Digital Preservation, Archival Science

and

Methodological Foundations for Digital Libraries[1]

Seamus Ross
HATII at the University of Glasgow,
Digital Curation Centre (UK) and DigitalPreservationEurope (DPE)
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 (DPE) 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 sustainable, authentic, accessible and understandable over time. Building on this work and taking the theoretical framework of archival science as bedrock, this paper investigates digital preservation and its foundational role if digital libraries are to have long-term viability at the centre of the global information society.

1 Introduction – The Significance and Scope of Digital Preservation

Good morning. It is a pleasure to return to another ECDL Conference and, in particular, to one in Budapest, one of my favourite cities.[2]

Libraries have long played a critical role in the creation and transmission of scientific knowledge and culture.[3] As they undergo a metamorphosis from the physical to the

[1] Please cite as: S. Ross (2007), *Digital Preservation, Archival Science and Methodological Foundations for Digital Libraries*, Keynote Address at the 11th European Conference on Digital Libraries (ECDL), Budapest (17 September 2007).

[2] Acknowledgements: I wish to extend my thanks to the Chairs of ECDL2007, László Kovás, Norbert Fuhr and Carlo Meghini, for inviting me to deliver the Opening Keynote Address. I am grateful to my colleagues in HATII, Sarah Jones, Perla Innocenti and Andrew McHugh, and to Professors Ross Harvey (Visiting Professor at HATII at the University of Glasgow 2007 and Digital Curation Centre UK Research Fellow), Dagobert Soergel (College of Information Studies at the University of Maryland) and Helen Tibbo (School of Information and Library Science at the University of North Carolina at Chapel Hill, NC) for their comments on this paper. I am indebted to my colleagues in DigitalPreservationEurope (DPE) who are collaborating on developing the Research Roadmap and especially to Holger Brocks of the FernUniversität in Hagen.

[3] L. Casson, *Libraries in the Ancient World*, (New Haven, CT: Yale University Press, 2001); W. Hoepfner (ed.), *Antike Bibliotheken*, (Mainz: Philipp von Zabern, 2002); M. Battles, *Library: An Unquiet History*, (London: Vintage, 2004).

virtual, they continue to serve this role, although their nature and reach may be very different in the future. 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 may require substantial continued investment of finance and effort, if the holdings themselves are to remain accessible over the longer term.[4] 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 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 discussion and encouraged more digital library researchers to contribute to addressing the challenges posed in the digital preservation research agenda.

Digital objects break. Digital materials occur in a rich array of types and representations. They are bound to varying degrees to the specific application packages (or hardware) that were used to create or manage them. They are prone to corruption. They are easily misidentified. They are generally poorly described or annotated; they often have insufficient metadata attached to them to avoid their gradual susceptibility to syntactical and semantic glaucoma. Where they do have sufficient ancillary data, these data are frequently time constrained. Beyond maintaining the intactness of the bit stream (which is fairly straightforward), 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 of objects are small, the diversity of their types is restricted, their complexity narrow, and the scale of digital libraries limited, there is widespread agreement that the handicraft approach will not scale to support the longevity of digital content in the diverse and large digital libraries that are emerging.

Digital preservation is about more than keeping the bits – those streams of 1s and 0s that we use to represent information.[5] 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 information about the context of its creation and use. Measured planning and the recognition that 'digital curation and preservation is a risk management activity at all stages of the longevity pathway' are critical aspects of the preservation process.[6] In undertaking preservation planning and

---

[4] S. Ross, 'Reflections on the Impact of the Lund Principles on European Approaches to Digitisation' in *Strategies for a European Area of Digital Cultural Resources: Towards a Continuum of Digital Heritage*, (Den Haag: Dutch Ministry of Culture, 2004), 88-98. http://www.digitaliseringerfgoed.nl/sites/cultuurtechnologie/contents/i000263/ocw_conclusieboekje.pdf or http://eprints.erpanet.org/103/01/sross_denhaag_dutch_paper.pdf; S. Ross, 'Strategies for Selecting Resources for Digitization: Source-Orientated, User-Driven, Asset-Aware Model (SOUDAAM),' in T. Coppock (ed.), *Making Information Available in Digital Format: Perspectives from Practitioners*, (Edinburgh: The Stationery Office, 1999), 5-27.

[5] S. Ross, 'Approaching Digital Preservation Holistically', in A. Tough and M. Moss (eds.), *Information Management and Preservation*, (Oxford: Chandos Press, 2006), 115-153; S. Ross, *Changing Trains at Wigan: Digital Preservation and the Future of Scholarship*, (London: British Library, National Preservation Office, 2000), ISBN 0712347178, http://portico.bl.uk/services/npo/pdf/wigan.pdf; S. Ross and A. Gow, *Digital archaeology? Rescuing Neglected or Damaged Data Resources*, (London & Bristol: British Library and Joint Information Systems Committee, 1999), ISBN 1900508516, http://www.ukoln.ac.uk/services/elib/papers/supporting/pdf/p2.pdf

[6] S. Ross and A. McHugh, 'Audit and Certification: Creating a Mandate for the Digital Curation Centre', *Diginews*, vol. 9, no. 5 (2005), http://www.rlg.org/en/page.php?Page_ID=20793#article1; S. Ross and A. McHugh, 'The Role of Evidence in Establishing Trust in Repositories', *D-Lib Magazine*,

action, individuals and organisations must adopt a level of risk that reflects their preservation objectives and capabilities both organisational and technical. Our approach to preservation must be variable and 'digital object responsive':

- for some materials held in digital libraries retaining the content will be 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 be able to reproduce the experience of use if we are to ensure that the right semantic representation and information is passed to the future.

As examples of these three classes of preservation, consider a digital library of literary texts, one of scientific reports linked to data sets, and 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 erroneously labelled as 'the original'. As every instantiation is a 'performance' representing a range of functions and behaviours, we need ways to assess the verisimilitude of each subsequent performance to the initial one and clear definitions of 'acceptable variance'.[7]

Although we have, as yet, no statistically substantiated grounds for making this claim, access over time to digital objects appears closely correlated to their continuous use for 'business' purposes, and to their perceived and actual recurring value. Recurring value arises from the use of digital objects for their evidential, information or commercial value. From an evidentiary perspective they might be used to:

- limit corporate liability;
- demonstrate primary rights to an idea, invention or property;
- meet compliance or regulatory requirements;
- achieve competitive advantage;
- facilitate education and learning; or
- support new scholarship.

