LIBER QUARTERLY 18 (3/4), December 2008 – ISSN: 1435-5205. P324–336
http://liber.library.uu.nl/
Igitur, Utrecht Publishing & Archiving Services

Supply and Demand: Special Collections and Digitisation

Ricky Erway

RLG Program Officer, OCLC Programs and Research,
777 Mariners Island Blvd. Suite 550, San Mateo, CA 94404, USA,
erwayr@oclc.org

Based on a powerpoint presentation at the LIBER preconference: Discovery to Delivery — A European Resource Discovery Space?, which is available at http://www.ku.edu.tr/ku/images/LIBER/erwaylibercomp.ppt.; © OCLC Online Computer Library Center, Inc.

Abstract

Based on the outcome of an RLG event entitled 'Digitisation Matters', the author presents a number of thought-provoking and sometimes provocative ideas to make more of libraries' special collections available on the web, stressing access and quantity as key factors.

Key Words: Digitisation; special collections; research libraries

I have been involved in a number of digitisation initiatives since the late 1980s all dealing with special collections. In each case, preservation was a key objective. Appropriately then, we would identify the highest quality standards feasible at the time and adhere to them, assuming that we were only going to get one chance to digitise, so we'd better get it right.

Since then, I have come to think that digitising special collections for preservation no longer makes sense.

The careful, high-quality capture projects that I and others worked on over the last twenty years resulted in some delightful digital collections. But collectively, our efforts were not making an impact. Innovative at the time, our gorgeous, expensive, hand-crafted websites attracted very few users. Times are changing and we need to keep up.

Google has completely changed the way we think about digitisation of books. We considered books the hardest part of library digitisation (with their sequence and hierarchy, hundreds of images, illustrations, and text conversion) — we did not imagine we would ever digitise *all* the books. But by focusing on quantity over quality, Google has made the seemingly impossible appear quite doable. And users are swarming. Google supplied — and changed demand.

Now, with millions of books flying off the shelves and 'into the flow', we need to change the way we think about digitisation of special collections. Soon students will only search online — what is not there, will not be considered.

Richard Ovenden, Keeper of Special Collections at the Bodleian Library at Oxford, says that when institutions are thinking about their priorities, one of the main drivers in this age of increasingly ubiquitous access to electronic information is going to be what makes their institutions stand out — what makes them unique.[1] He, and I, contend that that is our special collections.

As we increasingly share a collective collection of books, it is the special collections that will distinguish our institutions. Yet ironically, special collections run the risk of being marginalised. If they are not accessible, they are not used; if they are not used, they may go away. Neglect can lead to obsolescence.

Manuscripts, archives, ephemera, photographs, postcards, scrapbooks, clippings files, maps, architectural records — so many of our special collections are hidden, unprocessed in boxes in warehouses. No one knows what's there.

While unprocessed materials in special collections are absolutely hidden from view, even processed materials can be hidden. Our first order of business is to describe these collections, but then we must push those descriptions to the surface of the web where they can be discovered. No one is demanding to look in our supply rooms, as they have no idea what is there. Do we keep those treasures hidden or push them into the light of day? Describing them is just the start.

The Research Libraries Group (RLG) held an event in September 2007, called Digitisation Matters, to discuss ways to increase the scale of digitisation of special collections. We challenged our speakers to make very provocative suggestions.[2] The audience, of over 200 people, discussed the relative merits of the proposals and Jennifer Schaffner and I wrote the Shifting Gears essay,[3] based on the outcomes of that event.

This paper will highlight some of those ideas. They do not all make sense in every situation, but some of them will be applicable for most collections. The second part of the paper provides examples of places where the principles are being put into practice.

* * *

Due to their very specialness, we are committed to preserving the originals in our special collections. Why then must we approach digitisation from a preservation point of view? Of course, there are exceptions (like brittle books where we may, in fact, only get one chance to reformat them), but in the main, our digitisation should be in service to increased access. By increasing access we increase the perceived value of our collections. If we fail to make our collections better known, we may no longer have sufficient funds to, or even be employed to, continue collecting and preserving originals for our collections.

Here are some of the provocative recommendations that came out of the Digitisation Matters discussions:

- *It is time to think about digitisation in terms of access and begin to unlock our collections.* We need to safeguard the originals and set the digital copies free. This shift in scanning for preservation to scanning for access is an especially important point for RLG to make, since we were at the forefront of the do-it-once, do-it-right parade, waving high our recommendations for high quality practices and standards. Seeing how much content is available after twenty years and seeing how little it is used — especially in light of mass digitisation of books — has caused us to change our tune.
- *The selection has already been done*. Too much of our digitising time and effort is spent on deciding what to digitise. We have already carefully chosen the materials; they are in our collections for good reasons. So we should stop being so selective about what to digitise, which often amounts to trying to guess what our users need. It is better to scale up digitisation and make more of our great collections accessible, so that users can make their own selections.

