The International Journal of Digital Curation
Issue 1, Volume 1 | Autumn 2006

Digital Curation for Science, Digital Libraries, and Individuals

Neil Beagrie,

JISC/British Library Partnership Manager

Abstract

The creation, management and use of digital materials are of increasing importance for a wide range of activities. Much of the knowledge base and intellectual assets of institutions and individuals are now in digital form. The term digital curation is increasingly being used for the actions needed to add value to and maintain these digital assets over time for current and future generations of users. The paper explores this emerging field of digital curation as an area of inter-disciplinary research and practice, and the trends which are influencing its development. It analyses the genesis of the term and how traditional roles relating to digital assets are in transition. Finally it explores some of the drivers for curation ranging from trends such as exponential growth in digital information, to "life-caching", digital preservation, the Grid and new opportunities for publishing, sharing, and re-using data. It concludes that significant effort needs to be put into developing a persistent information infrastructure for digital materials and into developing the digital curation skills of researchers and information professionals. Without this, current investment in digitisation and digital content will only secure short-term rather than lasting benefits.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Definition and History of Digital Curation

The term digital curation is increasingly being used for the actions needed to maintain digital research data and other digital materials over their entire life-cycle and over time for current and future generations of users (Giaretta, 2005; Joint Information Systems Committee, 2003).

The term digital curation is very new and together with related terms such as digital preservation and digital archiving it is still evolving. It is important to recognise that these terms still can be perceived differently by different individuals and disciplines.

Inter-disciplinary dialogue between librarians and scientists has been seen by commentators as important to the long-term management, preservation, and use of scientific research (Hey & Trefethen, 2003; Messerschmidt, 2003). In the past, a major barrier to any such dialogue has been the differing interpretation and usage of terminology by the professions and disciplines involved. The use of terms such as “archiving”, “preservation”, and “data” can mean different things to different groups and there is often a deeply embedded local usage, which professions are reluctant to change.

These difficulties have led to the recent adoption amongst some specialists of the term “digital curation”. This is a relatively new term incorporating aspects of the existing concepts “data curation” and “digital preservation” used primarily by the scientific and digital library communities respectively. However its use was also intended to build bridges between them and reflect new approaches.

Although the exact terms used have varied (from “curation” to “digital preservation”, to “digital curation”) a range of commentators have been trying to convey the concept that we now need a new approach to creating and managing digital assets. This approach often confounds attempts to neatly categorise activities and demands involvement from and interaction between, a far wider group of individuals, roles and organisations. This involvement and interaction extends across authors and researchers, publishers and curators, and information and data management specialists (Beagrie & Jones, 2001; Gray, Szalay, Thakar, Stoughton, & vandenBerg, 2002).

The term “digital curation” was first used at the "Digital Curation: digital archives, libraries and e-science seminar" sponsored by the Digital Preservation Coalition and the British National Space Centre held in London on the 19th October 2001. This invitational seminar brought together international speakers from many different sectors to discuss leading edge developments in the field of data curation and digital preservation. The seminar was felt by participants to have established an essential cross-sectoral dialogue between archivists, library and information management specialists, and data managers in e-science (Beagrie & Pothen, 2001).

A contribution to this successful dialogue was the careful selection of the term “digital curation” used for the seminar. The new term benefited from some existing usage of the term “curation” by both the library and museum sectors, and the biological sciences. In all three sectors the term implies not only the preservation and maintenance of a collection or database but some degree of added value and knowledge.

In the library and museum sector curation centres on well-established concepts of added value from themed collection-building around physical objects (the sum being

greater than the parts); from the documentation accompanying individual objects and collections which provides the relevant context and history for research, learning, and discovery; and from the skills, domain expertise, and knowledge of the staff, the curators of the collections. The existing use of the term curation in museums and libraries largely applied to physical artefacts. "Digital curation" was used at the seminar as a term to explicitly transfer existing curatorial approaches to digital collections, and also to highlight some of the changes that are needed in approaches to curation of digital as opposed to analogue artefacts (for examples of both transferable practice and changes, see Beagrie and Jones, 2001).

The concept of collection building as part of curation, in other words selecting and maintaining a body of knowledge and evidence for specific disciplines or topics, can be seen elsewhere in many other disciplines and sectors e.g. in data centres for the social sciences or oceanographic and other environmental sciences.

In the biological sciences, the term curation had been applied to the maintenance and publishing of databases such as the human genome and was therefore already implicitly digital. In this context added value is derived from annotation, linkage, and the management, validation, and editorial input of domain specialists employed to curate and publish the database.