Recurring value may result from the re-exploitation of materials through leasing them, their sale in new kinds of packaging or contexts, or their release in some new and unexpected way. Certain data sets that are regularly exploited for commercial or research purposes, such as meteorological, diagnostic (especially medical), digital maps, or biological data sets (e.g. genomic or protein databases) are likely to benefit from a level of persistent care that will ensure their longer-term accessibility. 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 what Professor Helen Tibbo of

---

July/August, vol. 12, no. 7/8 (2006), (also published in *Archivi e Computer*, August 2006), http://www.dlib.org/dlib/july06/ross/07ross.html

[7] This approach is most elegantly described in UNESCO [National Library of Australia], *Guidelines for the Preservation of Digital Heritage*, (Paris: UNESCO, 2003), http://unesdoc.unesco.org/images/0013/001300/130071e.pdf Indeed, we have done little to provide mechanisms to establish 'verisimilitude' between initial and subsequent instantiations. A paper presented at ECDL 2007 by Lars Clausen of the Statsbibliotek in Denmark is a good example of the kind of work that needs to be done in this area: L. Clausen, 'Opening Schrödingers Library: Semi-automatic QA Reduces Uncertainty in Object Transformation', L. Kovás, N. Fuhr and C. Meghini (eds.), *Research and Advanced Technology for Digital Libraries*, 11th European Conference, ECDL 2007, LNCS 4675, (Berlin: Springer, 2007), 186-197.

the University of North Carolina at Chapel Hill has called 'benign neglect' before they become the subject of scholarly or commercial interest again.[8] As a result of the constant evolution of technology, the degradation of storage media and the ever-increasing pace of 'semantic drift', *digital objects do not, in contrast to many of their analogue counterparts, respond well to benign neglect.*

2 An Appreciation of the Problem

How widespread is the appreciation of the digital preservation problem? The answer is not encouraging. Just before ERPANET – a preservation activity funded under the European Commission's Fifth Framework Programme – ended in November 2004 it completed one hundred case studies involving companies and public sector organisations in an effort to investigate this question. Some seventy-eight of these case studies are publicly available on the ERPANET website.[9] These 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, including but not limited to research by InterPARES;[10] a survey of fifteen National Libraries;11 the DPE survey of archives and libraries in the EU Member States;[12] the AIIM surveys in 2004 and 2005;[13] the 2006 Digital Preservation Coalition UK survey 'Mind the Gap';[14] and surveys of national and local archives which Hans Hofman

---

[8] H.R. Tibbo, 'On the Nature and Importance of Archiving in the Digital Age,' *Advances in Computers*, vol. 57 (2003), 1-67.

[9] ERPANET conducted around 100 case studies between 2002 and the end of 2004, of which seventy-eight are published on the ERPANET website and are forthcoming in S. Ross *et al*., *ERPANET Case Studies in Digital Preservation* (Glasgow, 2007). See also S. Ross, M. Greenan and P. McKinney, 'Digital Preservation Strategies: The Initial Outcomes of the ERPANET Case Studies' in the *Preservation of Electronic Records: New Knowledge and Decision-making,* (Ottawa: Canadian Conservation Institute, 2004), ISBN 0-662-68620-9, 99-111. ERPANET, with funding from the Swiss Federal Government and the European Commission (IST-2001-32706), led by the Humanities Advanced Technology and Information Institute (HATII) at the University of Glasgow (United Kingdom) and its partners the Schweizerisches Bundesarchiv (Switzerland), ISTBAL at the Università di Urbino (Italy) and Nationaal Archief van Nederland (Netherlands), worked between November 2001 and the end of October 2004 to enhance the preservation of cultural and scientific digital objects.

[10] http://www.interpares.org; L. Duranti (ed.), *The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project*, (San Miniato: Archilab, 2005).

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

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

[13] AIIM – the Enterprise Content Management industry association – reports are only available to members or for a fixed fee, http://www.aiim.org/. The 2005 study 'Electronic Communication Policies and Procedures' was conducted by AIIM and Kahn Consulting.

[14] M. Waller and R. Sharpe, *Mind the Gap: Assessing Digital Preservation Needs in the UK*, (York: Digital Preservation Coalition, 2006), http://www.dpconline.org/graphics/reports/mindthegap.html. In Section 3, Background and Methodology, of the 'Mind the Gap' report we discover that the questionnaire on which this study was based had been sent to '900 individuals', but we are not told how they were selected. As a result, the reader has no information as to whether or not these 900 constitute a representative sample of any particular constituency and if so what community that is. The project team received 104 responses, a response rate of just over 10%. The authors do not explain whether or not the 10% is representative of the cohort of 900 and if so in what way it is representative of that cohort. The problems are further exacerbated by the way the results of the survey are subsequently used. For example, how are we to understand the sentence: 'In contrast, fewer than 20% of the organisations surveyed have some kind

reported on in *Enabling Persistent And Sustainable Digital Cultural Heritage in Europe.*[15] Basically, as a result of the ERPANET Case Studies, it is safe to conclude that:

- awareness of the issues surrounding digital preservation varied markedly across organisations, and even across different divisions of the same organisation;
- 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 where failure to adopt best practices creates higher levels of the regulatory risk exposure than in other sectors;
- an organisational strategic approach to preservation was rare;[16]
- the lack of preservation policies and procedures within organisations was 'an issue that still needs a lot of attention';[17]
- retention policies were not often noted but where they were, they too were not necessarily implemented across the entire organisation;[18]
- 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;
- costs were poorly understood;
- benefits to be derived from long-term preservation have proved elusive and arguments which might convince commercially minded business leaders of the benefits are restricted;[19]
- the value placed on the digital materials by organisations depended on how much the organisation relied 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;
- organisations were waiting for solutions to be delivered by technology developers, researchers and service providers.

---

of digital preservation strategy in place.' Is this 20% of the 900 or 20% of the 104 that responded? Each interpretation would tell a very different story in my view. It is fair to say that in general these kinds of methodological problems are inherent in much of the survey-based research that is done in electronic records management, digital preservation and data curation.

[15] H. Hofman and M. Lunghi, 'Enabling Persistent and Sustainable Digital Cultural Heritage in Europe: The Netherlands Questionnaire Responses Summary and Position Paper', presented at *Towards A Continuum of Digital Heritage–Strategies for a European Area of Digital Cultural Resources*, (2004), http://www.minervaeurope.org/publications/globalreport/globalrepdf04/enabling.pdf

[16] ERPANET Case Studies, (Glasgow, 2004), http://www.erpanet.org

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

[18] The findings of ERPANET in Europe are also borne out by evidence in the USA. In legal cases involving the securities and financial sectors more generally staff often report that they were ill advised about how they should handle records. For instance, *In re Old Banc One Shareholders Securities Litigation*, 2005 US Dist. LEXIS 32154 (N.D. Ill., 8 December 2005), '[b]ank 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.'

