

2nd LIBER Workshop on Digital Preservation

Florence, Italy, 7th–8th May 2012

About the event

‘Do not go at this alone’ was the conclusion from the first LIBER digital preservation workshop in 2009. Since then the Euro crises and budget cuts have made partnerships even more important. But with whom to collaborate, and how? How does it work and what responsibilities can you share? This workshop was hosted by the Fondazione Rinascimento Digitale in Florence, a not-for-profit agency that promotes good practice and supports dialogue in the use of internet technologies, and with LIBER the consortium of European Research Libraries.

WK represented the DPC and other DPC members were present in their own right – Joy Davidson, (HATII/ DCC), Liz Lyon (UKOLN/ DCC) and Patricia Killiard (CUL).

These notes are intended to provide an informal briefing for members of the DPC not able to attend the event. For an authoritative and comprehensive repost readers are encouraged to contact the organisers of the event or speakers directly.

Presentations and discussions

Inge Angevaere, NCDD – Setting the Scene: Digital Preservation and Partnering in the Global Information Space

Inge introduced the workshop and explained how it had been organised. She started by discussing the role of research libraries. Information sharing has always been at the core of research libraries because they assume that you can ‘talk to your predecessors’. But it is hard to turn this infrastructure into digital form and librarians have worried about what would remain of the research library once everything has been digitised. A new challenge has emerged at the same time – the need for digital preservation to ensure that data remains robust and accessible. Research libraries look after more than just ‘research data’ but this is a category of particular importance. Data lifecycles are extended and they involve the creation and planning ahead of creation. But this is IT intensive and it requires a kind of expertise that libraries don’t have. The library needs to be re-positioned and re-skilled, and that is where the collaboration is important. Technical challenges are a breeze in comparison to the organisational relationships we will need. ‘Connection not collection’ is a motto for the digital age, but it appears like a threat to local services and research libraries. Partnerships however are not just about similar institutions – it needs also to include the creators, publishers and the users of the data. These are very diverse and can be difficult to engage with. It’s not like the paper world. Where the workflow is internal – the digitised book for example – the management can be simple and planned well in advance. But where the workflow involves external parties, especially long chains of external parties and rights, then the management is a lot harder. Websites contain many different digital objects and many different contributors; research data may need complicated representation information and so forth. And this gives us a basis for other types of collaboration between institutions: preservation planning for example, or shared infrastructure

like LOCKSS. We can structure these relationships as equals and on a community basis like LOCKSS or establish relationships with specialists with service level agreements like PORTICO, or establish national agenda like KB. Even so there are 'homeless' data which need attention.

All of this means a new role for research librarians and needs for clear policy. They need to develop new competencies and knowledge and to develop policies that are consistent with their organisational mission. This means keeping up with trends and emerging solution, and it means advocacy at high levels within an organisation. KB makes preservation plans around 'parcels' of data and sketches out the plans based on expectations of use and access which are based on experience.

Don't wait for perfection: 'Dare to make your choices!'

Liz Lyon, UKOLN / DCC – Partnering for Research Data

Research data comes in all shapes and sizes. Genomic data, environmental data, satellite imagery, mobile phones. Some is large and homogeneous, some is small and heterogeneous. The real challenges are in the small heterogeneous collections. There are numerous research data stakeholders: national research academies, funders, data centres, universities, publishers, national libraries, large research projects and so on. Data also tends to make sense at a subject specific level and individuals tend to make their decisions influenced by their disciplinary needs. A research librarian needs to manage 7 key relationships: Director of Information Services; Data Librarian; repository manager; Computing Services; research support office; doctoral training centres; VP for research; and in addition there may be a relationship with a public engagement office where they exist. (More detail published in a recent paper in IJDC). There are tools and services available for different players from DCC. The data asset framework for example is helpful in establishing what data you have and what's valuable; Cardio helps you assess the extent to which you are ready for managing research data: these tools are really helpful for advocacy in addition to the information they present. Data Management Planning is helpful and the 'dmponline' tool allows an assessment before data is created, and fits with research grant applications. Advocacy and training can help ensure that data is produced in formats and with documentation that is consistent with institutional and discipline specific forms. DCC also has a set of guides on things like licensing, appraisal and citation of research data. Total Impact help researchers track references and citations to data. These tools and services are really useful though they don't address the core skill areas. Research shows that there are perceived skills gaps for subject librarians. A recent report from RLUK shows that gap between the skills needed and the skills currently available: for example RLUK shows that only 10 percent of staff have skills in preservation but that in the near future this will need to grow to 49%, and for subject specialist metadata this will have to grow from 2% to 16%. McKinsey report a shortage of around 190,000 data 'scientists' managing big data by 2019. There is a role for LIBER here - a massive gap in skills and expertise. We can define core components of data informatics such as discovery, domain ontologies, workflows, visualisation, analysis and preservation. We can increase the threshold of maths and sciences entry requirements for LIS students so that the profession is changed. Finally we need an international data informatics working group to explore promotion, recognition and reward. LIBER could lead on all of these.