Prior to the seminar, the term "curation" had already been adopted by John Taylor, then Director-General of the Research Councils, when referring to the information infrastructure needed for the proposed e-science programme, and in particular the acquisition and curation of very large valuable collections of primary data (Taylor, 2001). This provided some valuable political context and support for the adoption of the term.

Another consideration in adopting the term "digital curation" was the perception amongst many data creators and researchers of the terms archiving and preservation as an end-of-project activity which did not involve them – their role being confined to the research and creation and publishing of data and outcomes (Feeney, 1999, p14; Lievesley & Jones, 1998). Equally influential was a perception of preservation of digital materials as a separate activity with little connection to creation or promoting the re-use of these materials (Atkins et al., 2003, p43). Effectively these perceptions can lead to ad hoc and fragmented preservation in what has been termed "data mortuaries" rather than data archives.

The JISC and other bodies had long recognised the importance of a life-cycle approach to the maintenance of digital research. This approach recognised that different (and often differently interested) stakeholders become involved with data resources at different stages but that building relationships between these different stakeholders was vital for their maintenance and research value (Beagrie & Greenstein, 1998, p3; Beagrie, 2004). The concepts and ideas emerging behind the term digital curation offered the prospect of helping to build these relationships, promote pro-active stewardship, and avoid some of the negative perceptions associated with other terms amongst researchers.

Subsequently there have been a number of attempts to refine the definition of the term and related activities. For example, there was extensive discussion of the term and different sectoral perspectives on it by the Digital Data Curation Taskforce (Lord & Macdonald, 2003a).

The e-science curation report which followed on from the Taskforce later suggested:

> "This is a relatively new field, and terminologies are not yet

> stable. For this paper we use working definitions of three key activities: “curation”, “archiving” and “preservation” ….very broadly speaking, these are terms of increasing specificity in this context:
> preservation is an aspect of archiving, and archiving is an activity needed for curation. All three are concerned with managing change over time.” (Lord & Macdonald 2003b, p12)

The term digital curation was subsequently incorporated by the JISC and the e-science core programme into a call for proposals to establish a Digital Curation Centre (DCC) (JISC, 2003). At the 1st International Digital Curation Conference a session on "What is Digital Curation" debated the definition of data curation in terms of the remit of the DCC (Kerr, Reddington, & Wilkinson, 2005).The Digital Curation Centre’s own current definition of the term is:

> “Digital curation, broadly interpreted, is about **maintaining** and **adding value to**, a trusted body of digital information for **current** and **future use**.” (Giaretta, 2005).

This short introduction provides some history for, and documents the evolution and definition of, the term digital curation and the desire to foster new approaches and collaboration which underlay it. The remainder of this paper examines some of the key drivers behind emerging user requirements in this field. It is not intended to be a comprehensive discussion of individual drivers and of each topic area but to provide an overview and introduction to them.

Drivers and Requirements for Digital Curation

Information Growth

The escalating volumes of information generated and encountered by individuals and institutions is increasingly recognised. The School of Information Management and Systems at the University of California at Berkeley has estimated that in 2002 print, film, magnetic, and optical storage media produced about 5 exabytes of new information. World-wide this amounts to almost 800 MB of recorded information produced per person that year. Based on their studies conducted in 1999 and 2002 they estimated that new stored information grew about 30% a year between 1999 and 2002. Ninety-two percent of the new information was stored on magnetic media, mostly on hard disks (Lyman & Varian, 2003).

Worldwide growth in published information for both serials and monographs, and a growing shift from paper to electronic publication, are already widely recognized trends amongst research libraries. In a study undertaken for the Joint Voluntary Committee for Electronic Deposit Department in advance of UK legislation for legal deposit of electronic publications, these trends are predicted to continue and accelerate (Electronic Publishing Services, 2002). More recently the British Library in launching its new three-year strategy has estimated by the year 2020, 40% of UK research monographs will be available in electronic format only, while a further 50% will be produced in both print and digital. A mere 10% of new titles is expected to be available in print alone by 2020. It points out that this will be a seismic shift for the Library, its partners in publishing, and the information sector (British Library, 2005).

For scientific data, Hey and Trefethen argue that experiments and instruments currently being built will dramatically escalate the current rates and volumes of scientific data creation. They point out that e-Science data generated from sensors,

satellites, high-performance computer simulations, high-throughput devices, scientific images and so on will soon dwarf all of the scientific data collected in the whole history of scientific exploration (Hey & Trefethen, 2003). Examples include 'virtual observatories' containing astronomical data being funded in the USA, in Europe, and the UK. In the USA it is estimated that the planned Large Synoptic Survey Telescope alone will produce over 10 petabytes of data per year by 2008 (National Virtual Observatory, 2005). Another example taken from particle physics is the Large Hadron Collider (LHC) at CERN in Geneva. This will generate roughly 15 petabytes of data annually from 2007, which thousands of scientists from around the world will access and analyse (CERN, 2005).