[19] ERPANET, *Business Models Related to Digital Preservation*, (Glasgow, 2003) http://www.erpanet.org/events/2004/amsterdam/Amsterdam_Report.pdf, 17.

Preservation of digital materials is a dynamic and evolving process: the methods are changing, as are the technical requirements. It is hard, and the hype surrounding digital preservation has made it even 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 this period members of the archives, library, records management and research communities have worked relentlessly to create 'an agitating buzz' about 'things digital'.[20] Indeed, where preservation is concerned, the 'risk amplifiers' have taken the high ground from 'risk attenuators', as is evident from the growth in the number of publications, conferences and conference presentations during the past ten years that stress how essential it is that we overcome the obstacles to the longevity of digital materials. Through our discussions we have socially amplified the perception of risks associated with digital entities,[21] but mainly within our own community. It would seem appropriate to conclude that we have done this with the best of intentions. As curators of our cultural and scientific memory we want to ensure that we pass our information heritage to future generations in viable form. We recognise that the accountability of individuals and public and private institutions in the digital age depends on the preservation of digital materials. We acknowledge that reuse over time of digital materials will produce opportunities for the growth of creative and knowledge economies. We know that, as the transition from 'in vitro' to 'in silico' science gathers pace, the longer-term viability of this new scientific paradigm requires that we curate digital materials in ways that ensure their reusability. While we might conclude that a small band of 'agitated buzz makers' have alone socially constructed our views of preservation risk, we know from other domains that the process of establishing risk perceptions involves complex social and cultural processes and depends on more than just the actions of individuals.[22] Indeed, as a result we might even mistakenly conclude that, in creating 'an agitating buzz about things digital', individuals within the preservation community have in a post-modern sense socially constructed the impression and notions of preservation risk without a basis in reality.

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 truth, 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 – as I have done elsewhere – one is to contrast two classic statements of the digital preservation challenges, Roberts 1994 and Tibbo 2003, it is obvious that, although our understanding of the challenges surrounding digital preservation has become richer and more sophisticated, the approaches to overcoming obstacles to preservation remain limited.[23] Ross Harvey's comprehensive examination of the landscape of preservation, *Preserving Digital Materials*,[24] similarly points to only a few

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

[21] R.E. Kasperson, O. Renn, P. Slovic, H.S. Brown, J. Emel, R. Goble, J.X. Kasperson and S. Ratick, 'The Social Amplification of Risk: A Conceptual Framework'. *Risk Analysis*, vol. 8, no. 2 (1988), 177–187. See also R.E. Kasperson, 'The Social Amplification of Risk: Progress in Developing an Integrative Framework', in S. Krimsky and D. Golding (eds.), *Social Theories of Risk (Ch. 6)*, (Westport, CT: Praeger, 1992).

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

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

[24] R. Harvey, *Preserving Digital Materials*, (München: K.G. Saur, 2005). A reading of U.M. Borghoff, P.

implemented preservation methods, and the preservation approaches he examines appear to be best characterised as handicraft. The preservation community has not yet carried out sufficient underlying experimental and practical research either to deliver the range of preservation methods and tools necessary to support preservation activities or to provide us with sufficient data to reason effectively about preservation risks or how to manage them. We need to be able to reason about preservation risks 'in the same way as, say, an engineer might do in the construction industry, or a transport safety expert might, or an epidemiologist in a hospital might'.[25] While the work that DigitalPreservationEurope (DPE),[26] the Digital Preservation Cluster of the DELOS NoE[27] and the Digital Curation Centre (UK)[28] have done in the risk management area, such as the development of the DRAMBORA[29] (Digital Repository Audit Method Based on Risk Assessment) toolkit which enables organisations to reason about risk at the repository level, is worthy of mention, we need similar tools to reason about risk at the object levels as well.

3 Digital Libraries and Archival Science

Scientific communication and 'in silico' research required a new mechanism for managing its scholarly production, dissemination and preservation. Digital Libraries appeared as a solution; there are lots of them – in the realm of scientific communication, the ACM,[30] IEEE,[31] Springer[32] or Elsevier[33] 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 to use one that I prepared for the National Library of New Zealand as part of a review of their digital preservation initiatives, which as a result emphasises preservation. For our 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.'[34]

Rödig, J. Scheffczyk and L. Schmitz, *Long-Term Preservation of Digital Documents: Principles and Practices*, (Heidelberg: Springer, 2003) gives the same impression.

[25] S. Ross, (2006) 'Uncertainty, Risk, Trust and Digital Persistency', (see above note 20).

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

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

[28] http://www.dcc.ac.uk; C. Rusbridge, P. Burnhill, S. Ross, P. Buneman, D. Giaretta, L. Lyon and M. Atkinson, 'The Digital Curation Centre: A Vision for Digital Curation', in *Proceedings IEEE's Mass Storage and Systems Technology Committee Conference on From Local to Global: Data Interoperability - Challenges and Technologies*, (2005); an online version is at: http://eprints.erpanet.org/archive/00000082/01/DCC_Vision.pdf

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

[30] http://portal.acm.org/dl.cfm

[31] http://ieeexplore.ieee.org/Xplore/login.jsp?url=/Xplore

[32] http://www.springerlink.com

[33] http://www.elsevier.com/

[34] S. Ross, *Digital Library Development Review*, (Wellington, NZ: National Library of New Zealand, 2003), http://www.natlib.govt.nz/files/ross_report.pdf, 5. This definition is broad enough to encompass the new classes of 'digital libraries', such as YouTube (http://www.youtube.com) and Flickr (http://flickr.com), which are interactive, participatory, dynamic and user-driven.

But if we think more carefully about digital libraries we easily observe that they may be libraries by name, but they are archives by nature. The content they hold does not really need to be held elsewhere because net-based services mean it can be provided from a single source wherever and whenever it is wanted. Digital libraries, therefore, can hold 'unique' exemplars. 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 framework and have operationalised it in repository design, management and use over at least three centuries. The archival framework meets requirements surrounding the production, management, selection, dissemination, preservation and curation needs of information. It also supports a layering of services from repository services at the foundation to user services at upper levels. 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 the development of diplomatics.[35] Much of 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.[36] Beginning in the late 16th century there was an unprecedented information and documentary explosion and this trend has continued into the digital age. Over three centuries archival practice and science has responded well to the changing environment of information production and use. Its core principles of authenticity, trust, context, provenance, description and arrangement, and repository design and management evolved during this period and have become more and more refined as the communication and information production and use landscapes have evolved. Others such as appraisal have emerged more recently. While an effort to define a formal foundation for digital libraries within the context of archival science would require an exploration of each of these notions in turn, today I shall touch only on three topics: diplomatics as a tool, the concepts of authenticity, and provenance.