- *Do not let yourself get further behind.* We should consider scanning an entire collection as it is accessioned and if that is not feasible, we could scan small portions (samplings or a series), then see if the demand supports more effort.
- *Stop thinking about item-level description.* Learn from archivists and let go of the obsession to describe items. Think in collections and arrange and describe unique materials in subunits. We have got to give up our slavish devotion to perfection. But whatever the level of cataloguing, it must be made web accessible. Putting the most minimal description on the web will not restrict use anywhere nearly as much as limiting discovery to those who are able to show up in person and who know whom to ask for what.
- *When scanning for access, quantity is the goal.* It is great to have access to a few nice pictures, but the real discoveries are likely to happen when researchers have access to vast amounts of photos. We often take longer to select a few from a large number of photos than it would have taken to scan all of them. Lorcan Dempsey says, 'Quantity has a quality all its own. A focus on quality is one reason that libraries, archives and museums have not moved their collections in large quantities to the web. This reduces their visibility and impact as the web becomes central to research, learning and civic engagement. Scale matters and fragmented small-scale activities do not map well onto behaviors in a web environment.'[4]
- *Build infrastructure* — We should be engaging in on-going programmes, not special projects. We need the organisation and workflows to make our efforts ongoing, sustainable activities. This is work essential to our missions, so it should be built into budgets, into the overall technology platform, and into our processes. It is time to stop thinking in terms of projects and start thinking about ongoing processes.
- *Encourage our funders to support those programmes that operationalise digitisation*, rather than building project-based portals. Now that the private sector has shown how easy book-scanning can be, we have taken to wringing our hands about the challenges of special collections. Our private sector partners *made* it easy by creating equipment and processes that made book scanning more routine. The Internet Archive has developed technology using pro-sumer digital cameras and open source software that brings scanning costs down to around 10 cents a page. Now we need similar gear for special collections to automate the process and make scale possible. Let us encourage our

funding agencies to help us digitise our special collections by encouraging development of format-specific equipment.

- *Let the users guide you:* Start by scanning the collections that are heavily used in the library. Consider scanning on demand — let your researchers identify what is of interest, and as long as the material has been paged and is in your hands, you capture it and send it to the repository. Or scan signposts from a collection and let the level of interest guide future efforts. But once we have some special collections digitised, we have got to make access easy. Mostly we have been creating portals that lead to other portals that eventually lead to deep collections. Each collection has to be discovered and searched individually. How many of them do how many users ever find?
- *Discovery happens elsewhere*. Where are our users? The 2008 Web Trend Map[5] shows the 300 most visited sites where our users are going. Amazon, Yahoo, Google, Wikipedia, and YouTube are dominant, but they are not overwhelmingly dominant. However, there are no libraries in this picture. No OCLC, no JISC, no TEL … What does their absence tell us? The environment is dominated by large-scale information hubs. Users bypass the authoritative content of libraries in favour of just-in-time information from sources more convenient to their daily networked lives. Discovery happens elsewhere — we need to be there.

Those are some of the provocative proposals. The following are some promising signs from the community. These were not necessarily influenced by *Shifting Gears* (in fact many predate it), but they exemplify the recommendations in an encouraging way:

- The special collections library at the University of Texas has stopped providing photocopies on request.[6] Instead they scan on demand … and send the images to the digital repository. Eventually they will be able to serve the most requested items digitally to local, and remote, users. [It also helps to preserve the originals, saving them from repeated photocopying.]
- The University of Wisconsin has a project to streamline digitisation.[7] Their control model, using their old approaches to scanning and metadata, yielded a cost per page of $1.53. Their streamlined model, scanning in bulk with no individualised documents or item-level metadata, was $0.33/page. While users preferred the output of

the control model, they opted for more stuff over more description. The University is now looking at ways to automate the isolation of individual documents and at improved approaches to browsing. Josh Ranger, spokesman for the project, says: 'every dollar spent to The Archives of American Art at the Smithsonian is using two tactics at one time. They are scanning on demand for reading room requests — if they touch it, they scan it. Even if the request is only for a page, they scan the whole thing (an entire book or an entire folder of correspondence). They are building quite a resource of materials that, having been requested, are likely to be of interest again. And they are methodically scanning entire collections according to their archival arrangement — no cherry-picking, no decisions. The result is unprecedented access to the content and context of thousands of documents, photographs, diaries, sketches, writings, and rare published materials.[8]