There are also rapidly growing databases in bioinformatics including examples such as the Protein Data Bank, a database of protein structures. The Protein Data Bank has been in existence since 1972. Figure 1 shows the growth in deposited structures between 1972 and July 2005. Growth in its first two decades was relatively slow but the last decade has seen exponential increases. Arguably this reflects not just the increases in data being generated but the increasing recognition by the discipline of the importance of data, requirements to deposit data with publication of journal articles, and of "collection-based science" – a phenomenon explored further below.

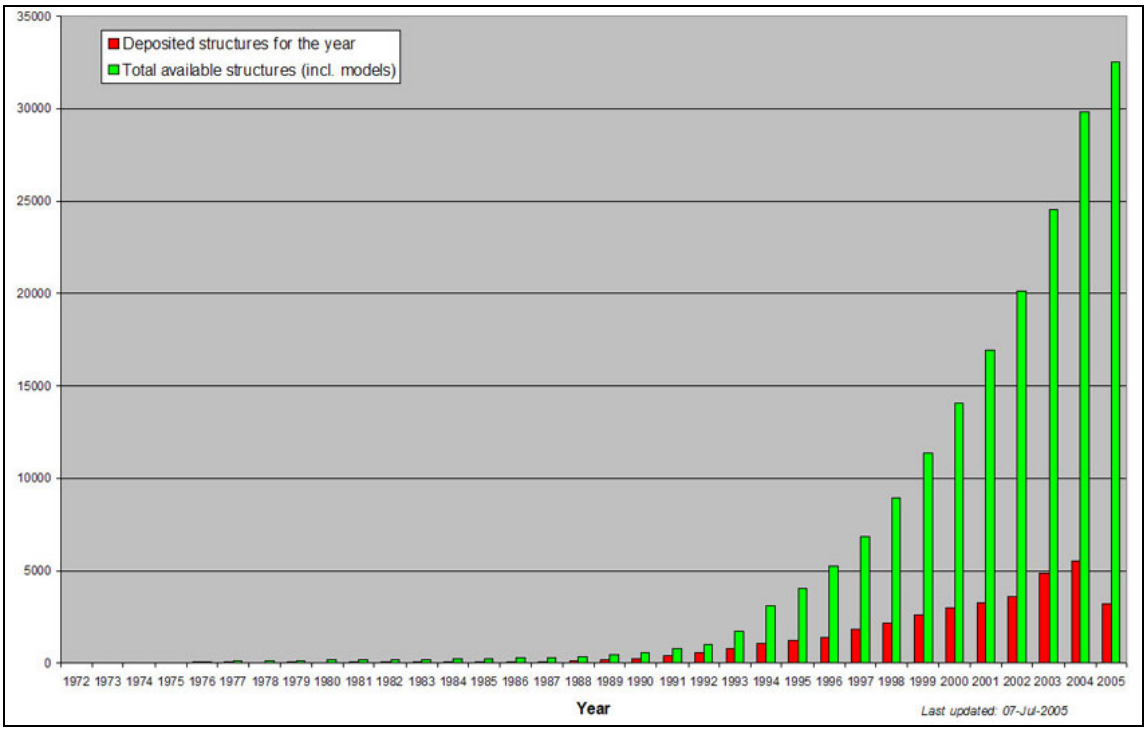

Figure 1 Protein Data Bank Content Growth 1972 – 7th July 2005 (Protein Data Bank 2005), © 2005 Research Collaboratory for Structural Bioinformatics.

As noted above, through sensors, experiments, digitisation and computer simulation, digital resources and data are growing in volume and complexity at a staggering rate. The cost of producing these resources is very high: satellites, particle accelerators, genome sequencing, and large-scale digitisation and electronic publishing collectively represent a cumulative investment of billions in digital research and learning.

What are the long-term implications of this information growth for repositories of data or publications? First it is clear that funding for repositories is unlikely to match the exponential growth in data and publications currently underway. A substantial part of the cost-base of repositories consists of skilled staff and these human resources and many existing workflows and practices will not scale appropriately. There will be a need for more automation of processes and metadata generation, software tools for this, and potentially the development of greater collaboration and shared services to lower the entry and operational costs for institutions.