Digital library users might wish to know where the digital materials came from, who created them, why they were created, where they were created, how they were created, 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), and they may need information as to how the digital object was maintained after its acquisition by the digital library (e.g. was it maintained in a secure environment? have changes in hardware and software had an impact on the digital object in question?). If they were to require or seek such data, they could legitimately expect to be able to acquire this information relatively easily. Their need for this knowledge increases in line with the increase in the time between the point at which the digital object was created and deposited in the digital library and when it comes to be used. Diplomatics, a core tool in archival science, provides the theoretical framework to investigate such questions.

---

[35] J. Mabillon, *De re diplomatica libri VI*, (Paris: Sumtibus Caroli Robustel, 1709). (Although originally issued in 1681 I am familiar only with the revised edition of 1709.)

[36] One need only think of the 80 million pages of documents in the Archivo General de Indias (Seville) representing the records from the Conquistadores to the end of the 19th century or the 14 kilometres of records of the East India Company beginning in 1600 (and its various reincarnations after 1858) to see the scale on which documents were being created during the period (see http://www.bl.uk/collections/iorgenrl.html).

In their rigour, transparency and methodological precision, the methods of Jean Mabillon, the Benedictine monk who solidified the foundations of diplomatics, mirror those of the generally better known scientific giants who were his contemporaries, including Robert Boyle, Edmond Halley, Robert Hooke, Antoni van Leeuwenhoek, Marcello Malpiqhi and, of course, Isaac Newton.[37] The 'information object' domains to which theorists and practitioners have applied diplomatics have evolved since the early thinking of Mabillon and Papenbroeck[38]. Early scholars, such as Thomas Madox, felt diplomatics was most appropriately applied to 'instruments' such as charters.[39] For nearly two centuries the prevailing intellectual wind, as represented in manuals of diplomatic practice and introductions to what we regard now as classic studies of documents, held that the concepts of diplomatics should really only be applied to juridical documents – the conservative view consistently reigned in more broadminded thinking.[40] But during the 20th century attitudes firmly changed. For instance, Georges Tessier, Professor of Diplomatics at L'École Nationale des Chartes from 1930 until 1961, argued that diplomatics was applicable to all classes of 'documents' and not just juridical ones.[41] This view has been increasingly adopted by other scholars. Luciana Duranti, Professor of Archival Science at the University of British Columbia, who has pioneered the revitalisation of diplomatics for the digital age, has argued for its relevance to electronic records.[42] Indeed, through her leadership of InterPARES 1 and 2 she has led a broadening of the conceptualisation of records from including 'records produced and/or maintained in databases and document management systems' to 'records produced and/or maintained in interactive, experiential and dynamic environments.'[43] Duranti has thereby broadened the types of objects to which diplomatics could be effectively applied. Leonard Boyle, the eminent scholar and, rumour has it, reluctant Prefect of the Vatican Library, in an elegant essay that I first read just thirty years ago this month while an undergraduate and that remains the finest succinct discussion of diplomatics of which I am aware, argued: … 'it seems much more realistic and far less precious and selective to describe diplomatics as the scholarly investigation of

---

[37] Jean Mabillon (b. 1632 – d. 1707) and Daniel van Papenbroeck (b. 1628 – d. 1714). Contemporaries: Robert Boyle (b. 1627 – d. 1691), Edmond Halley (b. 1656 – d. 1742), Robert Hooke (b. 1635 – d. 1703), Antoni van Leeuwenhoek (b. 1632 – d. 1723), Marcello Malpighi (b. 1628 – d. 1694), and of course Isaac Newton (b. 1643 – d. 1727).

[38] The groundbreaking work of Daniel van Papenbroeck (b. 1628 – d. 1714) is worthy of discussion, but time does not permit me to do it justice here.

[39] Thomas Madox, *Formulare Anglicanum*, (London, 1702).

[40] For example, the conservative views of J. Ficker, *Beiträge zur Urkundenlehre*, (Innsbruck, 1877-8) is a good example of this kind of conservative thinking.

[41] G. Tessier, *La diplomatique*, (Paris: Presses universitaires de France, 1952), (3rd edition, 1966). Georges Tessier, 'Diplomatique', in *L'Histoire et ses méthodes*, in Charles Samaran (ed.), *Encyclopédie de la pléiade 11*, (Paris, 1961), 633–76.

[42] L. Duranti, 'The long-term preservation of accurate and authentic digital data: the InterPARES project,' *Data Science Journal*, vol. 4 (2005), 106-118. (http://www.jstage.jst.go.jp/article/dsj/4/0/4_106/_article); L. Duranti, 'Concepts and Principles for the Management of Electronic Records,' *The Information Society. An International Journal*, vol. 17 (2001), 1-9; L. Duranti, 'Diplomatics: New Uses for an Old Science. Part VI,' *Archivaria*, vol. 33 (1991-92), 6-24; Diplomatics: New Uses for an Old Science. Part V, *Archivaria*, vol. 32 (1991), 7-24; 'Diplomatics New Uses for an Old Science. Part IV,' *Archivaria*, vol. 31 (1990-1), 10-25; 'Diplomatics: New Uses for an Old Science. Part III,' *Archivaria*, vol. 30 (1990), 4-20; 'Diplomatics: New Uses for an Old Science. Part II,' *Archivaria*, vol. 29 (1989-90), 4-17; 'Diplomatics: New Uses for an Old Science,' *Archivaria*, vol. 28 (1989), 7-27.

[43] L. Duranti and K. Thibodeau, 'The Concept of Record in Interactive, Experiential and Dynamic Environments: the View of InterPARES,' *Archival Science*, vol. 6, no. 1 (2006), 13-68 (Online: http://dx.doi.org/10.1007/s10502-006-9021-7)

any and every written documentary source, juridical, quasi-juridical, or non-juridical.'[44] Moreover, there is no reason to limit its applicability to information objects represented as 'physical documents;' it can equally well be applied to all information objects held in a digital library, whether still or moving images, audio, vector graphics, and data (and even data held in databases). Broadly speaking, diplomatics provides a critical apparatus to study any information object and this process was encapsulated for Boyle in seven mechanisms to investigate the veracity of an information object: *quis*?, *quid*?, *quomodo*?, *quibus auxiliis*?, *cur*?, *ubi*? and *quando*?[45]

Diplomatics assists us to assess a digital object's provenance, which relates its origin, lineage or pedigree. Provenance is central to archival practice and to our ability to validate, verify and contextualise digital objects.[46] Within the archival context the significance of knowledge about provenance came to be reflected in how documents were managed. So, archivists beginning in the late 18th and early 19th century rejected approaches to the organisation of information objects along such lines of pertinence as subject, content and physical place of creation in favour of respecting the environment of creation and the original order in which the documents had been created and used.[47] To be just a little more precise the significance of provenance within archival practice emerged not merely in response to the flood of documents that were arriving at the doors of archives, but from a combination of experience, the cultural milieu of the period which emphasised classificatory practices and evolutionary thinking, and a belief by historians that if material was to be retained in its original order researchers would be able to hear the voices in the documents more accurately, more richly and with a more precise semantic appreciation.[48] As Michael Roper, former Keeper of the Public Record Office (London), put it, 'the provenance or context of archives remained a vital means of assessing the source, authority, accuracy and value of the information which they contained for administrative, legal [….] research and cultural uses.[49] In fact, provenance is of critical importance to another archival concept, that of appraisal,[50] where the disposition of digital objects is

---

[44] L.E. Boyle, 'Diplomatics,' in J.M. Powell (ed.), *Medieval Studies: An Introduction, (*Syracuse, NY: Syracuse University Press, 1976), 69-101, at p. 75.

