From Libraries to 'Libratories'

by LEO WAAIJERS [1]

ABSTRACT

While the eighties of the last century were a time of local automation for libraries and the nineties the decade in which libraries embraced the internet and the WWW, now is the age in which the big search engines and institutional repositories are gaining a firm footing. This heralds a new era in both the evolution of scholarly communication and its agencies themselves, i.e. the libraries.

Until now libraries and publishers have developed a digital variant of existing processes and products, i.e. catalogues posted on the Web, scanned copies of articles, e-mail notification about acquisitions or expired lending periods, or traditional journals in a digital jacket. However, the new OAI repositories and services based upon them have given rise to entirely new processes and products, libraries transforming themselves into partners in setting up virtual learning environments, building an institution's digital showcase, maintaining academics' personal websites, designing refereed portals and – further into the future – taking part in organising virtual research environments or collaboratories. Libraries are set to metamorphose into 'libratories', an imaginary word to express their combined functions of library, repository and collaboratory. In such environments scholarly communication will be liberated from its current copyright bridle while its coverage will be both broader - including primary data, audiovisuals and dynamic models - and deeper, with cross-disciplinary analyses of methodologies and applications of instruments. Universities will make it compulsory to store in their institutional repositories the results of research conducted within their walls for purposes of academic reporting, review committees, and other modes of clarification and explanation. Big search engines will provide access to this profusion of information and organise its mass customisation.

INTRODUCTION

The digital library is nowadays taken for granted. Indeed, they produce catalogues with the aid of a keyboard and then post them on the Web; they supply scanned copies of articles as attachments and send out e-mails about acquisitions or expired lending periods. For their part publishers issue journals in digital jackets and facilitate editorial and refereeing processes using workflow applications. So, yes, libraries and publishers have been digitised. That is, they have digitised their centuries-old paper-based processes, taking twenty-five years or an entire generation to do so.

And then, out of the blue, emerged the question of the rationale of library catalogues in the age of full-text documents and powerful search and presentation engines. There has also been debate about the need for journals when the same search engines automatically produce the citation indices of articles, which although not uncontroversial are nevertheless a broadly accepted measure of their quality. And even outsiders prophesy the obsolescence of document supply in the open-knowledge environments now under construction. The centuries old processes themselves are suddenly being questioned: will they be able to satisfy the future needs of their users and financiers?

Any attempt to answer this question requires some insight into the needs referred to. Is such insight at hand? On the detailed level of a blueprint it is not, as the current situation is still too turbulent, but on the level of trends there is a good deal of literature available and any number of investigations are being carried out. A convincing and comprehensive study is Michael Nentwich's *Cyberscience*, published two years ago (Nentwich, 2003).

Education is evolving towards e-learning, i.e. the world of associative non-linear learning, which is highly interactive both at the personal and community levels and at the same time visually oriented and information-dense. We are talking here about Virtual Learning Environments or Learning Management Systems, the counterpart to which is Virtual Research Environments. In these environments, sometimes called collaboratories, distant researchers share and enhance datasets or text corpora, models and theories. Tony Hey is one of the proponents of this development, giving us an insight into the giant data streams emerging along with it (Hey & Trefethen, 2002).

Although we may not know exactly what the future information needs will be of the academic community, i.e. students, teachers and researchers, to me one thing is certain: open access to state-of-the-art knowledge is crucial in order for both research and learning environments to succeed. Limited access, be it the result of either technical or juridical implications, impedes solid growth in the human knowledge base. Put another way, there is no point going to great pains to overcome the technological obstacles facing ICT only to come up against the legal copyright barriers. An interesting example here is the Elsevier content stored in the e-Depot of our National Library: the costly technological infrastructure required for guaranteeing long-term access to this material is renowned. But in order to enjoy this access one has to travel to the library in The Hague and then possibly stand in a queue, as only one person at a time is allowed access; again a replica of the situation in the paper era.

I would like to go back to the question whether libraries and publishers will be able to meet the future needs of their users and financiers. Put this way, the question raises a problem for publishers arising from the bundling of user and financier needs. In my opinion such needs are conflicting: users want open access for their learning and research processes while financiers, who are the shareholders' representatives, want exclusive and highly prized products. The academic community is all too aware of which party has for several decades been on the winning side of this battle.