Although not all of the digital information being generated will have long-term value, often a significant component of it will. Long-term value and the volumes of information will vary between disciplines and different categories of material so selection for long-term curation and preservation may be a significant issue. As a consequence, selection, curation and long-term preservation of digital resources could be of increasing importance for a wide range of activities (Digital Preservation Coalition, 2006; Lord & Macdonald, 2003b).

e-Research and Collection-based Science

Alongside the growth in data, commentators have highlighted that the use and value of data is also changing. In the US the National Science Board has stated that:

> "It is exceedingly rare that fundamentally new approaches to research and education arise. Information technology has ushered in such a fundamental change. Digital data collections are at the heart of this change. They enable analysis at unprecedented levels of accuracy and sophistication and provide novel insights through innovative information integration. Through their very size and complexity, such digital collections provide new phenomena for study." (National Science Board, 2005).

Similar views have been expressed internationally through the International Council for Science:

> "Because of the critical importance of data and information in the global scientific enterprise, the international research community must address a series of new challenges if it is to take full advantage of the data and information resources available for research today. Equally, if not more important than its own data and information needs, today's research community must also assume responsibility for building a robust data and information infrastructure for the future." (International Council for Science, 2004).

This is having a profound effect on how science is being conducted in some disciplines. For example in astronomy the archiving and sharing of data is dramatically changing the pattern of publishing and the conduct of research. In the past individual astronomers would have booked time on telescopes and published research based on their observations. Now the development of "virtual observatories" allows research to be conducted on a digital collection of archived observational data (Gray et al., 2002). It is now possible for a larger number of researchers to access and utilize data from expensive instruments. Figure 2 below shows the growth of new data observations (ingest) for the Hubble Space Telescope (HST) and Far Ultraviolet Spectroscopic Explorer (FUSE) Data Archive against use by researchers (retrievals). It demonstrates how use of the archive has overtaken and continues to outgrow the growth of new data

observations and provides a good illustration of the growing importance of collection-based science in Astronomy.

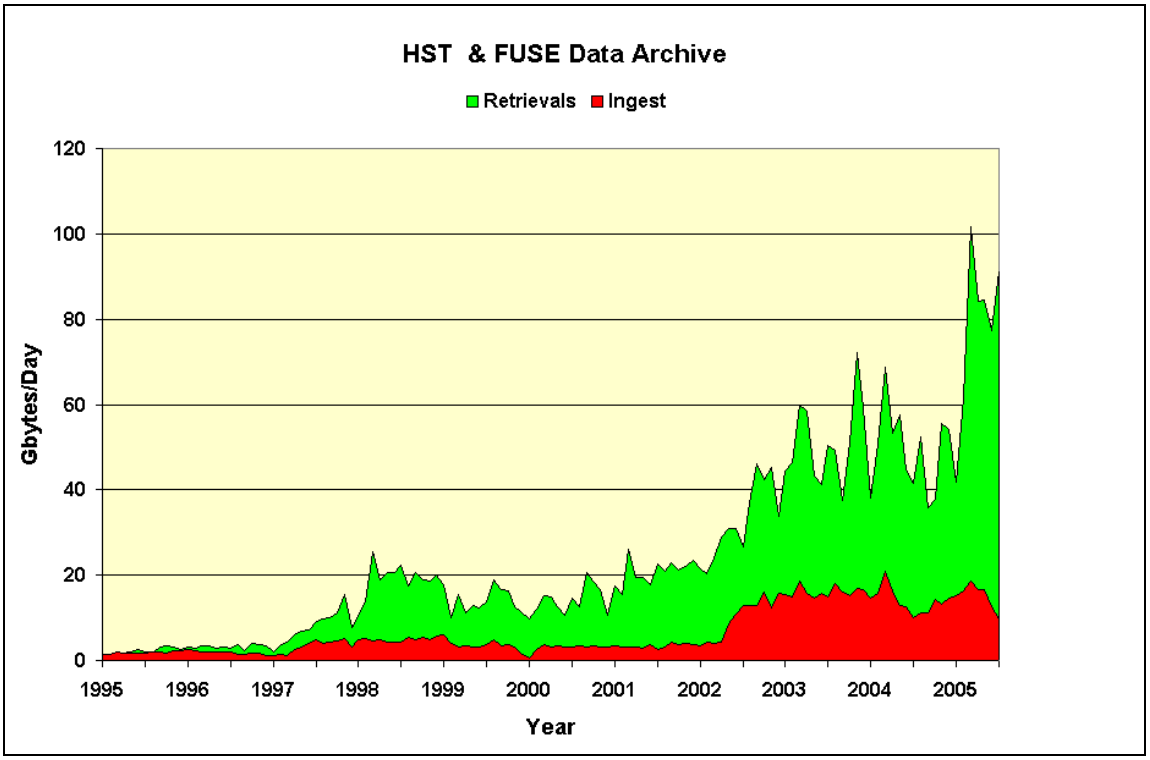

Figure 2. Hubble Space Telescope (HST) and Far Ultraviolet Spectroscopic Explorer (FUSE) Data Archive: Retrievals and Ingest 1995 – June 2005. © 2005 Space Telescope Science Institute