[45] Boyle (1976), 79-90. Had, for example, Hugh Trevor-Roper (Lord Dacre) adhered to these principles of analysis, which depend upon asking questions about *who, what, in what manner (e.g. form, formulae, style), with what support, aid or help, why (e.g. what purpose), where, and when*, when he acted as a member of the group engaged to determine the authenticity of the 'Hitler Diaries' in 1983, he might not have been led astray. One could cite dozens of other examples, including some in which the materials in question were held within archives. When these principles are applied they can assist scholars, as is evident in the study by L. Berlin and H. Craig Casey, 'Robert Noyce and the Tunnel Diode', *IEEE Spectrum*, (May 2005), 42-45. See especially page 43 where the authors describe the process of validating copies made from pages of Noyce's laboratory notebooks.

[46] See the essays in K. Abukhanfusa and J. Sydbeck (eds.), *The Principle of Provenance: Report from the First Stockholm Conference on Archival Theory and the Principle of Provenance (2-3 September 1993),*' Skrifter utgivna av Svenska Riksarkivet 10 (1994), ISBN: 91-88366-11-1.

[47] N. de Wailly (1841). See M. Duchein, 'Theoretical Principles and Practical Problems of Respect des fonds in Archival Science', *Archivaria*, vol. 16 (Summer 1983), 64-82.

[48] S. Muller, J.A. Feith and R. Fruin, *Handleiding voor het ordenen en bescrijven van archiven,* (Groningen, 1898).

[49] M. Roper, 'Archival Theory and the Principle of Provenance: a Summing-up', in K. Abukhanfusa and J. Sydbeck (eds.), (1994), 187.

[50] Ross Harvey, 'Appraisal and Selection', in S. Ross and M. Day (eds.), *DCC Digital Curation Manual*, (Glasgow: Digital Curation Centre, 2006), http://www.dcc.ac.uk/resource/curation-manual/chapters/appraisal-and-selection provides an excellent introduction to the issues of appraisal.

determined. Of course, in the digital age knowledge of provenance continues to be essential, as Peter Buneman and his colleagues at the University of Edinburgh have argued in the context of databases.[51] In the flexible digital libraries (and digital archives for that matter) we can both retain the knowledge of provenance at all levels of granularity and even repackage the entities along the lines of pertinence if this is required to meet specialised user needs or expectations.

Digital preservation aims to ensure the maintenance over time of the value of digital entities. As the research of the InterPARES Task Force on Authenticity concluded, '[w]hen 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.'[52] These twin concepts are encapsulated in the terms authenticity and integrity.[53] Digital objects that lack authenticity and integrity have limited value as evidence or even as a source for information. 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. How many of us would be comfortable if our doctor were to use a clinical trials data set in which he/she could not verify the authenticity of the materials it contained to plan a regime of treatment? The ability to establish authenticity of, and trust in, a digital object is crucial.[54] A well-documented chain of custody is one factor that helps with establishing authenticity.[55]

Authenticity has become a 21st-century challenge that reaches into every corner of modern life. Of course, authenticity means different things to different communities – indeed, even within a single domain its meaning can vary from rigid to flexible, as a contrast between the Warhol Foundation approach to validating 'authorship' in Warhol works[56] and the judgement in the UK legal case of Thomson vs Christie's demonstrates for the art world.[57]

---

[51] P. Buneman, A. Chapman and J. Cheney, 'Provenance Management in Curated Databases', in *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, (Chicago, IL: 2006), 539-550. http://portal.acm.org/citation.cfm?doid=1142473.1142534; P. Buneman, S. Khanna and W.C. Tan, 'Why and Where: A Characterization of Data Provenance', in *8th International Conference on Database Theory (ICDT 2001)*, (2001), 316-330, http://www.springerlink.com/content/edf0k68ccw3a22hu/; P. Buneman, S. Khanna, K. Tajima and W. Chiew Tan, 'Archiving Scientific Data', *ACM Transactions on Database Systems (TODS)*, vol. 29 (2004), 2-42, http://portal.acm.org/citation.cfm?doid=974750.974752

[52] InterPARES Authenticity Task Force, *Authenticity Task Force Report in The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project*, (Vancouver, 2004), http://www.interpares.org/book/index.cfm

[53] L. Duranti, 'Reliability and Authenticity: The Concepts and Their Implications,' *Archivaria*, vol. 39 (Spring, 1995), 5-10.

[54] S. Ross, 'Position Paper on Integrity and Authenticity of Digital Cultural Heritage Objects', *Integrity and Authenticity of Digital Cultural Heritage Objects*, DigiCULT Thematic Issue 1, (2002), 7-8; also available at http://www.digicult.info

[55] From the point of view of the police this is seen in: The National Hi-Tech Crime Unit produced for the Association of Chief Police Officers, (n.d.), *Good Practice Guide for Computer Based Electronic Evidence*, (version 3.0), http://www.acpo.police.uk/asp/policies/Data/gpg_computer_based_evidence_v3.pdf

[56] For example, R. Brooks, 'Worthless Warhol Alarms Art World', *Timesonline*, 22 January 2006, http://www.timesonline.co.uk/tol/news/uk/article717330.ece

[57] In the case of Thomson vs Christie's, 70% certainty that an object was what Christie's claimed it to be was good enough for the presiding judge, http://news.bbc.co.uk/2/hi/uk_news/england/norfolk/3727623.stm; S.G. Vyas, 'Is There an Expert in the House? *Thomson v. Christie's*: The Case of the Houghton Urns,' *International Journal of Cultural Property*, vol. 12 (2005), 425-441.

