessment. Digital Curation Centre (DCC) and DigitalPreservationEurope (DPE), (2007), http://www.repositoryaudit.eu

[19] Ross, S, *Digital Library Development Review*, National Library of New Zealand, (Wellington, 2003), http://www.natlib.govt.nz/files/ross_report.pdf, p. 5

framework and have operationalised that in repository design, management, and use. The archival framework meets requirements surrounding the production, management, dissemination, preservation, and curation needs of information. While these notions originate in the world of archival science they equally well belong to the world of digital libraries.

Modern archival science began in the 17th century with development of diplomatics initially by Papenbroeck and then by Mabillon.[20] Modern archival practice developed in the same early modern period in response to the need to manage distant conquests and distributed trans-national trading companies and economies. This early modern period experienced an information and documentary explosion. Over four centuries archival practice and science has responded well to the changing information production landscape. Its core principles of provenance, appraisal, context, trust, authenticity, and repository design and management have become more and more refined as the communication and information production and use landscapes have evolved.

While an effort to define a formal foundation for digital libraries in archival science would require an exploration of each of these concepts, I shall her only touch on three topics today: diplomatics as a tool, and the twin concepts of authenticity, and provenance.

Digital Library users need to know where the digital materials came from, who created them, how they came to be deposited, how they were ingested (e.g. under what conditions, using what technology, how the success of the ingest was validated), why they were created, where they were created, how they were created, and they need information as to how the digital object was maintained after its creation (e.g. was it maintained in a secure environment, was the software used to store and represent it changed). There need for this knowledge increases as the distance between the point at which the information was created and deposited in the digital library and it comes to be used becomes greater. These questions are ones which Diplomatics, a core tool in archival science, provides the theoretical framework to investigate in it seven core tests for information objects: quis, quid, quomodo, quibus auxiliis, cur, ubi and quando?

Of course the origin of the world diplomatics itself ought to concern us, because it continues to evolve. In fact although Mabillon's methods in their rigor, transparency, and methodological precision mirrored those scientific giants who were his contemporary, he failed to give a good definition to the term diploma. It's meaning has been debated into the late 20th century with the conservative view constantly reigning in more broad minded thinking. Ficker (1878), von Sickel (c. 1880s), Redlich (1907), Steinacker (1927), and many others moved to narrow the applicability of diplomatics to juridical documents.[21] But just over 50 years ago, Georges Tessier (1952) argued that diplomatics was applicable to all classes of 'documents' and not just to juridical ones.[22] Luciana Duranti, who has pioneered the revitalisation of diplomatics for the digital age, has argued for its relevance to electronic records. Broadly speaking diplomatics provides a critical apparatus to study any information object. There is no reason to limit its applicability to information objects represented as documents it can equally well be applied to all information objects held in a digital library, such as images, audio, and databases.

For information objects diplomatics assists us with assessing a digital objects provenance which relates an information object to its origin, lineage or pedigree. Provenance is central to archival practice and to our ability to validate, verify, and contextualize digital objects. It captures the pedigree or lineage of a digital entity. Within the archival context the significance of knowledge about provenance came to be reflected in how objects were managed. So archivists beginning in the late 18th and early 19th century archivists rejected approaches to the organization of information objects along such lines as subject, content, and place of creation in favour of

[20] Mabillon, Jean., *De re diplomatica libri VI*, Paris, 1681, 1709 and 1789.

[21] Ficker, J., (1877-8). *Beiträage zur Urkundenlehre*, 2 vols., Innsbruck. von Sickel, T., (1861-82). "Beiträage zur Diplomatik IñVIII," *Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften*, Vienna;

[22] Tessier, G., (1952), *La diplomatique*, Paris : Presses Universitaires de France.

respecting its environment of creation and its original order. To be fair the significance of provenance within the archival community emerged from experience and in part in response to the flood of documents that were arriving at the doors of archives in the 18th century and to the cultural milieu of the 19th century Europe which emphasised classificatory, evolutionary, and analytical thinking. of 19th century Europe. By leaving objects in the order in which they were found, the felt 'would facilitate a deeper understanding of their inherent meanings'.[23] In fact provenance is of critical importance to another archival concept, that of appraisal, where the disposition of digital objects are determined. In archival appraisal the relevance of provenance to the identification of evidential value is central. Of course in the digital age knowledge of provenance continues to be essential, as Peter Buneman of Edinburgh University has argued in the context of database, we can both retain the knowledge of provenance at all levels of granularity and even repackage the entities along lines of pertinence to user requirements.

