LIBER QUARTERLY 19 (1), April 2009 – ISSN: 1435-5205. P1–12
http://liber.library.uu.nl/
Igitur, Utrecht Publishing & Archiving Services

Taking Care of Digital Collections and Data: 'Curation' and Organisational Choices for Research Libraries

Inge Angevaare

Coordinator, Netherlands Coalition for Digital Preservation (NCDD) and
Editor of LIBER Quarterly,
PO Box 90407, 2509 LK The Hague, Netherlands,
inge.angevaare@kb.nl

Abstract

This article explores the types of digital information research libraries typically deal with and what factors might influence libraries' decisions to take on the work of *data curation* themselves, to take on the responsibility for data but market out the actual work, or to leave the responsibility to other organisations. The article introduces the issues dealt with in the LIBER Workshop 'Curating Research' to be held in The Hague on 17 April 2009 (http://www.kb.nl/curatingresearch) and this corresponding issue of *LIBER Quarterly*.

Key words: digital curation; digital preservation; research libraries.

Introduction

Digital data are fragile. Some would argue that perhaps they are no more fragile than printed books and journals are, but over the past centuries we have learnt to deal with printed materials, and as yet we have much to learn about preserving digital information, which makes it — at least momentarily — much more fragile.

Digital data require specific care, they require so-called 'curation', which includes 'preservation', to stand the test of time. As these terms are not yet household terms in the LIBER community, I quote this definition from a 2003 JISC brochure which laid the foundation for the UK Data Curation Centre (DCC):

> ' The term "**digital curation**" is increasingly being used for the actions needed to maintain and utilise digital data and research results over their entire life-cycle for current and future generations of users. Data curation in science may be defined as the maintenance of a body of trusted data to represent the current state of knowledge in some area of research. Implicit in this are the processes of digital archiving and **digital preservation**, but it also includes all the processes needed for good data creation and management, and the capacity to add value to generate new sources of information and knowledge. In most research fields, capturing "knowledge" is more than just the archiving and preservation of source data and associated metadata. It generally involves interaction between creators and providers of data, the archivers of data, and most importantly the consumers of data. Successful curation of data requires data users to be able to utilise the data using their current tools and methodologies.' [JISC, 2003].

This article explores what categories of digital data research libraries typically deal with and how librarians might go about deciding how to organise the preservation of these different categories: leave the responsibility of curation and preservation to others, take responsibility but market the actual work out to other organisations, or take complete responsibility and implement a digital repository within the organisation.

Digital Publications: from Storage to Access

The digital age has presented research libraries with a whole new set of challenges. The first was, of course, to make the transition from printed books and journals to digital publications. This required adaptation to new acquisition methods, especially (big-deal) licensing negotiations, as well as the

implementation of new means of access, by means of the internet. One key responsibility for libraries, however, was almost imperceptibly moved elsewhere: that of storing the information. In order for libraries to serve their patrons, on-site ownership is no longer a prerequisite. Access has become the key, access to networks of information which in themselves remain with the publishers.

It cannot but have been a bit unsettling for research libraries to no longer own the information on which they rely, but rather trust their suppliers' word for a 'perpetual access' clause in the licensing agreements, but such has become the way of the world for a large portion of the records of science. The publishers, in turn, send their publications to emerging safe places such as the e-Depot of the KB, National Library of the Netherlands, Portico, and LOCKSS, quite often signing long-term archiving agreements with more than one of these organisations in order to spread the risk of data loss across multiple preservation strategies. In other instances national deposit legislation ensures that digital information is preserved for the long term, mostly in national libraries.

At the risk of simplifying matters, for the purpose of this article it is important to point out that most of these archiving agreements deal with *publications*, mostly in PDF format — and it has been established that of the many digital preservation challenges facing us, securing permanent access to comparatively well-organised publishers' PDF's is probably not the hardest job to tackle.[1] Most importantly, though, the main responsibility for preserving this content seems not to lie with research libraries, but with publishers and national and international preservation initiatives.[2]

Digitised Local Library Collections

Digitised (local) collections are the second category of digital information to be considered. As it is libraries themselves who take the initiative to digitise (special) collections to make them more readily available to the user community, it might almost be taken for granted that they would assume responsibility for curating the data involved (which have been created at great expense) and making sure they remain available for future generations. However, such

is not always the case. In the past few years I have witnessed many a presentation at the LIBER Annual General Conference dealing with preserving valuable (printed) collections for future use. And just as things would get interesting for one who is the coordinator of a digital preservation coalition, the presentation would end (in one form or another) with the following final statement: 'We digitised the collection and thereby preserved it. Thank you for your attention.'