The inability to separate the authentic from the inauthentic in the case of counterfeit drugs is creating a 'global public health problem causing death, disability and injury'[58] and the continuing growth in the production of such counterfeit products as handbags, trainers and watches raises concerns over the protection of intellectual property rights and economic returns. At the heart of establishing authenticity lies trust and this is an area where we are just beginning to understand the issues.[59]

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…'[60] Authenticity 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 from the perspective of the user:

- How does a user know that a digital object is an authentic instantiation of the version that was ingested (e.g. deposited) into the digital library? What tools will a user need to have at her/his disposal in this world of digital diplomatics if the user is to be able to make an independent judgement about authenticity?[61] What information, functions and services should the digital library provide to enable the user to be able to authenticate a digital object?
- Confronted with digital objects, those of us who were engaged in the InterPARES 1 Taskforce on Authenticity concluded that most users begin from a position of presuming that if an object is said to be authentic by the supplier then it is – 'Presumption of Authenticity'. Unless some evidence emerges that causes them to question the authenticity of an object, users generally assume that, because the object is held by an archives or a library, its authenticity is beyond question.
- 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.

---

[58] WHO, *Combating Counterfeit Drugs: A Concept Paper for Effective International Cooperation World Health Organization*, Health Technology and Pharmaceuticals, 'Combating Counterfeit Drugs: Building Effective International Collaboration', International Conference – Rome, Italy, 16-18 February 2006, (drafted by Michele Forzley), (Rome, 27 January 2006) http://www.who.int/medicines/events/FINALBACKPAPER.pdf, p. 1. See also: http://www.who.int/medicines/services/counterfeit/overview/en/ and FDA, *COMBATING COUNTERFEIT DRUGS: A Report of the Food and Drug Administration*, (Rockville, MD: FDA, 2004), http://www.fda.gov/oc/initiatives/counterfeit/report02_04.html

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

[60] T. Cooke, 'Archival Science and Postmodernism: New Formulations for old Concepts', *Archival Science*, vol. 1, no. 1 (2000), 3-24.

[61] The technologies to assist with digital forensics are emerging, as W. Wang and H. Farid, 'Exposing Digital Forgeries in Interlaced and Deinterlaced Video', *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3 (September 2007), 438- 449 shows.

Not wishing to confuse the issues at this stage, but it is worth recognising the distinction between authentic and reliable information.[62] Not all 'authentic' material held by a digital library need be 'reliable'. Once material comes to be held in a digital library or repository it must be immutable if we are to accept it as authentic. In fact, many digital libraries contain unreliable information, but even unreliable data can tell its own story if its provenance, context and purpose can be ascertained. Additionally, we might raise the issue of content quality in terms of digital libraries; quality is a property of digital objects that needs attention alongside authenticity and reliability.[63]

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. A growing number of researchers are moving into this discussion area and debate in this area is likely to become more and more lively.

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 action 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 emphasise 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 (2002-2006) of proceedings from ECDL and JCDL is that most digital library research tends to focus on the here and now. The addition of a digital preservation cluster to DELOS Network of Excellence was a visionary move by Costantino Thanos and Vittore Casarosa (Istituto di Szienza e Tecnologie dell'Informazione - ISTI, Consiglio Nazionale delle Ricerche CNR at Pisa); it reflected their recognition that digital libraries were not just about communicating with the present but that they are mechanisms to facilitate communication with the future.[64] Until recently, however, preservation has not been seen as central to digital library design and development. Those of us working in the team ably led by Donatella Castelli (also of ISTI at CNR Pisa) to develop the DELOS Digital Library Reference Model are only just coming to grips with how to incorporate

[62] See the work of InterPARES 1, http://www.interpares.org/ip1/ip1_index.cfm

[63] D.M. Strong, Y.W. Lee and R.Y. Wang, 'Data quality in context,' *Communications of the ACM*, vol. 40, issue 5 (May 1997), 103-110. DOI= http://doi.acm.org/10.1145/253769.253804; A. Martinez and J. Hammer, 'Making quality count in biological data sources', *IQIS '05: Proceedings of the 2nd international workshop on Information quality in information systems*, (Baltimore, MD: ACM Press, 2005), ISBN 1-59593-160-0, 16-27. DOI= http://doi.acm.org/10.1145/1077501.1077508. Of course, as A. Even and G. Shankaranarayanan have demonstrated, the same data may be assessed by different users to have different degrees of data utility depending upon context of use (Utility-driven assessment of data quality, *SIGMIS Database*, vol. 38, no. 2 (May 2007), 75-93. DOI= http://doi.acm.org/10.1145/1240616.1240623).

[64] DELOS: Network of Excellence on Digital Libraries (G038-507618) funded under the European Commission's 6th Framework IST Programme, http://www.delos.info and http://www.dpc.delos.info

preservation functionality and capabilities into what is emerging as an outstandingly robust framework for digital libraries.[65]

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 and interrelatedness of the digital entities they will contain. The current generation of solutions, many of which centre on migration and emulation, are unrealistic and focus too heavily on narrow aspects of the problem – they are the kinds of solutions that we have described above as artisan. The ingest of heterogeneous materials into a digital library (e.g. the digital materials created by contemporary writers 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 (the objects themselves and their context), represent, and manage digital entities radically change.

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 community has yet emerged. Since 1989 at least twelve have been published.[66] One of the tasks of DigitalPreservationEurope (DPE) has been to look at the

---

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

[66] NHPRC, *Research Issues in Electronic Records: Report of a Working Meeting*. (St Paul, MI: Minnesota Historical Society for the United States National Historical Publications and Records Commission, 1991); M. Hedstrom, 'Understanding Electronic Incunabula: A Framework for Research on Electronic Records,' *The American Archivist*, vol. 54 S. (1991), 334-355; J. Garrett and D. Waters (co-chairs), *Preserving Digital Information: Final Report and Recommendations*, (Commission on Preservation and Access and the Research Libraries Group, 1996), ftp://ftp.rlg.org/pub/archtf/final-report.pdf; Ann Arbor Report, *Electronic Records Research and Development: Final Report of the 1996 Conference held at the University of Michigan, Ann Arbor, 28-29 June 1996,* (Ann Arbor, MI: School of Information, Bentley Historical Library, and National Historical Publications Records Commission, 1997); D. Lievesley and S. Jones, *An Investigation into the Digital Preservation Needs of Universities and Research Funders*, (London: BLRIC Report no. 109, 1998), http://www.ukoln.ac.uk/services/papers/bl/blri109/datrep.html; NSF and LC, *It's About Time: Research Challenges in Digital Archiving and Long-term Preservation*, 12-13 April 2002. (sponsored by the National Science Foundation (NSF) and the Library of Congress (LC), 2002), http://www.si.umich.edu/digarch/NSF%200915031.pdf; CLIR Report, *The State of Digital Preservation: An International Perspective*, (Washington, DC: Council on Library and Information Resources, 2002), http://www.clir.org/pubs/reports/pub107/pub107.pdf; M. Hedstrom and S. Ross (eds.), *Invest to Save: Report and Recommendations of the NSF-DELOS Working Group on Digital Archiving and Preservation*, (National Science Foundation's (NSF) Digital Library Initiative & The European Union under the Fifth Framework Programme by the Network of Excellence for Digital Libraries (DELOS), 2003), http://delos-noe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/digitalarchiving/Digitalarchiving.pdf; P. Lord and A. Mcdonald, *e-Science Curation Report*, (JCSR Report, 2003), http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf; Cyberinfrastructure, *Revolutionizing Science and Engineering Through Cyberinfrastructure*, (Washington DC: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, January 2002), http://www.communitytechnology.org/nsf_ci_report/report.pdf; DigiCULT, *The Future Digital Heritage Space: An Expedition Report*, DigiCULT Thematic Issue 7, 2004, http://www.digicult.info/downloads/dc_thematic_issue7.pdf; D. Giaretta and H. Weaver, (eds.), *Report of the Warwick Workshop, 7 & 8 November 2005: Digital Curation and Preservation: Defining the research agenda for the next decade* (2005), http://www.dcc.ac.uk/events/warwick_2005/Warwick_Workshop_report.pdf; R. Heery and A.