- What is the 'locale' of the research library? Won't these services end up being highly distributed around disciplinary needs and not end up as recognisable research libraries? Perhaps, though it's hard to imagine this not as a blended landscape. It is still a dramatic change with great opportunities for those that make the change most effectively.
- Institutional versus national infrastructure? There is too little information about the costs of the data repository, though Keeping Research Data Safe provides some answers to this. The real question is 'who pays' and no one seems to have answered this properly in the UK. There are obvious data centres in some disciplines and there is no need for a competition between these. In fact the two can work together well provided they are in touch with each other and properly planned.

Marcel Ras, KB and Randy Kiefer, CLOCKSS – Partnering for Academic Journals

Two terms to play with permanent access and digital preservation – the two are connected but they are subtly different. E-journal content is attractive for all sorts of reasons such as time availability, space savings and so forth, but they have a different business model. That means the library does not own content they have paid for and 'post cancellation access' is now problematic. Access can be lost so service providers have entered this space. Usage of ejournals continues to grow and the budget for e-journal purchases is growing. More money is spent on e-journals than on print by a factor of 6. But only 15% of the collections (Cornell and Columbia) are archived in either Portico or Lockss. Libraries need to be concerned about the archives that will support access should the publishers go bust, and need to develop relationships with archival services which help them understand the services on offer. The International EDepot exists to guarantee continuous access to academic publications. Operational from 2003 it has concentrated its efforts on ingest and preservation and is now moving to a post-cancellation service which can be triggered. CLOCKSS has a similar service provision though is set up really differently with a global alliance to look after e-journal content.

Gianluca D'Amato (Universita Catolica del Sacro Cuoro Milano) – Towards distributed digital preservation

Sociologists refer to the 'communication society' but they really mean the 'recorded communication society'. A document is a social object and recording is part of the social process of communication. Archives and libraries existed to support that recording and provide resilience for the long term and in a way to provide access to these collections. This worked for paper because the media was mostly robust. Is it really sensible to provide the same model in the digital age? The two roles – access and preservation seem to diverge but access depends on preservation at a point in time, and assumes allied features like provenance, integrity and authenticity. So they are complimentary not contradictory. But recognising that the roles will be different, it is not at all clear which organization does what. Paper libraries and archives depend on commitment of specific agencies (archives and libraries) as well as spontaneous initiatives by interested parties who have been able to preserve collections though a sort of passive preservation. We know that passive preservation will not work for the digital age, and the policy framework sometimes is not up to the task. So we need to establish some principles – custodial responsibility, appropriate redundancy, open source, what to preserve and so forth. A particular issue is the need to support distributed action and the

Distributed Digital Preservation Model offered by Lockss is therefore helpful because it is explicitly collaborative, based on open source software and it replicates content to numerous partners. A Global Lockss Network provides an implementation that UCSC have used to support their own preservation efforts.