And libraries; are they able to meet academic needs? "Will there be a need for library services beyond licence management?" is the question posed by the LIBER conference announcement. The question seems remarkable since no one observing current library trends and activities can possibly overlook the mushrooming of repositories in the global library community. To date, the OAIster registry [2] lists over 500 academic repositories, with a new one being added every working day. Together these repositories contain 6 million digital objects. Indeed, these include more than a few duplicates while many are images but, on the other hand, the first research datasets are also emerging in OAIster. Just three years ago none of this existed. Since all these repositories comply with the OAI Protocol for Metadata Harvesting (OAI-PMH) they together form an interoperable global knowledge grid that has enormous potential.

REPOSITORIES

Listening to the reasons for setting up repositories within their own institutions [3], libraries have unanimously argued that repositories offer better long-term digital curation than that provided by authors' own laptops, that storage of material in repositories lays the foundation for its reuse for educational purposes, that repositories could make research results available much faster than dissemination through traditional channels, that the accurate time stamping of publications stored in repositories provides a solid basis for laying priority claims, and that institutional repositories currently offer the only opportunity for storing compound documents i.e. publications that include primary research data, images, models or simulations in a retrievable way.

It seems that libraries are supported by their financiers. The growing list of signatories to the Berlin Declaration [4] is a good indicator, while the forthcoming Research Council's UK draft position statement is another. The minutes of a recent EC workshop discussing the hundreds of millions of Euros budgeted for the 7th Research Framework Programme (FP7) say, "Looking to the future, the deployment of Digital Repositories is likely to become far more pervasive throughout Europe and the size of the holdings is likely to become more inclusive. Hence, although there is a considerable way to go at both the institutional and national levels, it seems essential to plan now for a pan-European infrastructure within the time-frame of FP7." An article in the April 2005 issue of *Science* reads, "While moves in the United States to make scientific research results available – for free – at the click of a mouse have generated intense debate, European research organizations have quietly been forging ahead. Slowly but surely, they are starting to build and connect institutional and even nationwide public archives that will, according to proponents, be the mega-libraries of the future, allowing anyone with an Internet connection to access papers produced by publicly funded research." (Vogel & Enserink, 2005).

Therefore, I would extend the original conference question at least as follows: "Will there be a need for library services beyond licence AND repository management?" Even if the answer to this is no, the outlook for libraries is still exciting, though not in terms of licences. It is my view (and I recognize there are others) that a licence is simply an act of surrender by libraries that has to be renewed every 3 to 5 years. And what is politely referred to as licence negotiations is merely a euphemism for the begging of favours. No, the truly exciting part is repository management. This is where a whole new world is opening up. Selecting an open-source OAI application [5] and installing it on a server is only the beginning: stocking an institution's repository with research output produced by that institution is really what it's all about. Librarians have to impress upon their university managers that the institutions responsibility

does not stop with furthering the creation of new knowledge, but also includes communicating it. In my perception, universities have neglected this latter responsibility for too long, leaving it to individual researchers. Ultimately, this has led to the serials crises. Now, for the first time ever, detailed texts on scholarly communication are appearing in universities' policy documents. This means that institutions have to define strategies that address the issues of copyright, quality and secrecy. Coriyne McSherry's book *Who owns academic work?* may become mandatory reading for institutional managers. It does not answer but rather raises questions such as: how far may authors go in giving away copyrights to their publications and data; does material in institutional repositories have to meet certain quality standards and what results may or must be kept secret? (McSherry, 2001)

Aside from the strategic component, there are practical issues attached to repositories. Every repository starts with Dublin Core as its metadata standard. [6] But sooner or later you run into the limitations of this standard. Long-term preservation requires technical metadata, while compound documents or coherent clusters of text, data and images require metadata that reflect the structure of documents. For any document this metadata should contain information about its status, version, provenance and usage rights. Our current Dublin Core standard is far too primitive for containing this wealth of information.