- The EU Commission's i2010 strategy for the Information Society[9] announced a flagship initiative on Digital Libraries and identified the need to reduce costs for mass digitisation of specific materials, saying digital libraries are not just a short-term goal but a process.
- The Carolina Digital Archives[10] at the University of North Carolina got a lot of money to digitise a whole collection, instead of cherry-picking 'interesting' samples. They say, 'Since we have so much time and money for a relatively small project (22,000 pages), we're able to spend a lot of time on research and experimentation on mass digitisation of archival materials. The findings will inform other projects, and the workflow will then be used on a much larger scale.'
- The Hugh Morton Photograph collection at the University of North Carolina includes 200,000 slides. The high demand for use of the collection argued for a non-traditional approach to processing. The collection is unorganised, so they are sorting it chronologically and then by format, right down to the type of film shot, to allow for automated digitisation and processing. Looking at each slide on a light box for five seconds (which alone would take almost a year) to make selection decisions as to what to digitise, would cost more than scanning all the slides. One of the processors said 'all I am doing is taking one huge pile of stuff, sorting it into smaller piles, sorting those piles into smaller piles, and, finally, describing the piles.' They are planning to make parts of the collection available incrementally as they

complete them. And then let the users do the cherry picking, according to their needs. The staff on the Morton project wanted to keep people informed about their progress (and offer glimpses into the collection's wealth). So they developed a blog[11] to meet those needs. They are moving towards methods that emphasise access over preservation, aim for quantity over quality, and that let the users decide which 'cherries' they want to pick — and they are sharing what they learn with others.

- We are making progress with purpose-built scanners. Things have been pretty much worked out for microfilm scanning — it is fairly customary to capture 100 frames a minute. At RLG, we are in conversations with a vendor who is working on efficient document handling for photos, postcards, and other flat items. They estimate a throughput rate of about 250,000 images in three months. The Stokes Imaging System consists of a capture station, various material handling devices, and imaging workflow software. The distance and angle of the camera can be computer-positioned to take advantage of the various handling devices. High-quality images are captured and access quality images are automatically generated. Material handling is the limiting factor in increasing throughput, so Stokes has developed several devices to expedite this process while assuring the originals will not be damaged. The manuscript cradle and the book cradle allow non-damaging digitisation of images from precious materials, at a rate of 100 images per hour, while assuring consistent, sharp focus. De-skewing and cropping is automated. The auto-slide loader, which handles a stack of 175 slides in 45 minutes, and the roll film transport, which handles up to 400 feet of film at a rate of 6 frames per minute, both pass the original through an electrostatic cleaner with fine brushes and position it for capture. The cut-film platen is designed to facilitate the handling of negatives and transparencies up to 11 x 14 inches. A conveyor captures similar-sized reflective materials at the rate of six colour images per minute. Stokes is currently working on a robotic postcard handler and is interested in hearing about other needs.
- The Koninklijke Bibliotheek (National Library of the Netherlands) recently released a report[12] that looks at reframing digitisation practices. Because of the scale of projects and the economics of storage, they are purposefully reconsidering file formats and quality require-

ments, aiming for 'good-enough' results. They plan to distinguish between master image files which must be stored for all 'eternity' (for preservation reasons) and objects which are stored for access. This will allow for more pragmatic and efficient storage. The KB has also overhauled their whole approach to digitisation to emphasise quantity over quality.

- And we are seeing some encouraging signs from funders. At RLG, we have talked with US government funding agencies and philanthropic foundations. They are looking for more effective ways to fund digitisation and ways to help change our practice, to make more rare and unique material available. The US National Archives funding body[13] is seeking digitising projects that can repurpose existing metadata. The Mellon Foundation, through CLIR,[14] is launching an aggressive programme to describe hidden collections, the first and most important step in increasing access.
- The German Research Foundation, or DFG, is the central funding organisation that promotes research at publicly financed research institutions in Germany. They revised their practical guidelines[15] for funded digitisation projects. They want the materials that were once difficult to get a hold of or vulnerable to damage to be easily viewed at home or on library computers. They want digital documents linked to other online resources — such as catalogues, bibliographies, secondary literature etc. — so the potential of the Internet is fully leveraged. Thus the objective is not only to make these materials available, but to integrate them into a network. As a general rule, they say, there is little need for experimenting with novel techniques, since a wealth of experience is already available. Digitisation is a standard service to provide, rather than a distinctive feature. Digital access to cultural heritage should be the rule rather than the exception. They support libraries, archives and museums in their endeavour to provide digital versions of as many research-relevant public-domain holdings as possible within a short period of time. They say that all projects should be designed such that their results will be available to researchers quickly and for the long term. This entails the provision of free access to digital copies on the Internet.
- An RLG-sponsored study[16] to assess the public/private partnerships for mass digitisation determined that one of the important capabilities that has been given up is the right to freely create aggregations of

the digitised content. The guidance in this report should help institutions negotiate better deals in the future. A high-level expert sub-group of the European digital library initiative produced a similar report in May 2008.[17]