Digital preservation aims to ensure the value of digital entities. As the work of the InterPARES Task Force on Authenticity concluded, 'When we work with digital objects we want to know they are what they purport to be and that they are complete and have not been altered or corrupted.'[24] These twin concepts are encapsulated in the terms Authenticity and Integrity. As digital objects are more easily altered and corrupted than say paper documents and records, creators and preservers often find it challenging to demonstrate their authenticity. As digital objects that lack authenticity and integrity have limited value as evidence or usefulness as an information resource. How many of us would wish to use a biological dataset where we could not verify the authenticity of the materials it contained. The ability to establish authenticity of and trust in a digital object is crucial.[25] A well-documented chain of custody helps with establishing authenticity.

Authenticity means different things to different communities—within the domain of history of art or antiquities the notion of authenticity can be either very rigid, as in the case of the Warhol Foundation approach to validating 'authorship' in Warhol works or flexible as in the judgement of the UK legal case of Thomson vs Christie's where 70% certainty that an object was what Christie's claimed it to be was good enough for the presiding judge.[26] It can main or kill as in the case of tainted drug manufacturing or cause copyright and business concerns as in the case of handbags and watches. At the heart of establishing authenticity lies trust and this is an area where as Clifford Lynch has noted we are just beginning to understand the issues.[27]

We live in a post modernist world, and as the innovative archival theorist, Terry Cooke, has poignantly noted: 'The postmodernist tone is one of ironical doubt, of trusting nothing at face value, of always looking behind the surface…'[28] While this is a topic that could be the subject of much new research at both practical and theoretical levels here we can only draw attention to the issue:

- ♦ As a user, how do I know that a digital object is an authentic instantiation of the version that was deposited within the digital library?
- ♦ Confronted with digital objects most users begin from a position which presumes

[23] Natalis de Wailly (1841), S. Muller, S., Feith, J.A., and Fruin, R. (1898), *Handleleiding voor het ordenen en bescrijven van Archiven, Groningen*.

[24] InterPARES Authenticity Task Force, *Authenticity Task Force Report in The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project*, (Vancouver, 2002), http://www.interpares.org/book/index.cfm,.

[25] Ross, S., 'Position Paper on integrity and authenticity of digital cultural heritage objects', *Integrity and Authenticity of Digital Cultural Heritage Objects, Thematic Issue* 1, 2002, 7-8. also available at http://www.digicult.info

[26] http://news.bbc.co.uk/2/hi/uk_news/england/norfolk/3727623.stm

[27] MacNeil, H., 'Providing Grounds for Trust: Developing Conceptual Requirements for the Long-term Preservation of Authentic Electronic Records,' *Archivaria*, 50, 2000, 52-78. MacNeil, H., 'Providing Grounds for Trust II: The Findings of the Authenticity Task Force of InterPARES', *Archivaria,* 54, 2002, 24-58.

[28] Terry Cooke, (2000), 'Archival Science and Postmodernism: New Formulations for old Concepts' *Archival Science*, vol. 1, no. 1 (2000): 3-24.

authenticity ('Presumption of Authenticity'). They assume that unless there is evidence to the contrary if the holder of a digital object says that the object is authentic, that it is.

- There are few ways that a user could even begin to determine whether a digital object is what it purports to be where they lack access to the details of the process by which the digital object was created, ingested, and managed. They can only do this if institutions have adequately and transparently documented the processes of digital entity ingest, management, and delivery.

Not to confuse the issues at this stage, but it is worth recognising the distinction between authentic and reliable information. Once material comes to be held in a digital library or repository it must be immutable if we are to accept it as authentic. But many digital libraries contain unreliable information—in fact even unreliable data can tell its own story if its provenance, context, and purpose can be ascertained.

We are just coming to grips with archival science and diplomatics as components of a theory of information object management and a foundation for digital libraries. An growing number of researchers are moving into this discussion area.