In addition, every repository starts as an entity on its own, but it must obviously be embedded in the institutional, national and international infrastructures as well. This necessitates transparent relations with adjacent applications. Storage must be the imperceptible effect of registering publications for an annual institutional report, an overview supplied to a visiting committee, or the Research Assessment Exercise. Long-term preservation is achieved by automatically forwarding publications to a national library's e-depot. When harvesting publications, object and author identification need to be in place to avoid duplications or other annoying irregularities.

This means there is a good deal of interesting work to be done by libraries. And perhaps other bodies besides. And once an institution has its repository in place, certification can be requested from DINI, the German Initiative for Networked Information or, possibly, from the Research Library Group (RLG) in the US.[7]

Finally, in addition to management and repository practicalities, there is the question of authors themselves. After all, it is their intellectual products we are talking about. Authors are under increasing pressure, particularly by funding agencies, to place their publications in repositories. Take the following quote from a press release earlier this year: "The eight UK Research Councils, under the umbrella of Research Councils UK (RCUK), have proposed to make it mandatory for research papers arising from Council-funded work to be deposited in openly available repositories at the earliest opportunity." [8]

Authors have to be convinced that depositing is in their own interest. In doing so it is most important to demystify the issue. For example, they need to be told that current research shows that open access publishing increases the number of citations and hence impact factors, that the project RoMEO site proves that publishers are gradually giving in on copyrights, that experiences of authors, who formulate their own copyright statements, teach us that publishers accept them, that parallel publishing on the Internet stimulates sales of a book's paper version - and so on and so forth. A project like the Netherlands' Cream of Science demonstrates that it is possible to overcome the hurdles and make top authors, even Nobel Prize winners, enthusiastic about placing their work in repositories (Feijen & Kuil, 2005). It has also shown that so-called objections sometimes amount to no more than librarians' perceptions of author viewpoints. And that it is occasionally impossible to publish the complete oeuvre of an author simply because his or her publications have become lost. This in itself constitutes another powerful argument for depositing materials in repositories.

Although the establishment and maintenance of a harvestable and well stocked repository is a valuable and exciting job for its own sake, it is certainly not the end of the matter.

SERVICES

Fundamental to the OAI protocol is its stratification in a data layer and a services layer. Once a repository has established a firm data layer, the issue of services needs to be examined. Where the data layer is an infrastructure, established in the public domain and operating on the supply side, services are developed in response to a demand. Any player - commercial, public, community or individual - can start a service. And such a service is subject only to the limits of your imagination. In practice you obviously have to accept the limitations of technology, money and

human resources, and in that order too. I mean that technology is the easiest aspect to tackle and people the most difficult. So, to achieve success, you should approach the task in the opposite order, starting first with people, moving on to money, and finally tackling the technology issue.

The most basic service is simply to allow a number of repositories to be harvested, getting a search engine to order the yield and present it to the world. This is what Scirus, Yahoo and several other search engines do, although they trawl the Web in addition to drawing on institutional repositories. Most interesting in this regard are the Websites of individual scientists, as more and more authors post the officially published versions of their articles on their own sites. DAREnet in the Netherlands offers the same type of service nationwide. [9] That is, DAREnet offers the openly accessible content of all Dutch academic repositories. DAREnet now contains 50.000 publications from the country's 13 universities, the Royal Netherlands Academy of Arts and Sciences (KNAW), and the Netherlands Organization for Scientific Research (NWO).

Services like Google Scholar and Scopus have added a new dimension to this service by giving the citation index of each article as well. And since Google and Elsevier use completely different business models, Google Scholar is able to provide its services for free to the end user, while Scopus is very expensive. Two effects of these citation-enhanced search engines come to mind: firstly, suppose you are writing an article and you need some additional information. You then use an academic search engine and get a list of potential candidate articles. You browse a number of abstracts and you conclude that two articles may really give what you need. In one case a click on the word 'full text' does what it promises, delivering you the full text. In the other case, however, the click produces an order form that asks for your credit card number. In all likelihood you will use the first article. This article will then be cited in your own one and thus rise one step on the citation ladder. This mechanism means that openly accessible articles will gradually drive down toll-gated ones. And then there are the effects of citation-enhanced search engines on journals. The function of a journal is to bundle articles per subject, time stamp them, give access to the full text and render prestige via citations. But this is exactly what citation-enhanced search engines do, but they do the job even more accurately as they give the exact number of citations per article, where journals only attribute a so-called impact factor to an article, which is an average of the citation numbers per article in the journal. Therefore the advent of citation-enhanced search engines means that the added value of journals is being seriously questioned. Added to this is the fact that journals are slow and costly vehicles of knowledge.