- The US National Archives have managed to arrange partnerships with private partners to do digitisation on their own terms. They developed their own criteria[18] internally and now go into negotiations with potential private partners knowing exactly what their walk-away points are. And they shared the criteria with the community for comment — as well as the agreements.
- The National Archives in the UK say[19] it is now primarily a digital archive. 100 times more documents are delivered over the web than on paper. They have made a strong argument for industrial-scale digitisation, believing that only wholesale digitisation of complete collections delivers full benefits. They have in excess of 60 million digital documents. The entire archive would cost 5 billion pounds to digitise. Dan Jones, the head of business development, does not despair at the size of the task. He quotes the French Marshal Lyautey who asked his gardener to plant a tree. The gardener objected that the tree was slow growing and would not reach maturity for 100 years. The Marshal replied, 'In that case, there is no time to lose; plant it this afternoon!'
- Emory University announced its strategic aims[20] for the next five years: they have two main areas of focus — digital innovation and special collections. They acknowledge that they are not going to have one of the great general collections in the world, so they see the value in putting their special collections at the forefront of their strategy.
- As Martine de Boisdeffre, President of the European Regional Branch of the International Council on Archives said, 'Users expect to be able to connect the different types of cultural heritage material. To make this possible, organisations need to provide their metadata to Europeana. So many excellent digital resources lie below the surface of the web at present, and aren't easily located by search engines. Europeana will make this material accessible as never before.'[21] We need to be sure that our portals have open, two-way doors. The supply ought to meet the demand on its own turf.

- The Library of Congress worked with flickr to create a Commons[22] to offer a taste of the hidden treasures in the world's public photography archives, and to see how user input and knowledge can help make these collections even richer. With this tactic, minimal description is good enough … (and it could become great). Users add links to biographies or add detail about a building site. The Powerhouse Museum in Australia, the Brooklyn Museum and the Smithsonian have joined the Library of Congress in the flickr commons. Other institutions like Boston Public Library are creating regular flickr photostreams for many of their collections.
- The National and State Libraries Australasia say: 'We are now in a position to explode and reshape our core services, resourcing and infrastructure; to explore radical new approaches across all parts of our work; and to fundamentally shift our libraries to the digital world.[23] Their new outlook includes statements such as:
  - Access is our primary driver.
  - Digital is mainstream.
  - No job will be unchanged.
  - Some things we have always done, we will no longer do.

* * *

None of this suggests that we discourage access to the originals in our collections. Tony Grafton, in a November New Yorker article[24], talks about the two paths that scholars will want to travel, the digital and the paper-based. Special collections need to digitise with an eye to access, so that our materials, so important and necessary for scholarship, will be visible. We must increase the discoverability of more of our collections in the online environment. If more researchers know about them, more people will come to use the original collections in our institutions. We will be ensuring we have the supply to meet the increased demand.

I think the demand is an exciting challenge and I think our supply is fabulous. I would like to see us move quickly to seize the opportunity to connect them.

Notes

[1] Ovenden, Richard. ‘Special Collections in the next 10 years’ presented at the CURL Research Support Task Force meeting, ‘Special Collections: The Way Forward’ at the Wellcome Collection. February 5, 2008.

[2] Digitsation Matters speakers: Susan Chun, Cultural Heritage Consultant (talk presented by Michael Jenkins, Metropolitan Museum of Art); Sam Quigley, Vice President for Collections Management, Imaging & Information Technology/ Museum CIO Art Institute of Chicago; Barbara Taranto, Director, Digital Library Program, New York Public Library; Sharon Farb, Director, Digital Collections Services, UCLA; Bill Landis, Head of Arrangement and Description & Metadata Coordinator, Manuscripts and Archives, Yale University Library; James Eason, Principal Archivist for Pictorial Collections, The Bancroft Library, University of California, Berkeley; James Hastings, Director, Access Programs, NARA [Mr. Hastings was unable to attend the conference, so Ricky Erway, Program Officer, OCLC Programs and Research, spoke in his stead.] For more information (and recordings) see www.oclc.org/programs/events/2007-08-29.htm

[3] Erway, Ricky, and Jennifer Schaffner (2007). Shifting Gears: Gearing Up to Get Into the Flow. Report produced by OCLC Programs and Research. Published online at: <www.oclc.org/programs/publications/reports/2007-02.pdf>.