digital preservation landscape and to come up with a research agenda that might be taken forward under the Seventh Framework Programme of the European Commission, as well as at national levels within the Member States of the European Union. Based on an extensive crosswalk of existing preservation research agendas, the DPE Research Roadmap's objective is to provide a concise overview of the core issues that have to be addressed in future digital preservation research.[67] To construct the framework, my colleague Holger Brocks (of the FernUniversität in Hagen) led participants in the DPE Research Roadmap Working Group (RAWG) to examine the challenges of preservation from five vantage points: digital object level, collection level, repository level, process level and organisational environment which also encapsulates creation and use. So, for instance, at the object level we focus on migration, emulation, experimentation and acceptable loss; at the collection level we examine interoperability, metadata and standardisation; and at the process level we look at issues such as automation and workflow.

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. Secondly, it recognises that, as those working in the discipline came to better understand the preservation obstacles, they extended the research domain into areas that were originally peripheral to digital preservation. This has actually hampered progress because it has fragmented research activity much too broadly. In response, DPE has proposed narrowing the research agenda and argued that as a research community we must capitalise on ancillary work carried out in other domains such as semantic-enabled information infrastructures, grid-based resources and service-oriented architectures. The DPE team have agreed that there are really nine themes that should characterise our research in preservation. These nine themes also bring digital preservation in line with traditional preservation activities in the analogue world. In addition, there is one core methodological approach that researchers in preservation need to adopt.

The nine themes are:

1. Restoration. Digital objects break. This can occur when storage media become damaged, software and hardware become obsolete, applications become inaccessible either through loss of access or through technological developments, or bit streams become corrupt. When they break and they are unique and valuable, 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,[68] but we need more experimental research in this area to develop effective and user-friendly restoration technologies. How

---

Powell, *A Digital Repositories Roadmap: Looking Forward* (2006), http://www.eduserv.org.uk/upload/foundation/pdf/rep-roadmap-v15.pdf; N. Beagrie, *e-Infrastructure Strategy for Research: Final Report from the OSI Preservation and Curation Working Group*, (Edinburgh: National e-Science Centre, November 2006, but published in 2007), http://www.nesc.ac.uk/documents/OSI/preservation.pdf

[67]DigitalPreservationEurope, *DPE Digital Preservation Research Roadmap* (2007), http://www.digitalpreservationeurope.eu/publications/dpe_research_roadmap_D72.pdf

[68] Companies such as OnTrack Data Recovery (http://ontrackdatarecovery.com) or DriveSavers (www.drivesavers.com) have developed a rich array of data recovery technologies. The methods and processes are getting better, as Scott Gaidano, co-founder of DriverSavers, points out: 'eight years ago [1997], 50 percent of our drives could not be restored. Now up to 90 percent of the data can be salvaged from 85 to 90 percent of drives,' E.A. Taub, 'Bad habits keep data recovery firms alive,' *International Herald Tribune*, 16-17 July 2005, 14.

do we verify the completeness of a restored digital object? What is an acceptable level of loss at different syntactical and semantic levels? How do we restore content, context and experience?

2. Conservation. Whereas restoration offers ways to handle objects that have become severely damaged or exist only in fragmentary form, 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 preventive 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 simply have not done the research.

3. Collection and repository management. 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. For example, how do we construct collections in the digital age? What kinds of service layers do users of digital libraries require and how will these be maintained over time?

4. Preservation as risk management. We have argued elsewhere that digital preservation is a risk management problem.[69] 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.

5. Preserving the interpretability and functionality of digital objects. Our understanding of the properties that digital objects must retain over time if the objects are to remain semantically meaningful, authentic, reliable and usable, whether for rendering or analysis, remains limited. How do we validate verisimilitude of content, context and performance? What metrics do we have for measuring consistency of functionality and behaviour of digital objects over different digital library technical systems and environments?

6. Collection cohesion and interoperability. Digital libraries and repositories handle collections of digital objects as opposed to just discrete entities. It is the integrated nature of these collections that provides some degree of contextuality to the individual objects. Moreover, collections often only gain real value when they can be integrated with collections held by other repositories. The research that has been done into interoperability across generations of systems, time and repositories has been insufficient.

7. Automation in preservation. The sheer quantity of 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.[70] Areas where

[69] Ross and McHugh, (2006a), 'The Role of Evidence in Establishing Trust…..' (see above note 6), and Ross (2006), 'Uncertainty, Risk, Trust and Digital Persistency' (see above note 20).

[70] J.F. Gantz *et al.*, *The Expanding Universe: A Forecast of Worldwide Information Growth Through 2010*, (An IDC White Paper, sponsored by EMC, 2007),

automation has promise include: metadata extraction,[71] preservation planning and action,[72] 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 simply have not done the underlying research, experimentation and testing.

8. Preserving the context. Establishing 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 kinds of processes are necessary to construct context and meaning? Research into contextuality is needed.