If only life were that simple. It cannot be repeated often enough that digitisation is not the same as digital preservation. Digitisation constitutes no more than the first step, it is the creation of a digital object which is then subject to all the well-known threats of the digital age:

- media decay
- hardware or software obsolescence
- organisational discontinuity, and, last but not least
- human error.

Nobody has as yet found the answer to all of these threats for all times to come, nor is anyone likely to find such solutions. Instead, digital curation and preservation are about *risk minimisation* in a moving wall environment where data flows and technologies are changing constantly. In practice, making a back-up of the data and storing it off-site is often seen as an adequate measure to minimise the risks, but it must be emphasised that this is only a first step that by no means addresses all of the threats listed above.

If digital information is to survive, it must be taken care of continuously. In the case of digital data there is no such thing as the 'benign neglect' of the printed era, in which old books could miraculously be rediscovered after many years in dust-ridden attics. Digital information is entirely dependent on a properly functioning hardware and software environment. The Australian National Archives therefore coined the phrase 'performance model' for what happens when a digital object is accessed: object+(hardware+software)=performance. [National Archives of Australia, 2002].

Securing the success of a digital object's 'performance' requires more than just keeping a number of bits running on a server. It requires an organisation which will not only monitor the survival of the bitstream, but which will also scout the world for technological developments which might affect the

object's capacity to perform, and develop strategies to make sure the object will play on the next generation computers. Well-known examples of such strategies are migration, emulation and normalisation — quite complicated technical processes for which, and this is important, ready-made and complete commercial solutions have not yet been developed. In other words: any research library contemplating curating and preserving their own digitised collections must to some degree be willing to co-develop the technology involved, especially if it must fit into an existing access system.

On a positive note: digitised local collections usually contain only a limited number of file formats (PDF, .jpg or .tiff), which, in curation terms, are relatively simple file formats to preserve. Another important factor which distinguishes these collections from the next category, that of research data, is that it is usually libraries themselves who create them. Thus they may be expected to come with a structured set of metadata which facilitates data management and preservation.

And yet, really securing long-term access to such collections is still quite an undertaking, especially in terms of:

- the financial investment involved — not so much the storage media themselves, which get cheaper all the time, but the organisation which is needed to *manage* digital objects and keep them safe [see also Paul Ayris's & Marcel Ras's articles in this issue; also Sustaining the Digital Investment, 2008];
- the expertise needed — as mentioned previously, although commercial vendors are now entering the marketplace, both the incorporation in the library's systems and the running of a digital archiving system require a lot of local and technical knowledge.

The question now arises whether research libraries should really take on this task, especially in times when budgets are tight and are not expected to get much better within the foreseeable future. I would argue that most *national* libraries have no option; they mostly have legal obligations to act as deposit libraries for printed and digital materials. However, *research* libraries do, in my opinion, have options:

- Simply store the digital collections somewhere within the organisation and hope the best of it. This in fact is a much-practised option, but it is a risky one. Yet, if the original physical collections remain in tact, one might consciously take such risks, reasoning that digitising some lost items anew at some point in time might in the long run be less expensive than preserving the entire digital collection. From a digital preservation policy standpoint this option, however, can only work if the choice has been made based on careful consideration of the risks involved; e.g., has the question been answered whether the physical collections are in fact stable?
- Find a third party to host the collections, either as a national service or for a fee. For some research libraries this might be a very viable option, especially when the organisation is considered too small to take on the development of a digital repository. Two factors are important here: find a *trustworthy* repository, one that really is capable of applying the care needed, and integrate access to the information stored into the library system.[3] A special caveat is called for here: quite often third-party commercial vendors sell simple back-up storage facilities under some guise of long-term durability. I have seen optical storage media on sale which were supposed to last a thousand years — but which would of course not provide any safeguard against hardware & software obsolescence. In a thousand years the data might still be there, but no computer would be able to process them anymore.
- Build your own digital archive. This is by far the most ambitious and invasive measure to take, as is described in Marcel Ras's article in this issue of *LIBER Quarterly*. More on this option in the last section of this article.

Digital Research Data

Where research libraries lost the responsibility for preserving publishers' e-journals, they may well have come to be held responsible for another task: that of storing and curating research data. This is a challenge indeed, as digital research data are the most complicated category of digital information to curate: both data producers and types of objects come in many shapes and

sizes and include complex digital objects such as (live) databases. However, in an era in which the likes of Google seem to corrode libraries' traditional reasons for existence, here might well lie an important task that could revive the library's unique position at the very heart of the university's information network.