4 Research Agenda

Given the core dependency of digital libraries on guaranteeing the authenticity, integrity, interpretability, and context of the digital material across systems, time, and context digital preservation/curation must be at the heart of any future digital library research agenda. If Digital libraries are to function in this new technological environment they will need to be transparent, accessible, and responsive to user needs and expectations. Contemporary research in digital libraries tends to emphasize such research topics as personalisation, architecture, representation, retrieval, presentation and access. And the investigation of digital preservation has been limited. My impression from browsing through the past five years proceedings of ECDL and JCDL is that most digital library research tends to focus on the here and now. The addition of a preservation cluster to DELOS NoE was a visionary move by Costantino Thanos and Vittore Casarosa in recognition that digital libraries were not just about communicating with the present, but that they were mechanisms to communicate with the future.[29] Preservation is rarely seen as central to digital library design and development, those us working in the team led by Donatella Castelli to develop the DELOS Digital Library Reference Model are only just coming to grips with how to incorporate preservation into what is emerging as an outstandingly robust framework for digital libraries.[30]

That said while some might argue that research in the area of digital preservation has been innovative, in reality it has been far from sufficient to underpin projected digital library developments and the increasing complexity of digital entities. The current generation of solutions which centre on, for example migration and emulation, are unrealistic and focus too heavily on narrow aspects of the problem—they are the kinds of solutions that I described above as artisan. The challenges of effectively ingesting heterogeneous materials into a digital library (e.g. the digital materials created by contemporary writers after they die, or the data sets generated by scientific teams) will only be viable if the processes can be automated and authenticated. Even where it is possible to ingest and effectively document the digital materials drawn into a digital library, these materials will remain in an environment susceptible to constant technological change. As a result digital curation must be continuous and dynamic, this can only happen if it is automated, and the ways we describe (e.g. the context of the objects and the objects themselves), represent, and manage digital entities radically changed.

Despite all the discussions in recent years about what kinds of research are needed in the area of digital preservation, no concise and well-developed strategy that represents the views of a broad

[29] http://www.delos.info

[30] http://www.delos.info/index.php?option=com_content&task=view&id=345

community has yet emerged. Since 1989 twelve have been published. One of the tasks of DigitalPreservationEurope (DPE) has been to look at the digital preservation landscape and to come up with a research agenda that might be taken forward under FP7, as well as at national levels within the Member States. Based on an extensive crosswalk of existing preservation research agendas, the DPE Research Roadmap's objective is to provide a concise overview on the core issues which have to be addressed in future digital preservation research.[31] To construct the framework my colleague Holger Brocks lead us to examine the challenges of preservation from five vantage points: digital object level, collection level, repository level, process level, and organisational environment which covers creation and use. So for instance at the object level we focus on migration and emulation, experimentation, acceptable loss, and at the collection level we examine interoperability, metadata, standardisation, and at the process level we look at issues such as automation.

First and foremost the DPE research agenda responds to the lack of progress that has been made in the delivery of preservation solutions, methods, and techniques over the past twenty years. Second it recognises that as those working in the discipline came to better understand the issues they extended the research domain into areas that were originally seen as peripheral to digital preservation. In response we have narrowed the research agenda and argued that as a research community we must capitalise on ancillary work done in other domains such as semantic-enabled information infrastructures, grid-based resources, and service oriented architectures. We have agreed that there are really nine themes and one core methodological approach that researchers in preservation need to adopt—these nine themes also bring digital preservation inline with traditional preservation activities in analogue world.

Digital objects break. This can occur when storage media become damaged, software and hardware become obsolete, applications become lost, or bit streams become corrupt. When they break they must be restored. What processes can we use to ensure the syntactical completeness of digital objects and what methods will enable us to address semantic opaqueness. Computer forensics research has led to some restoration methods, but we need more experimental research in this area to develop effective restoration technologies.

Whereas restoration deals with objects that have broken methods for conservation enable us to address challenges that may arise with digital entities before the damage has become too severe, much as we might conserve a post-1830s printed book by de-acidifying it before brittle book syndrome takes hold or adopt preventative medicine. Transcoding, migration, emulation, virtualisation, information extraction, metadata enhancement, and semantic annotation technologies are all examples of methods that we might deploy to facilitate the conservation of digital objects. Here again there are few methods that we can take-off the shelf, we just have not done the research.

Operational and organisational research into the management of digital objects, collections, and repositories is needed. Research needs to focus on planning, enacting, executing, managing, and monitoring of organisational processes for repositories.