9. Storage technologies and methods. On the one hand this is an engineering problem and on the other it 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.[73] The reason is that metadata issues 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.[74] Every aspect of preservation research from characterisation of digital objects to preservation planning to user needs analysis requires experimental research. Some of the post-2003 research and support activities related to digital preservation in Europe, such as the Digital Curation Centre (DCC) in the UK,[75] DigitalPreservationEurope (DPE), CASPAR (Cultural, Artistic and Scientific

---

http://www.emc.com/about/destination/digital_universe/pdf/Expanding_Digital_Universe_IDC_WhitePaper_022507.pdf The current growth rate continues to exceed predictions. For example, contrast the data in Gantz *et al*. with that in P. Lyman and H.R. Varian, *How Much Information*?, (Berkeley, CA: University of California at Berkeley, School of Information Management and Systems, 2000), online: http://www.sims.berkeley.edu/research/projects/how-much-info/internet.html

[71] For example, Y. Kim and S. Ross, 'The Naming of Cats: Automated Genre Classification', *The International Journal of Digital Curation*, vol. 2, no. 1 (2007), http://www.ijdc.net/./ijdc/article/view/24/27, ISSN: 1746-8256

[72] For example, S. Strodl, A. Rauber, C. Rauch, H. Hofman, F. Debole and G. Amato, 'The DELOS Testbed for Choosing a Digital Preservation Strategy', in *Proceedings of the 9th International Conference on Asian Digital Libraries (ICADL'06)* (Kyoto, Japan, 27-30 November 2006), (Berlin: Springer, 2006), 323-332; S. Strodl, C. Becker, R. Neumayer and A. Rauber, (2007), 'How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure', in *Proceedings of the 2007 Conference on Digital Libraries,* (Vancouver, 2007), 29-38, http://doi.acm.org/10.1145/1255175.1255181

[73] W. Duff, 'Metadata in Digital Preservation: Foundations, Functions, and Issues,' in F.M. Bischoff, H. Hofman and S. Ross (eds.), *Metadata in Preservation*, (Marburg: Veröffentlichungen der Archivschule Marburg, Institut für Archivwissenschaft, no. 40, 2004), ISBN 3-923833-77-6, 27-38.

[74] This is not to suggest that there has been no experimentation to date, but to point out that it has been limited. Examples include M.L. Nelson, J. Bollen, G. Manepalli and R. Haq, 'Archive Ingest and Handling Test: The Old Dominion University Approach,' *D-Lib Magazine*, vol. 11, no. 12 (2005); W.Y. Arms, R. Adkins, C. Ammen and A. Hayes, 'Collecting and Preserving the Web: The Minerva Prototype,' *RLG DigiNews*, vol. 5, no. 2 (2001). A good summary of the publications related to both practitioner and researcher led studies in preservation is provided in the quarterly DPC/PADI 'What's New in Digital Preservation', http://www.nla.gov.au/padi/quarterly.html

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

knowledge for Preservation, Access and Retrieval),[76] PLANETS (Preservation and Long-term Access through NETworked Services),[77] the Digital Preservation Cluster of the DELOS Network of Excellence in Digital Libraries (DELOS-DPC)[78] and numerous other projects I might have mentioned, 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 are the take-away points with which I want to leave you?

First, as a community we need to re-think how we are approaching research into digital preservation and curation.

Second, we need to engage digital libraries researchers in this process, and especially those with a strong computing science and engineering background.

Third, research in digital preservation must 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 be more 'experimental' than it has been up to now, something that, as I have noted, a number 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.

Fourth, not only do we need to try to better understand what we might do to alleviate obstacles to the longevity of digital materials, we must do more to define the uncertainties related to digital preservation and to convert these uncertainties into known, measurable and mitigatable 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 – we just do not know in a measurable way in the context of digital objects which risks are actual.

Digital libraries must adopt a theoretical stance; recent discussions about curricula for undergraduate and postgraduate education in digital libraries make this lack of a theoretical knowledge base really evident. Indeed, the team led by the School of Information and Library Science at University of North Carolina (Chapel Hill), and Department of Computer Science at Virginia Tech conducting the US National Science Foundation project to develop a curriculum for education in digital libraries have reported that, 'research and development in the DL area will flourish only if it has a firm theoretical foundation'.[79] As I noted above, library science has not demonstrated that it has the

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

[77] http://www.planets-project.eu

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

[79] Work to develop curriculum to facilitate the education of digital librarians is described by J. Pomerantz, B.M. Wildemuth, S. Yang and E.A. Fox, 'Curriculum development for digital libraries', in *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries* (Chapel Hill, NC, USA, 11-15 June 2006), JCDL '06, (New York, NY: ACM Press, 2006), 175-184. DOI = http://doi.acm.org/10.1145/1141753.1141787. See also M. Moss and S. Ross, 'Educating information management professionals – the Glasgow perspective', *Journal of Education for Library and Information*

theoretical foundations and knowledge base that are capable of providing the framework for handling digital entities and for underpinning digital libraries. Moreover, as digital libraries are more akin to archives than they are to traditional libraries we need to seek their theoretical foundations in the domain of archival science and their practices in archival and records management environments. Archival science, with its principles of uniqueness, provenance, arrangement and description, authenticity, appraisal, and its tool sets such as diplomatics and palaeography,[80] may offer us a framework for a theoretical foundation for digital libraries.

Perhaps you are surprised that I have not come here today and told you that those of us working in digital preservation have solved all, or at least most, of the challenges, and we are just waiting for those of you working in digital libraries to ask us to integrate our solutions into your work. Since we have not overcome the obstacles to preservation of digital materials, I can not hold out such a promise. 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 they 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.[81] Thank you.

Prof. Seamus Ross, Budapest, 17 September 2007 (revised 30 September 2007)

Science, (forthcoming).

[80] I might have examined the issues surrounding digital palaeography here as well. In the same way that using knowledge about different scripts (say Insular round compared with Caroline minuscule) a palaeographer can make inferences about the origin and production of documents, a digital palaeographer will be able to use information about the characterisation and nature of digital objects to draw conclusions about the process of production, use and authenticity. The boundaries of diplomatics and digital palaeography still need to be defined for the digital age, much as they did in the 17th century.

[81] As I noted in my closing keynote talk, 'Connecting and Communicating our Digital Heritage with the Future', at the conference organised under the auspices of the Austrian Presidency of the European Union (Residenz zu Salzburg, 21-22 June 2006) on *An Expedition to European Digital Cultural Heritage: Collecting, Connecting – and Conserving?*, prizes for competitions have acted as a powerful incentive to research and development in other contexts, http://dhc2006.salzburgresearch.at/images/stories/info/mp3/6_ross.mp3 For instance, the Orteig Prize for a non-stop trans-Atlantic flight (e.g. New York to Paris) and Ansari X-Prize for private sector space flight using a reusable craft stimulated investment by individuals and organisations that far exceeded the value of the prizes. In 2005, when designing DigitalPreservationEurope (DPE), I proposed that we should include a Digital Preservation Challenge, http://www.digitalpreservationeurope.eu/challenge/. The first challenge was launched in the spring of 2007 and the awards were announced at ECDL2007 in Budapest.