[4] Dempsey, Lorcan. ‘The Special Web.’ Lorcan Dempsey’s weblog. October 21, 2007. [blog] <orweblog.oclc.org/archives/001461.html>.

[5] Information Architects Japan. Web Trend Map 2008 Beta. http://informationarchitects.jp/web-trend-map-2008-beta/

[6] The University of Texas at Austin, Harry Ransom Center, Using the Ransom Center Collections, <www.hrc.utexas.edu/research/info/>.

[7] Ranger, Joshua. ‘More Bytes, Less Bite: Cutting Corners in Digitization.’ MAC Fall Symposium. October 7, 2006.’ <www.midwestarchives.org/2006_Fall/presentations/Ranger%20Omahapresentationranger.doc>.

[8] The Smithsonian Institution, Smithsonian Archives of American Art, Collections Online, Primary Sources on American Art and Artists. <www.aaa.si.edu/collectionsonline/>.

[9] European Commission, Information Society. i2010: European Digital Libraries Initiative, ‘Europe’s cultural and scientific heritage at a click of a mouse’. <http://ec.europa.eu/information_society/activities/digital_libraries/index_en.htm>.

[10] University of North Carolina, Carolina Digital Library and Archives, CDLA Projects. <cdla.unc.edu/index.html>.

[11] University of North Carolina, A View to Hugh: Processing the Hugh Morton Photographs and Films. [blog] <www.lib.unc.edu/blogs/morton/>.

[12] Gillesse, Robert, Judith Rog, Astrid Verheusen. Alternative File Formats for Storing Masters 2.0.doc. Koninklijke Bibliotheek/National Library of the Netherlands, Research and Development Department. March 7, 2008. <www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/Alternative%20File%20Formats%20for%20Storing%20Masters%202%201.pdf>.

[13] The National Archives, National Historical Publications and Records Commission (NHPRC). Grant Announcement: Digitizing Historical Records. <www.archives.gov/nhprc/announcement/digitizing.html>.

[14] Council on Library and Information Resources (CLIR), Current CLIR Activities, Cataloging Hidden Special Collections and Archives: Building a New Research Environment (2008) <www.clir.org/hiddencollections/index.html>.

[15] The Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). 'Practical Guidelines for the Cultural Heritage Funding Programme. 03/07. <www.dfg.de/forschungsfoerderung/formulare/download/12_151e.pdf>.

[16] Kaufman, Peter B., and Jeff Ubois. 'Good Terms - Improving Commercial-Noncommercial Partnerships for Mass Digitization.' D-Lib Magazine. November/December 2007. <dlib.org/dlib/november07/kaufman/11kaufman.html>.

[17] European Commission Information Society. i2010 European Digital Libraries Initiative. High Level Expert Group on Digital Libraries Sub-group on Public Private Partnerships 'Final Report on Public Private Partnerships for the Digitisation and Online Accessibility of Europe's Cultural Heritage.' May 2008. http://ec.europa.eu/information_society/activities/digital_libraries/doc/hleg/reports/ppp/ppp_final.pdf

[18] The National Archives. 'Draft Nara Digitizing Plan Available For Public Comment.' September 10, 2007. <www.archives.gov/comment/digitizing-plan.html>.

[19] United Kingdom Serials Group (UKSG). Live Serials. 'Mass digitisation of historical records for access and preservation - Dan Jones, Head of Business Development, National Archives' April 2008. <liveserials.blogspot.com/2008/04/mass-digitisation-of-historical-records.html>.

[20] Emory Libraries. An Overview of the Five Year Strategy for the Emory Libraries <web.library.emory.edu/about/publications/Five_Year_Strategy_2007.html>.

[21] Europeana Press Release: These boots were made for...' European Digital Library Foundation. The European Library. Feb 2008. <www.theeuropeanlibrary.org/portal/organisation/press/documents/Europeana_Feb08_press_release_Final.doc>.

[22] flickr. The Commons: Your opportunity to contribute to describing the world's public photo collections. <http://www.flickr.com/commons>.

[23] National & State Libraries Australasia (NSLA). 'The Big Bang: Creating the new library universe.' June 2007. <http://www.nsla.org.au/publications/papers/2007/pdf/NSLA.Discussion-Paper-20070629-The.Big.Bang..creating.the.new.library.universe.pdf >.

[24] Grafton, Anthony. 'Future Reading: Digitization and its discontents.' The New Yorker. November 5, 2007. <http://www.newyorker.com/reporting/2007/11/05/071105fa_fact_grafton>.