Eric Meyer (OII) – Partnering for Web Archiving

The question is how to preserve the web in such a way as to make it useful for researchers. The Internet is here to stay even if the pages are unstable and it's a pretty important phenomenon of its own, but only if we understand what researchers want or need will our archives be any use. Preservation is not enough. The Internet Archive offers a search interface to the 'saved' internet and is good for individual pages and references. But new devices change this, such as smart phones. Web Archives don't really take account of the new types of interface that people have with new types of web pages and very many transactions that underpin it. Facebook, for example is never the same: there is no home page and no canonical view to archive. It's also getting more and more complicated and more and more interdependent. Even more true about gaming and about virtual worlds. Therefore web archives need to learn more from the live web. Tools to analyse the live web simply don't work for web archives. The APIs for web archives are mostly rubbish if they exist at all. How would OII look at collections in 2020? Life-logging? Cross-searching? Webtracks? Images of a changing world? Collecting priorities need to respond in real time to the changing world. Web logs are also fascinating to help understand web traffic and touch on the key issues of the social sciences about people's behaviour and interactions. One of the problems is that users are seen as passive consumers but that's not true: they use and consume and change the data that they engage with. Moreover, the way they interact changes the way the resources are configured: these feedback loops are a lot faster and a lot more complicated than the concept 'user' suggests. The relationship that users have with online resources is not the same interaction that they have with lending a book. Key theme – the digital age is not just about the change in media but it also re-invents the consumers. Web archives require national legislation and that is counter-intuitive because the internet is pretty diverse and has lots of connections.

Barbara Sierman (KB) – Preservation Policies: necessary and beneficial

40 million tons of electrical equipment is thrown away in the US every year. Massive volumes of data are produced every year. Libraries need to sort some of this junk and work out what can be recycled(!). Preservation policy is a written statement authorized by the repository management that describes the approach to be taken by the repository for the preservation of objects accessioned into the repository. A policy describes the intentions of the organisation and how they will be realised. It is guidance for the whole organisation and it needs to be effective – ie approved and committed by the top management, visible, published, implemented and updated. It should be part of the genetic code of the organisation. It's not just paperwork: it's the start. They mean change of staff is possible, they allow exchange of knowledge and education. They harmonize activities, embed them in workflows and make responsibilities clear. Policy also helps clarify relationships with external agencies – outsourcing processes like digitisation or storage is only really possible when responsibilities and expectations are clearly defined and stated upfront. Digital is different from analogue and there is need for very great change in processes. Very different skills,

very different risks and all too easy that we wait for someone else to shoulder the responsibility for the collection. In 2009 PLANETS studied the preparedness for preservation in a variety of institutions and noted that there is a wide base of awareness, but that tools and services are under-developed, implementation was weak. An improvement on previous surveys but it also became clear that agencies with a policy framework were in better shape. There is a large audience for a preservation policy including partners outside your agency. Policy is increasingly a research and development subject of its own. There are some useful places to start – PLATTER and Digital Preservation Policies Study.

Joy Davidson (HATII / DCC) – Training the right staff

What skills are needed to help define and implement policies? Focus on practical skills rather than formal education, and co-operation between staff within an agency, based on lessons learned by the DCC working with UK research institutions. This has been prompted by EPSRC requiring institutional roadmaps required by 2012 and compliance required by 2015 – nine distinct expectations to be met in the next three years. The funding bodies have quite a lot of requirements but very little support, and the funders expect that the institutions will pick these up. There is an implication for loss of income and loss of competitive advantage for those who fail to take this seriously. Responsibility for preservation is shifting to the institution. Local support is therefore vital. Add to this the research integrity issues that institutions have faced means there is a strong and growing reputational risk for those who are not able or willing to share data. All of this points to a need for skills. This includes identifying what researchers are doing and what they already have available, such as the Data Audit Framework. Codes of practice and existing bodies within the organisation are likely to be critical to success and have the levers to put policy into practice. This might include the institutional training policy.

Giovanni Bergamin (Biblioteca Nazionale Centrale di Firenze) – The Digital Stacks Project, Magazzini Digitali

Digital stacks are the long term electronic deposit infrastructure. It started in 2006 and is now developing to a more complete implementation of the relevant legislation. The stacks grow, it is hard to predict usage, and deletion or modification is not permitted. There are three sites – in Florence, Rome and Venice (where the dark archive is stored), though the users interact with a single interface. The service uses the ISO 27001 security standard and there are three different data centres all of which are remote from each other and three sites were commissioned at the end of 2011. The digital stacks are governed by law so there is a strong regulatory environment and there are 4 partners – the three national libraries and also FRD that are held together by a series of agreements. There is also a general agreement between the ministry and the publishers which adds to the regulatory environment including mandating OAIS. Only registered users can access the digital stacks and only on the premises of the library. Access is available in principle for other libraries but under a different licence. The core is funded by the government but there is expectation for other sources of funding such as from publishers as a ‘perpetual access’ service. There is a fear of competition and piracy from publishers but there is also recognition of the role of the deposit libraries. In the view of Digital Stacks, long term digital preservation is offered as a public service which ensures viability, renderability, authenticity and availability. The project is using

the 'archivematica' approach using micro-services to provide an OAIS compliant repository. Access is complicated by the regulation. Access can be via a laptop and browser to connect to an access service which provide the 'frames' not the underlying files. As far as possible the system prevents people from downloading files to their own computers in order to protect copyright.