But such a position comes at a price. At last year's LIBER Annual General Conference Sijbolt Noorda, the President of the Dutch Research Universities Association, criticised research libraries for not adapting quickly enough to the digital age. He said: 'Very few research libraries developed into sustainable integrated e-support services for research and teaching & learning.' [Noorda, 2008]. Obviously, not everyone in the audience agreed with Noorda's statement, but his reasoning is well worth noting:

> 'the disparate nature of research cultures and traditions, national preferences, professional usage and language networks stand in the way of simple solutions across the board, both in e-science, e-learning and in digitally re-mastered scholarly publishing' [Noorda, 2008].

Implicitly, Noorda argued that the services offered by research libraries are often too generic to be of real value to the research community. This factor might also account for the fact that quite a few institutional repositories, which more often than not are hosted by the university's library, attract much less content than they had hoped. A study by the UK Research Information Network seems to point in the same direction [To share or not to share, 2008].

And yet research libraries have at least three crucial attributes which make them uniquely positioned to curate the output of academic research:

- they have a mission that includes long-term preservation;
- they have structural funding;
- they have a network in the research community.

Admittedly, the second attribute is a questionable one, as libraries seldom have enough funding. However, it is of a structural nature, and various studies have identified the *lack* of structural funding as one of the major obstacles for permanent access to the records of science. In the present-day academic

community, temporary research grants and project-based funding are dominant and as 'reliable preservation can suffer no gaps' [Sustaining the Digital Investment, 2008, p. 2], the data resulting from academic research are often lost when such projects end. Structural funding, no matter how modest, may be the better safeguard in the end.

As for the library's network in the research community: this attribute may never be taken for granted. As indicated by Noorda, library and data management services must be specifically attuned to your own research community, and actively finding out about their needs and wishes must always be a top priority. The most widely used standard in digital preservation, the Open Archival Information System (OAIS) framework [OAIS, n.d.], has reserved a special place and terminology for the user group: the 'designated community'. In view of the many variables at play in digital curation and the many different traditions in the research community, keeping this community in full view at every step on the way is crucial.

Such a full view might ultimately even lead to a decision *not* to get involved as research library, because while research libraries were still grappling with the notion of a digital future, some research communities proved themselves early adaptors and embraced the digital future by organising themselves around domain-specific information networks, often uniting researchers from around Europe or around the globe. Such networks are often well-established and well-attuned to researchers' needs. Research libraries have little to offer these communities, as their one great need, sustainable funding on sometimes quite a large scale, is one that research libraries cannot meet.

The research landscape is patchwork landscape, where local, national and international data networks intertwine with generic and domain-specific networks. Just recently, Chris Rusbridge of the UK Digital Curation Centre posted some interesting thoughts on the DCC weblog as to how they could interrelate [Rusbridge, 2009]. The trick of course is to find your own specific place in that landscape and to cultivate it.

A Provisional List of Do's and Don'ts

This article is but an introduction to the field of digital curation and preservation, and therefore many issues must remain unaddressed. I would like, however, to end on a practical note, so here is a provisional list of do's and don'ts for research libraries who are trying to decide how to handle the digital collections in their care and the digital needs of their research community:

- Find your 'designated community'.
- Sit down and make a plan, formulate a policy about what you are going to offer your designated community. Often the mere act of sitting down and writing a plan forces you to analyse your strengths and weaknesses and to get your priorities straight.
- Make sure the plan includes an inventory of the digital collections you have in your custody and an estimate of what is coming your way in the foreseeable future.
- Talk with your designated community about their needs and include those in your plan.
- Deselect and deselect again. Although some archivists still champion the cause of saving 'everything' (arguing: 'who are we to decide what the future will need?'), most analysts have agreed that saving 'everything' is neither feasible nor desirable.
- Scale matters. The first digital object you curate is outrageously expensive, the millionth hardly costs a penny. If you do not have enough scale yourself, go to the next bullet.
- Find partners, preferably within your own domain and of similar size. This may be tricky, as you will probably be competitors when it comes to attracting top researchers and top students. Yet experience shows that collaboration works best when partners are alike and gets more difficult as organisations are further apart in traditions and purposes [Zorich, Waibel & Erway, 2008].
- Find umbrella organisations with networks of expertise. These may be regional, national, or international [see: Lossau & Peters, 2008].
- Pamper your designated community. Let researchers be researchers; do not ask of them that they adapt to your (metadata) schemes, but strive to provide tools and methods that make it easy for them

to integrate data management in their workflow [To share or not to share, 2008].
- Find your own specific quality and place in the network of curation organisations.