In general, the new academic search engines materialise a form of mass customization of knowledge. In the future the application of emerging semantic web techniques may further improve the precision of their search results.

PROFESSIONAL TO PROFESSIONAL

Not everybody is satisfied with a daily portion of Google. Professionals in various fields need more. Here lies the basis for professional-to-professional services. A few observations:

What do teachers and students need? The answer is multimedia content in their Virtual Learning Environments, in such a way that they can reuse this content in different circumstances and exchange it with colleagues while avoiding being vendor-locked by their Blackboards or WebCTs. To meet these requirements, content should be both granulated and highly structured. As a consequence, complex metadata must be applied such as the new standard DIDL or Digital Item Declaration Language; the Dublin Core format must be replaced by more informative ones, such as IEEE LOM. [10] In short, there is room for professional services that go far beyond what the Googles of this world have to offer.

What do researchers need? To date huge data files have been created as the outcome of observations and measurements, through the scanning of giant text corpora or as a result of extensive inquiries over a long period. This data needs to be analysed, used for testing theories or models and augmented with new data. The Human Genome Project is an inspiring example of such a new research approach, referred to as a Virtual Research Environment or Collaboratory. Here the requirements are sufficient bandwidth to transport data, long-term preservation and accessibility of big data sets and seamless workflows between researchers, to list but a few. Again, ample opportunity for professional services.

What do politicians and managers need? They want the world to know what important and elegant research results are the outcome of (public) monies invested. They want to profile their country, university or institution. Their imaginations may go in the direction of windows that display not only the cream of science but its entire production,

not merely as a compilation that can be searched but rather one that can be enhanced with fingerprints of the expertise of the authors and their institutes and fleshed out with citations and information about relevant awards. Wouldn't this be wonderful? Here, the service required may be a sophisticated version of what Google Scholar already offers. Nevertheless, we are talking about a customized professional service.
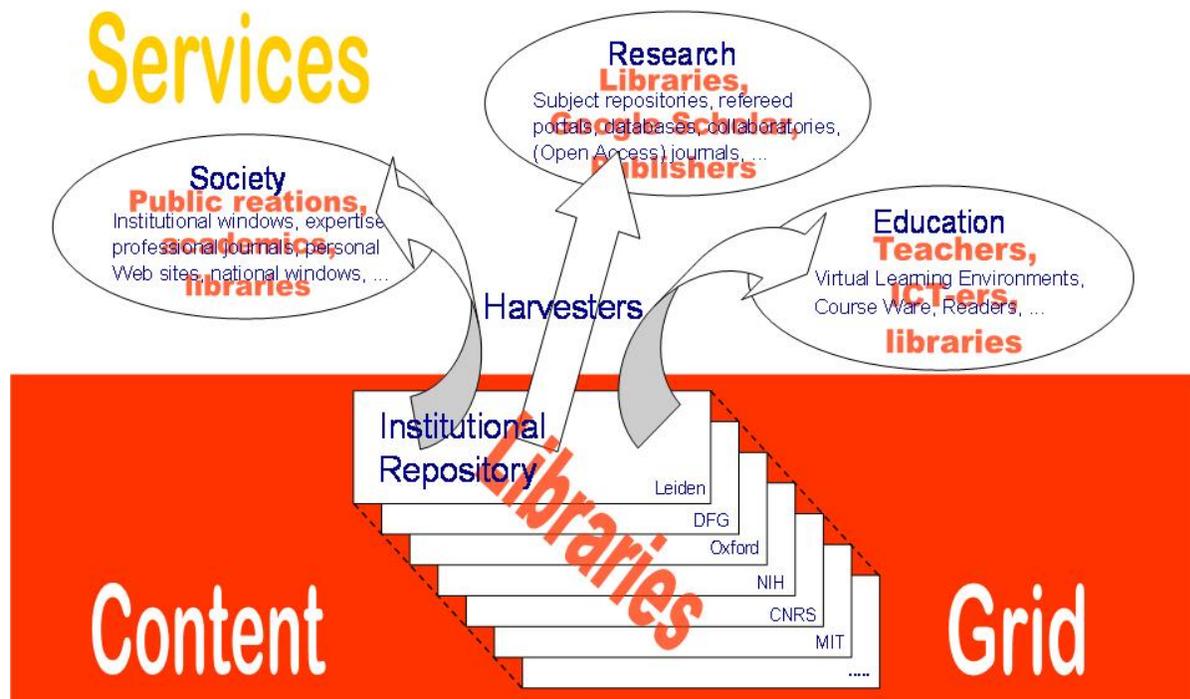
As I have already said, repository-based services are limited only by the bounds of your imagination. Other examples of services now emerging are personal news feeds, refereed portals and overlay journals where universities themselves organize quality control by setting up editorial boards and networks of reviewers, thus throwing off the yoke of publishers' monopolies, or the construction of knowledge bridges between the academic content of repositories and the demand for innovation in society.