We have argued elsewhere that digital preservation is a risk management problem. Hence decision making instruments are needed which will enable digital preservation practitioners to translate the uncertainties involved in digital preservation into quantifiable risks that can be managed.

Our understanding of the properties that digital objects must retain overtime if the objects are to remain semantically meaningful, authentic, reliable and usable whether for rendering or analysis remains limited.

Repositories handle collections of digital objects as opposed to just discrete entities. It is the integrated nature of these collections that provide some degree of contextuality to the individual objects. Moreover collections often only gain real value when they can be integrated with

[31] http://www.digitalpreservationeurope.eu

collections held by other repositories. The research that has been done into interoperability across generations of systems, time, and repositories has been limited. More must be done.

The sheer quantity digital objects with which digital libraries need to deal means that we need to do much more in terms of automation of processes than we have done in the past. Areas where automation has promise include: metadata extraction, preservation planning and action, and selection and appraisal. To date the tools that support automation of processes are quite limited, require human intervention, and do not scale. Again we just have not done the underlying experimentation and structured research.

Gaining the semantic meaning of digital objects and even collections depends upon retention of contextual information. How was the object created? How was it used? What was the legal or social context of its value? What kind of processes are necessary to construct context and meaning. Research into contextuality is needed.

The ninth area is storage technologies and methods—on the one hand this is an engineering problem and on the other this is a deployment problem. The digital library community has much to offer the preservation community through its research into the GRID and its collaborative initiatives in the domain of eScience.

You may wonder why issues such as metadata are absent from this list and this is because they cut across many research lines from interoperability to contextualisation.

Until recently, much preservation research has been theoretically led and little of it has actually involved well-designed experimentation. Every aspect of preservation research from characterisation of digital objects to preservation planning to user needs analysis can benefit from experimentation. Some of the newer research and support activities related to digital preservation in Europe, such as The Digital Curation Centre (DCC) in the UK[32], DigitalPreservationEurope (DPE), CASPAR (Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval)[33], PLANETS (Preservation and Long-term Access through NETworked Services),[34] Digital Preservation Cluster of the DELOS Network of Excellence in Digital Libraries (DELOS-DPC),[35] an numerous other projects I might mention reflect the realisation that we need to be much more experimentally driven in our research endeavours if we are to progress the digital preservation research agenda.

5 Conclusion

So what take-away points do I want to leave you with.

As a community we need to re-think how we are approaching digital curation research. We need to engage digital libraries researchers in this process. Digital libraries are more akin to archives than they are to traditional libraries. Research in digital preservation could in general be more rigorous, methodologically founded, repeatable, verifiable, contextualised, and more effectively reported; that is it could conform better to the ‘scientific paradigm’. It needs to more ‘experimental’ than it has been up to now, something which as I have noted a host of new research projects are attempting to inspire. These experimental results will provide us with mechanisms to predict more accurately the likelihood of certain conditions arising, and a better appreciation of how to measure the implications of uncertainties associated with digital objects and longevity pathways. So, not only do we need to try to better understand what we might do to alleviate obstacles to the longevity

[32] http://www.dcc.ac.uk

[33] http://www.casparpreserves.eu/

[34] www.planets-project.eu

[35] http://www.dpc.delos.info

of digital materials, but we must do more to define the uncertainties related to digital preservation and to convert these uncertainties into known, measurable, and mitigateable risks. We should of course make a genuine distinction here between perceived risk and ‘actual’ risk; an actual risk represents an assessed and measurable risk.

I might humbly suggest that digital libraries must adopt a theoretical stance. As I noted above library science is devoid of theoretical foundations and of a knowledge-base that is relevant to the budding digital world. Archival science with its principles of uniqueness, provenance, arrangement and description, authenticity, appraisal, and its tool sets such as diplomatics, may offer us a framework for a theoretical foundation for digital libraries.

Perhaps you will be surprised that I have not come here today and told you that those of us working in digital preservation have solved the problems and we are just waiting for those of you working in digital libraries to ask us to integrate our solutions into your work. But we have not solved the problem. So my final message is that the value of digital libraries rests very much in their ability to communicate our cultural and scientific knowledge to the future, if we are to do this we must address the digital preservation challenges and to do this we need to be more collaborative, better co-ordinated, and even competitive.

Prof Seamus Ross, Budapest, 17 September 2007