The Role of Data Publishing and the Longevity of "Active" Data

"Unwillingly, and sometimes unknowingly, projects become not only Authors, but also Publishers and Curators. The Consumers interact with the projects directly. Scientists are familiar with how to be an Author, but they are just starting to learn, out of necessity, how to become a Publisher and Curator. This involves building large on-line databases and designing user interfaces. These new roles are turning out to be demanding and require new skills." (Gray et al., 2002)

As noted above (Gray et al., 2002), in some subjects databases are supplementing or partly replacing journal publications as a medium of scholarly communication. These databases are built and maintained with a great deal of human effort but the scale of effort and supporting infrastructure varies: some may have discipline-wide scope or dedicated "curators" such as the Protein Data Bank or Hubble Space Telescope Data Archive but many others may be relatively small and project-based.

New forms of data publishing pose many challenges both technical and organisational. It is worth stressing that these changes and challenges in data publishing are not solely confined to research data. Similar trends can be seen in traditional publishing and in the Web as electronic publication increasingly involves dynamic, on the fly generation rather than static fixed versions of content. An excellent example of this is the large-scale publications of the Ordnance Survey, the national mapping agency for the UK. The Ordnance Survey has been publishing large-scale maps at different scales in paper editions since 1791.

Computerization of survey information at the Ordnance Survey was first designed to assist in the workflow of paper publication. However computerization is now complete and new survey information is now added continuously as it is captured to the Ordnance Survey National Topographic Database (NTD). For large-scale mapping traditional paper editions have now been discontinued. The NTD is the map: continuously updated and printed remotely on demand as users require (Fleet, 1999).

It is clearly likely that many of the largest and most significant emerging data publishing developments (e.g. the National Topographic Database) will have a very long active phase over many decades, perhaps centuries, and across many different generations of hardware and database software. These changes are very challenging and are blurring the boundaries between traditional professional skills and the roles of different organisations over the life-cycle of information. It involves greater shared responsibilities and liaison between those who would previously have seen their roles as solely involving creation, publishing or long-term preservation.

Digital Preservation

> ”Digital preservation refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary.” (Beagrie & Jones, 2001, p10)

As the volume and complexity of digital information grows, there has been growing realisation of the complexity of the activity needed to ensure long-term access to digital materials and the extent to which this differs radically from preservation activities in the paper environment.

In the right conditions papyrus or paper can survive by accident or through benign neglect for centuries or, in the case of the Dead Sea Scrolls, for thousands of years. It takes hundreds of years for languages and handwriting to evolve to the point where only a few specialists can read them.

In contrast, digital information will not survive and remain accessible by accident: it requires ongoing active management from as early in the life-cycle as possible. The information and the ability to read it can be lost in a few years. Storage media such as punched paper tape, floppy disks, CD-ROM, DVD evolve and fall out of use. Digital storage media have relatively short archival life-spans compared to other media. As the volumes, heterogeneity, and complexity of digital information grows, this requirement for active management becomes more challenging and more critical to a wider range of organisations.

This threat is not solely technological: it can also involve social factors and organisational risks particularly over extended periods of time. A real-life example of these different factors at work is given in a recent response by a UK-based Professor:

> “….I have data files from projects from years ago which are on disks I no longer have a drive for on computers I no longer have access to or are no longer made or the software/operating system changes would make it extremely difficult to access any more…. the nature of research work means a lot of short-term researchers over the years …Also as PIs move around and collaborate with many people in other organisations it is pretty difficult to go back more than a few years with confidence that data will be adequately archived.” (Lord & Macdonald, 2003b, p 17)

Digital preservation risks and losses take many forms. The threat of total loss of a broad swathe of the scientific record and cultural heritage in digital form has been

referred to as the risk of a "digital dark ages" (Kuny, 1998). Statistics on current losses are difficult to compile although there are a number of well-known individual examples of loss or near loss such as the BBC Domesday Disks. Wider overviews are rare. In part this is because few organisations wish to publicise losses. Also sometimes the information can be recovered or substituted in some way, e.g. a paper copy. In such cases the loss is often more subtle: information has effectively been degraded through loss of functionality, linking, or documentation, substantially reducing its real value.

We do have current statistics in some areas for this form of information loss. It is well established that Web documents and many links on the Web are ephemeral in nature. Some authors describe this in terms of a Web document half-life, others use terms like 'linkrot' or persistence (Koehler, 2004). Markwell and Brooks have studied citations of Web documents in the online literature employed by the scientific community and more specifically in biochemistry and molecular biology for education purposes. They found significant erosion in URL viability and estimated URL half-lives for these specific science education resources of some 4.6 years (Markwell & Brooks, 2002, 2003). In another study of the online citations in issues of 5 leading communications journals from 2000 to 2003, Bugeja and Dimitrova (2005) found that 33 percent of the links failed to work in the summer of 2004. They estimated the half-life of the citations they have studied to be under four years.