CONCLUSION

Going back to the central conference question - "Will there be a need for library services beyond licence management? - my answer certainly is yes as far as repositories are concerned. The world's libraries have grasped this. And if they had not done so pro-actively, they would have been told to do so. Institutions need repositories and someone has to manage them. That's all there is to it.

More thrilling, however, are the possibilities opening up for services. No doubt there is a growing need for a wide number of content services. Some major commercial players, such as Elsevier, Google, Yahoo and others, have already gained a foothold in this market. Happily, this time there is a market. That is to say it has not been monopolised, or not yet at any rate. For the time being the parties involved are concentrating on the mass customisation of academic knowledge. This means there is still room for other players in the field of professional-to-

professional services. Will libraries step in? Have universities learned their lesson from the past, when they left scholarly communication to third parties and continue to suffer the consequences even today? A repetition of this historical mistake may usher in a world in which not only publications but also data, models and learning content are monopolised. If the only reason for action would be to avoid such a situation, it would be sufficient in itself. But the world of repository-based services is also an exciting one, in which suppliers must interact intensively with researchers, teachers and managers alike. Libraries taking part in the process will undergo a metamorphosis: from paper-based thinking to the digital paradigm, from importers of global knowledge to exporters of local knowledge, from suppliers of a visible collection to invisible partners in academic processes, from libraries to 'libratories', my concoction to express the combined function of libraries, repositories and collaboratories. So, the final question is: "What could libraries put an end to?".

But I will leave this question for the conference to answer. If you feel this is unsatisfactory, bear in mind that that is exactly what your clients, i.e. scientists, always do: replace one question with another and then leave the stage. Thank you.

NOTES

1. This article was the keynote speech at the LIBER 34th Annual Conference "Strategic choices: current thinking", 5 July 2005, in Groningen, the Netherlands. It is also published in First Monday, 10(2005)2. http://firstmonday.org/issues/issue10_12/waaijers/index.html

2. OAIster is a project of the University of Michigan Digital Library Production Service. Its goal is to create a collection of freely available, previously difficult-to-access, academically-oriented digital resources that are easily searchable by anyone. See: http://oaister.umdl.umich.edu/o/oaister/

3. For a state of the art overview see: Gerard van Westrienen and Clifford Lynch: "Academic Institutional Repositories. Deployment Status in 13 Nations as of Mid 2005", *D-Lib Magazine*, 11(2005)9. http://www.dlib.org/dlib/september05/westrienen/09westrienen.html

4. "Our mission of disseminating knowledge is only half complete if the information is not made widely and readily available to society. New possibilities of knowledge dissemination not only through the classical form but also and increasingly through the open access paradigm via the Internet have to be supported. We define open access as a comprehensive source of human knowledge and cultural heritage that has been approved by the scientific community." See: http://www.zim.mpg.de/openaccess-berlin/berlin_declaration.pdf