Ivan Bosserup (National Library of Denmark) – Preservation of Email and Attached Documents

Ivan introduced a pilot project from the Royal Library of Denmark called myarchive.kb.dk. Correspondence is an important historical source and it forms a part of a lot of traditional archives. But traditional correspondence is declining rapidly and whereas they have simply 'printed' email, this is neither a very successful process nor very sustainability. So capturing correspondence in electronic form is the future, and doing it happily and invisibly. Ideally they will let people organise their own archives while they are live and make sure there is a simple workflow which ensures it will be exported to the national library (and partners) for preservation in due course. A range of authors will be given small hard disks to capture their email which they will curate and organise in the first instance then send to the library. A pilot project ran from 2009-10 based on webmail and this was a success. Feedback showed that people used this as a tool in their daily work as well as for the long term, easy distinction between 'rubbish' 'useful for a while' and 'need to keep' which were the three categories.

Bram van der Werf (OPF) – Developing digital preservation tools and practices: partnership

Developing and sustaining software almost always means partnerships. At route there is a 'make or buy' question about digital preservation tools. Software breaks however so management through time means some kind of partnership – either a commercial partnership with a vendor who can fix it, or a consultant who will fix it for you, or a collaboration that provides skills in return for some other service. Retaining all the skills in house is not practical for most agencies. Technical staff are very mobile and this creates vulnerability for agencies, and it can undermine the fitness for purpose and certification of an agency – so it's a serious problem which will make its way all the way to senior management. Orphan data and orphan processes are problematic. We know pretty well that orphan data can be looked after: but orphan tools or processes or software are harder to manage. The question in both cases however is who owns the problem and are the people who own the problem also the people who can solve the problem. Can they understand all the dependencies and resolve the issues that may arise? Libraries are used to managing relatively large volumes of controlled data, archives are more used to smaller amounts of chaotic data.

Chiara Cirinna (FRD) – APARSEN: towards a virtual centre of excellence

APARSEN aims to counter fragmentation in the research and development in digital preservation.

Matt Kibble (ProQuest) – Serving God and Mammon: cultural curation and national access

Matt introduced to a new project between ProQuest and the National Library in Florence on their early book collections. This project shows how a public cultural heritage agency can work with a for-profit commercial agency in such a way as to deliver return on investment while not sacrificing the values of the public good. ProQuest's Early European Books project means that ProQuest pays for

the digitisation of the collection and the national library receives the master copies for their own purposes. The whole collection is in scope, and that means shelf by shelf digitisation. The volumes are online and available for free in Italy from ProQuest. There is an embargo on international access which allows ProQuest to recoup the costs through subscriptions, though in time the access will be free for all. July 2012 the scanning studio was set up, and there is now a workflow to build up content. Most libraries have not needed extra staff, provided that there is sufficient planning to ensure that library services are not disrupted. A portable scanning studio is created in the library and a rapid throughput of about 5,000 pages or 24 books per day is the norm. QA and editing happens offsite. Spines, edges and blank pages are also captured and care is taken about the conservation which is managed throughout the process. A great deal of interesting material about book history is also produced as a result, and it's interesting to note how international these Italian collections turn out to be. The aims of the library and ProQuest align very well. A commercial supplier provides the capital and there's little need for public funds to be spent, while access for library's users is guaranteed.

Joy Davidson (HATII / DCC) – The Keeper's Registry

Keepers project is an e-journals preservation registry: who is looking after e-journal content and what are they doing. Search by ISSN and work out who is looking after what content.

<http://thekeepers.org/> It shows that there has been good progress: around 20,000 titles have preservation agencies associated with them, whereas there are around 98,000 titles in the ISSN serials register.

About this document

Version 1	Written at conference	7-8/05/2012	WK
Version 2	Distributed	8/05/2012	DPC members, FRD, LIBER