Such studies underline the fact that solutions to digital preservation challenges and development of a persistent information infrastructure must also involve persistent identifiers and resolver services for them.

Retention and Compliance

As noted above, the digital challenges affecting memory organisations e.g. libraries and archives, that have to think in centuries actually begin to manifest themselves in a decade or less, hence similar issues are beginning to impact on companies. Increasing regulation, compliance, and accountability across all sectors but particularly in banking, pharmaceuticals, medicine, and aerospace, mean companies must often retain digital information and keep it accessible for a decade or more. Also for some companies in broadcast and media their business assets are now largely or solely digital.

More broadly, legislation on data protection and freedom of information in the UK are focusing attention on records management generally and on electronic records (Bailey, 2005).

Changing business practices are also influencing retention and management of information. For example, the manufacturing and construction industries are switching to a system where instead of selling expensive products such as buildings and military hardware outright, they effectively lease them for 30 years, taking on the overall responsibility for maintaining, repairing and upgrading them. This new approach means that firms must understand how to deal with important digital information safely and securely for many years (McMahon, 2005).

Personal Digital Collections

> "Within 5-10 years, personal stores of a terabyte will cost a few hundred dollars, allowing persons to be immortal in terms of the media they've encountered. For "famous" people, one will be able to access his or her entire life." (Bell & Gray, 2001)

Digital challenges increasingly affect not just institutions but individuals. People are capturing and storing an ever-increasing amount of digital information about or for themselves, including emails, documents, portfolios of work, digital images, and audio and video recordings (Beagrie, 2005).

This abundance of personal data and collections presents numerous challenges to individuals, including: how physically to secure such material sometimes over decades; how to protect privacy; how to organise and extract useful knowledge from this rich library of information and to use it effectively; and for material intended to be shared, how effectively to present and control access by different groups of users.

Emerging demand has seen the growth of a number of commercial services to help individuals (and often their employers) to begin to address the challenges of managing personal data on PCs. Several companies now offer online backup of digital data to a remote secure repository using synchronisation and encryption software as a safeguard against data loss. Others are offering web-hosting of selected personal data such as address books and contact details, which can then be centrally maintained and accessed from different devices.

A desire to share digital images and documents has also led to rapid growth in software for individuals to publish blogs or publish digital images captured via mobile cameras and phones. Sharing of such information may be between immediate family and friends, interest groups or open to all individuals on the Web. Services such as Nokia's Lifeblog or Flickr for sharing, categorising, and searching digital images are seeing sharp increases in their user base and provide a number of tools for individuals which are proving highly popular. These behaviours capturing and sharing various life memories have been described as "life caching".

More recently, the Internet Archive and other partners have established Ourmedia. Individuals creating video, music, photos, audio clips and other personal media can store their content for free in perpetuity on Ourmedia's servers, as long as they are willing to share their works with a global audience. Ourmedia's goal is "to expose, advance and preserve digital creativity at the grassroots level." This is the first such service to offer explicitly long-term preservation as well as hosting services for personal and community content (Ourmedia, 2005).

Although such services are in their infancy there is also a growing interest from Computing Science in these areas. In the UK "Memories for Life" was recently recognized as a Grand Challenge for Computing Science by the UK Computing Research Committee and by the UK Foresight Cognitive Systems Group and a research network has been funded (Memories for Life, 2005). Opportunities for interdisciplinary research into memory are being created, between the life sciences, social sciences and physical sciences. The Memories for Life research network provides a broad inter-disciplinary arena where the scientific, social and technological aspects of personal and collective memory including digital memory and its preservation and transmission can be explored and reviewed (O'Hara et al., 2006).

Awareness of the curation issues that may surround personal digital collections and information is by no means widespread, but it is an area which seems very likely to grow and have increasing impact in years to come.

Conclusions

In conclusion, digital curation has implications for many different sectors as they move from paper to digital environments. For society and individuals, it can be argued that digital knowledge if it is to be useful and useable must be continuously updated,

maintained, and accessed. The emerging field of digital curation is central to this process.

Significant effort needs to be put into developing persistent information infrastructures for digital materials and into developing the digital curation skills of researchers and information professionals. Without this, current investment in digitisation and digital content will only secure short-term rather than lasting benefits.

Acknowledgements

I would like to thank Maggie Jones, Philip Lord, and Helen Hockx-yu for reading and commenting on an earlier version of this paper.