5. "The OAI-Protocol for Metadata Harvesting (OAI-PMH) defines a mechanism for harvesting records containing metadata from repositories. The OAI-PMH gives a simple technical option for data providers to make their metadata available to services, based on the open standards HTTP (Hypertext Transport Protocol) and XML (Extensible Markup Language). The metadata that is harvested may be in any format that is agreed by a community (or by any discrete set of data and service providers), although unqualified Dublin Core is specified to provide a basic level of interoperability. Thus, metadata from many sources can be gathered together in one database, and services can be provided based on this centrally harvested, or "aggregated" data. The link between this metadata and the related content is not defined by the OAI protocol. It is important to realise that OAI-PMH does not provide a search across this data, it simply makes it possible to bring the data together in one place. In order to provide services, the harvesting approach must be combined with other mechanisms." Source: "OAI for Beginners - the Open Archives Forum online tutorial" at http://www.oaforum.org/tutorial/.

6. "The Dublin Core is a metadata standard for describing digital objects (including webpages) to enhance visibility, accessibility and interoperability, often encoded in XML. It was so named because the first meeting of metadata and web specialists which saw its birth was held in the town of Dublin, Ohio in the United States." Source: http://en.wikipedia.org/wiki/Dublin_core

7. A draft RLG checklist for certifying digital repositories is currently under construction

8. Research Councils UK, "RCUK Announces Proposed Position on Access to Research outputs," news release, 28 June 2005, at http://www.rcuk.ac.uk/press/20050628openaccess.asp.

9. See note 6.

10. For a comparison of Dublin Core and IEEE LOM, see: http://www.ischool.washington.edu/sasutton/IEEE1484.html

REFERENCES

Feijen, Martin and Annemiek van der Kuil: "A Recipe for Cream of Science: Special Content Recruitment for Dutch Institutional Repositories," *Ariadne,* (October 2005)45. http://www.ariadne.ac.uk/issue45/vanderkuil/.

Hey, Tony and Anne Trefethen. *The Date Deluge: An e-Science Perspective*. UK e-Science Core Programme, 2002. http://www.rcuk.ac.uk/escience/documents/report_datadeluge.pdf

McSherry, Corynne: *Who owns academic work? : battling for control of intellectual property*. Cambridge, MA [etc.] : Harvard University Press, 2001, 275 p.

Nentwich, Michael, Cyberscience : research in the age of the Internet. Vienna : Austrian Academy of Sciences Press, 2003.

Vogel, Gretchen and Martin Enserink: "Europe Steps Into the Open With Plans for Electronic Archives". *Science*, 308(29 April 2005)5722, 623-624.

WEB SITES REFERRED TO IN THE TEXT

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. http://www.zim.mpg.de/openaccess-berlin/berlin_declaration.pdf

Cream of Science. http://www.darenet.nl/en/page/language.view/keur.page

DAREnet. http://www.darenet.nl/en/page/language.view/home

DIDL - Digital Item Declaration Language. http://xml.coverpages.org/mpeg21-didl.html

DINI − die Deutsche Initiative für Netzwerkinformation. http://www.dini.de/dini/zertifikat/

 FP7 - Seventh Framework Programme. http://www.cordis.lu/fp7/home.html

Google Scholar. http://scholar.google.com/

Human Genome Project. http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

LIBER 34th Annual Conference. http://liber.ub.rug.nl/

OAIster. http://oaister.umdl.umich.edu/o/oaister/

OAI-PMH - The Open Archives Initiative Protocol for Metadata Harvesting. http://www.openarchives.org/OAI/openarchivesprotocol.html

RCUK - Research Councils UK. http://www.rcuk.ac.uk/

RLG - Research Library Group. http://www.rlg.org/

RoMEO - Rights MEtadata for Open archiving. http://www.lboro.ac.uk/departments/ls/disresearch/romeo/

Scopus. http://www.scirus.com/srsapp/