And if all else fails:

- Never store important digital information on floppy disks, cd-rom's or local computers, but store it on more robust hardware.
- Make a back-up of your information regularly and make a deal with a colleague that you will take care of each other's back-ups.
- Make an inventory of the digital information you have in your custody and keep it up to date.
- Find a trustworthy custodian for your digital data. This may be a national library or a national archiving organisation; it may also be a colleague that has implemented a digital repository. If there is no official way to gauge an archive's trustworthiness, look at the organisation as a whole and ask yourself: is this organisation itself likely to be around fifty or a hundred years from now?
- Concentrate all your efforts on *access*, because in the end *access* is what matters. All else is but a means to make *access* possible.

References

JISC (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*, http://www.dcc.ac.uk/docs/6-03Circular.pdf, accessed 15 February 2009.

Long-term Preservation: Results from a survey investigating preservation strategies amongst ALPSP publisher members (2008), prepared by Sarah Durrant, http://www.alpsp.org/ngen_public/article.asp?id=&did=47&aid=27202&st=&oaid=-1

Lossau, Norbert and Dale Peters (2008), 'DRIVER: Building a Sustainable Infrastructure of European Scientific Repositories', LIBER Quarterly 18/3–4, p. 437–438, http://liber.library.uu.nl/publish/articles/000267/article.pdf

National Archives of Australia (2002), *An Approach to the Preservation of Digital Records*, http://www.naa.gov.au/Images/An-approach-Green-Paper_tcm2-888.pdf

Noorda, Sijbolt (2008), 'The Impact of Digitization from an Academic Point of View', powerpoint presentation at the 2008 LIBER Annual General Conference, Koç University, Istanbul, 1 July, http://www.ku.edu.tr/ku/images/LIBER/istanbul_noorda2.ppt

Open Archival Information System (OAIS), http://en.wikipedia.org/wiki/Open_Archival_Information_System and references listed there.

Rusbridge, Chris (2009), 'A National Research Data Infrastructure?', weblog, 5 February, http://digitalcuration.blogspot.com/2009/02/national-research-data-infrastructure.html

To share or not to share: publication and quality assurance of research data outputs (2008), report prepared by the Research Information Network, http://www.rin.ac.uk/data-publication

Sustaining the Digital Investment: Issues and Challenges of Economically Sustainable Digital Preservation (2008), Interim Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf

Zorich, Diane M., Gunter Waibel and Ricky Erway (2008): *Beyond the Silos of the LAMs: Collaboration among Libraries, Archives and Museums*, Report produced by OCLC Programs and Research, http://www.oclc.org/programs/publications/reports/2008-05.pdf

Websites Referred to in the Text

DANS, Data Archiving and Networked Services, http://www.dans.knaw.nl/en/

DCC, Digital Curation Centre, http://www.dcc.ac.uk/

DRIVER, Digital Repository Infrastructure Vision for European Research, http://www.driver-support.eu/en/

e-Depot of the KB, National Library of the Netherlands, http://www.kb.nl/dnp/e-depot/e-depot-en.html

LOCKSS, http://www.lockss.org/lockss/Libraries#Netherlands

NCDD, Netherlands Coalition for Digital Preservation, http://www.ncdd.nl/en/index.php

Portico, http://www.portico.org/

Notes

[1] This is entirely comparatively speaking — I am well aware of the many obstacles still to be overcome in both a technical and an organisational sense.

[2] This is the de facto situation. However, a recent survey of the Association of Learned and Professional Society Publishers (ALPSP) revealed: 'Publisher views on who should take responsibility for long-term preservation also reveal some interesting contradictions: despite presently supporting a range of preservation schemes, a significant majority of publishers indicated they would in fact prefer other groups and institutions to take this responsibility on. National libraries in particular were a popular choice.' [Long-term Preservation, 2008].

[3] Various tools and methods have been developed to measure a repository's trustworthiness, see, a.o., Barbara Sierman's article in this issue. A very basic tool is the Data Seal of Approval developed by Data Archiving and Networked Services (DANS) of the Netherlands. See 'Data Seal of Approval, Dissemination, Assessment and Procedures', powerpoint presentation by Henk Harmsen of DANS at the Digital Preservation Workshop, The Hague, 30 January 2009, http://www.datasealofapproval.org/files/20090130_Harmsen.ppt. At the very least the archive must have a long-term mission and sustainable funding, and it must offer guarantees with regard to authenticity and quality of the data.