References

Atkins, Daniel E., et al. (2003). *Revolutionizing science and engineering through cyberinfrastructure:* Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Retrieved July 15, 2005 from Alliance for Community Technology Web site: http://www.communitytechnology.org/nsf_ci_report/

Bailey, S. (2005). Assessing the impact of the Freedom of Information Act on the FE and HE sectors. *Ariadne, 42*. Retrieved June 2, 2006 from: http://www.ariadne.ac.uk/issue42/bailey/

Beagrie, N., & Greenstein, D. (1998). *A strategic policy framework for creating and preserving digital collections*. British Library Research and Innovation Report 107. London: British Library. Retrieved from UKOLN Web site July 15, 2005 from:

http://www.ukoln.ac.uk/services/elib/papers/supporting/pdf/framework.pdf

Beagrie, N., & Pothen, P. (2001). The digital curation: Digital archives, libraries and e-science seminar. *Ariadne, 30*. Retrieved July 15, 2005 from:

http://www.ariadne.ac.uk/issue30/digital-curation/

Beagrie, N., & Jones, M. (2001). *Preservation management of digital materials: a handbook*. London: British Library.

Beagrie, N. (2004). The continuing access and digital preservation strategy for the UK Joint Information Systems Committee (JISC). *D-Lib Magazine, 10* (7/8). Retrieved July 15, 2005 from:

http://www.dlib.org/dlib/july04/beagrie/07beagrie.html

Beagrie, N. (2005). Plenty of room at the bottom? Personal digital libraries and collections. *D-Lib Magazine, 11*(6). Retrieved July 15, 2005 from:

http://www.dlib.org/dlib/june05/beagrie/06beagrie.html

Bell, G., & Gray, J. (2001). Digital immortality. *Communications of the ACM, 44*,(3), 29-31. Retrieved July 19, 2005 from:

http://research.microsoft.com/~gbell/CACM_Digital_Immortality.pdf

British Library. (2005). British Library predicts 'switch to digital by 2020'. Press release. 2005, June 29. Retrieved July 15, 2005 from:
http://www.bl.uk/news/2005/pressrelease20050629.html

Bugeja, M., & Dimitrova, D. (2005). *Half-life of Internet footnotes*. Retrieved July 15, 2005 from:
http://www.halfnotes.org/

CERN. (2005). *LHC computing grid*. Retrieved July 15, 2005 from:
http://lcg.web.cern.ch/LCG/

Digital Preservation Coalition (DPC). (2006). *Mind the gap: assessing digital preservation needs in the UK*. York: Digital Preservation Coalition. Retrieved June 2, 2006 from:
http://www.dpconline.org/docs/reports/uknamindthegap.pdf

Electronic Publishing Services. (2002). *The impact of the extension of legal deposit to non-print publications: Assessment of cost and other quantifiable impacts*. Study report. Retrieved July 15, 2005 from Association of Learned and Professional Society Publishers Web site:
http://www.alpsp.org/2004pdfs/LegalDepositofNon-PrintPublications.pdf

Feeney, M. (Ed.). (1999). *Digital culture: Maximizing the nation's investment – a synthesis of JISC/NPO studies on the preservation of electronic materials*. London: National Preservation Office.

Fleet, C. (1999). Ordnance Survey digital data in UK legal deposit libraries. *LIBER Quarterly, the journal of European research libraries, 9* (2). Retrieved July 15, 2005 from:
http://liber-maps.kb.nl/articles/fleet11.htm

Giaretta, D. (2005). *DCC approach to digital curation, version 1.23*, 2005, May 28. Retrieved July 15, 2005 from:
http://dev.dcc.rl.ac.uk/twiki/bin/view/Main/DCCApproachToCuration

Gray, J., Szalay A. S., Thakar, A. R., Stoughton, C., & vandenBerg, J. (2002). *Online scientific data curation, publication, and archiving.* Microsoft Research Technical Report MSR-TR-2002-74. Retrieved July 15, 2005 from:
http://research.microsoft.com/research/pubs/view.aspx?msr_tr_id=MSR-TR-2002-74

Hey, T., & Trefethen, A. (2003). The data deluge: an e-science perspective. In F. Berman, G. Fox, & A.J.G. Hey (Eds.), *Grid computing: Making the global infrastructure a reality*. New York: John Wiley and Sons. Retrieved July 15, 2005 from:
http://www.ecs.soton.ac.uk/~ajgh/DataDeluge(final).pdf

International Council for Science. (2004). *ICSU Report of the CSPR Assessment Panel on Scientific Data and Information*. Retrieved July 15, 2005 from:
http://www.icsu.org/Gestion/img/ICSU_DOC_DOWNLOAD/551_DD_FILE_PAA_Data_and_Information.pdf

Joint Information Systems Committee. (2003). *JISC Circular 6/03 (revised): An invitation for expressions of interest to establish a new Digital Curation Centre for research into and support of the curation and preservation of digital data and publications*. Retrieved December 15, 2005 from:
http://www.jisc.ac.uk/uploaded_documents/6-03%20Circular.doc

Kerr P., Reddington, F., & Wilkinson, M. (2005). Digital curation: where do we go from here? *Ariadne, 45*. Retrieved June 2, 2006 from:

http://www.ariadne.ac.uk/issue45/dcc-1st-rpt/.

Koehler, W., (2004). A longitudinal study of Web pages continued: a report after six years. *Information Research, 9*(2), paper 174. Retrieved 15 December from: http://InformationR.net/ir/9-2/paper174.html

Kuny, T. (1998). The digital dark ages? Challenges in the preservation of electronic information. *International Preservation News, 17*. Retrieved December 15, 2005 from: http://www.ifla.org/IV/ifla63/63kuny1.pdf

Lievesley, D., & Jones, S. (1998). *An investigation into the digital preservation needs of universities and research funders:* a JISC/NPO Study within the Electronic Libraries (eLib) Programme on the Preservation of Electronic Materials. London: LITC South Bank University. Retrieved June 2 , 2006 from:

http://www.ukoln.ac.uk/services/papers/bl/blri109/datrep.html.

Lord, P., & Macdonald, A. (2003a). *Digital Data Curation Task Force*. Report of the Task Force Strategy Discussion Day Tuesday, 26th November 2002 Centre Point, London WC1. Retrieved July 15, 2005 from: http://www.jisc.ac.uk/uploaded_documents/CurationTaskForceFinal1.pdf

Lord, P., & Macdonald, A. (2003b). *e-Science curation report: Data curation for e-science in the UK – an audit to establish requirements for future curation and provision.* Report prepared for the JISC Support of Research Committee (JCSR). Retrieved July 15, 2005 from: http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf

Lyman, P., & Varian, H. R. (2003). *How much information? 2003*. Retrieved July 15, 2005 from University of California at Berkeley, School of Information Management and Systems Web site:

http://www.sims.berkeley.edu/how-much-info-2003

McMahon, C. (2005). *Immortal information, Engineering Grand Challenge Project.* Presentation to the DCC International Conference 2005. Retrieved December 15, 2005 from DCC Web site:

http://www.dcc.ac.uk/docs/dcc-2005/c-mcmahon-dcc-2005.ppt

Markwell, J., & Brooks, D. W. (2002). Broken links: the ephemeral nature of educational WWW hyperlinks. *Journal of Science Education and Technology, 11*(2), 105-108.

Markwell, J., & Brooks, D. W. (2003). 'Link rot' limits the usefulness of Web-based educational materials in biochemistry and molecular biology. *Biochemistry and Molecular Biology Education, 31*(1), 69-72.

Memories for Life. (2005). Retrieved December 15, 2005 from: http://www.memoriesforlife.org/

Messerschmitt, D. (2003). Opportunities for research libraries in the NSF Cyberinfrastructure Program. *ARL Bimonthly Report, 229*. Retrieved July 15, 2005 from:

http://www.arl.org/newsltr/229/cyber.html

National Science Board. (2005). *Long-lived digital data collections: Enabling research and education in the 21st century*. Pre-publication Draft Approved by the National Science Board, May 26, 2005. Retrieved July 15, 2005 from: http://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf

National Virtual Observatory. (2005). *About virtual observatories*. Retrieved 15 July 2005 from: http://us-vo.org/about.cfm

O'Hara, K., Morris, R., Shadbolt, N., Hitch, G. J., Hall, W., & Beagrie, N. (2006). Memories for life: A review of the science and technology. *Royal Society Interface Journal, 3* (8), 351-365. Retrieved June 2, 2006 from: http://www.journals.royalsoc.ac.uk/(edbtkybv2awx0e552vocqcbv)/app/home/contribution.asp?referrer=parent&backto=issue,1,11;journal,2,9;linkingpublicationresults,1:111337,1

Ourmedia. (2005). *Frequently Asked Questions*. Retrieved July 15, 2005 from: http://www.ourmedia.org/mission/faq

Protein Data Bank. (2005). *Current holdings content growth*. Retrieved July 15, 2005 from: http://www.rcsb.org/pdb/holdings.html

Taylor, J. (2001). *The UK E-Science Programme*. Powerpoint presentation to e-science London meeting 27th July 2001. Retrieved November 13, 2006 from: http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/research/esci/jtaylor.pdf