

The institutional repository in the digital library

We begin by looking at the concept of institutional repositories within the broader context of digital libraries. ‘Digital libraries’ can mean many things, but we consider them to be libraries first and foremost, and built upon the enduring principles of information management which have lain at the heart of the practice of librarianship for hundreds of years. We look also at the significance of the qualification which defines the scope of this book – the institutional repository. Libraries are themselves repositories, and have always dealt in the management of repositories for their users. With libraries now routinely managing repositories of various types in digital format, what does it mean to qualify ‘repository’ with ‘institutional’?

We examine the particular value of institutional repositories to research material, and look also at the other types of material for which institutional repositories are increasingly being used. There are considerable implications for librarians in managing digital material as full-text, where the digital item is the item being curated and managed over time, rather than a printed item with some digital metadata. The institutional repository movement has played a large part in making librarians face up to these implications in their entirety, and one of the first to be encountered is the question of metadata. What needs to be described for a digital object to be made findable in the present and into the future, when the environment which sustains and creates it may change and change again many times?

We complete this first chapter with a consideration of the real motivation behind the dissemination of research on the Web – research impact. Studies are beginning to show just how much more impact research can have when it is ‘unlocked’ from commercial journals and made available for everyone to find on the Web.

Digital libraries in a digital world

In the past ten years, the concept of the 'digital library' (or the 'electronic library') has been increasingly used, and now crops up relentlessly in the professional literature. This is not surprising, as the combination of low-cost computing and high-speed networking now affects all areas of life in the developed world. 'Digital banking', 'online shopping' and 'digital television' are transforming the ways in which we transact our daily business and consume entertainment. We also book holidays online, gamble on the Internet and conduct hundreds of other activities online. Increasing numbers of people work from home, using telecommunications to recreate their office environments in virtual space. As content goes online, and the means of access to it becomes as available and familiar as clicking on the television set, so it is a natural expectation that libraries too will join the interconnected web world.

Librarians are, however, well aware that there are also dangers surrounding the concept. It is often stated that the World Wide Web, or the Internet itself, is one huge electronic library. This is only true in the most general sense that it requires navigation aids in order to discover particular content. In fact, the Internet is no more a library than is a city or a country. Of course the Web contains masses of documents of all types, and in that sense it is like a library – but all libraries – even 'universal libraries' such as the Library of Congress – are based on selections. The Library of Congress's website admits that it does not collect everything, and nor would it want to: 'The Library's collections are based on the Jeffersonian ideal that all subjects will be of interest and value to Congress and, by extension, to the scholar and researcher'.

On the basis that it reflects the culture of a nation, universal libraries sometimes collect material which it is hard to imagine being of interest to scholars and researchers. The National Library of Australia, for example, reported in 2002 that it was now harvesting pornography published on the Australian web domain, for the use of researchers (BBC, 2002). It had not at that stage developed a policy on how to allow access to the material, however. It was also careful to confirm that it would be collecting only legal pornography. The Internet, as is well-known, contains both legal and illegal material.

Ross Atkinson emphasises the key library activity of selection:

The network is not a digital library. We cannot sit back and imagine that what is on the network is in the digital library ... A

library, digital or otherwise, is always a highly selective subset of available information objects, segregated and favored, to which access is enhanced and to which the attention of client-users is drawn in opposition to objects excluded. (Atkinson 1996)

Definition of a 'digital library'

So it is false to think of the Internet as itself a digital library. As we, therefore, have to refine the concept if we wish to talk about digital libraries, how do we do so? Writing in *Library Journal*, Cloonan and Dove (2005) reminded readers of the 'Five laws of library science' expounded by the great Indian librarian Shiyali Ramamrita Ranganathan, in 1931. To this day, many librarians accept Ranganathan's five laws as a perfect conceptual summary of the aims of librarianship. Written in a period before gender-neutral language was expected, the five laws are:

1. Books are for use.
2. Every book its reader.
3. Every reader his book.
4. Save the time of the reader.
5. A library is a growing organism.

Cloonan and Dove then look at Google, the world's most popular search engine and, therefore, the most obvious candidate to be a universal catalogue of the Internet, and ask whether it meets the test of Ranganathan's Third Law. Does Google find, from the mass of digitised documents which exist on the Internet, not only the specific item for which a user may be searching, but also related items which they may want to consult without previously knowing about them as they begin to search? In the days of print librarianship, the Third Law was met by the use of robust cataloguing and indexing standards, including cross-referencing from within catalogue records. Librarians have yet to replicate these standards and their reach in the digital world. The authors conclude:

Most information seekers using Google never go past the first page of results. Google's criteria for what goes on that first page are popularity and payment for placement. It is unlikely Google will

change that. Library resources should match Google's ease of use but not its criteria for first page listing. Library tools must exhibit all the qualities of what Ranganathan calls a 'well-arranged collection'. (Cloonan and Dove, 2005)

It is tempting to use the term 'digital library' about any collection of digital objects which have some means of navigation and retrieval, but approaching the question using Ranganathan's Laws would suggest that a collection of items is not a digital library merely by virtue of the items being digital. Rather, a digital library is such by virtue of its being a library first and foremost. An academic or research library is organised for use in the pursuit of human advancement. The fact that its contents – or a large proportion of them – are digital is merely an accident of history. Digital libraries are, therefore, much more than aggregations of documents on the World Wide Web, whose navigability and discovery services can be left to commercial companies interested in maximising income from advertisers with product information which can be made particularly eye-catching. A true digital library has to be organised for its purpose, and must not be randomly heterogeneous and indexed as a commercial by-product. Nor should it be simply a desperate response to digital deluge and budgetary inadequacy, as Robin Alston suggests:

If a librarian found juggling resources difficult in 1980, when the first storm clouds began to appear, by 1990 not even those who approached budgeting with imagination could balance the books. The concept of the digital library was born in desperation. (Alston, 2002)

Michael Lesk, however, in his authoritative book on the subject, disagrees:

The answer should not be despair but organization. A digital library, a collection of information that is both digitized and organized, gives us powers we never had with traditional libraries. (Lesk, 1997)

Yet even today there is still no common consensus to define a 'digital library'. One reason for this is that the term was adopted by the computing science community while librarians were still talking about 'electronic libraries' and 'hybrid libraries'. The National Science Foundation's 'Digital Libraries Initiative', launched in 1996, funded six

projects, one each in environmental planning and geographic information systems, spatially-referenced map information, digital video creation, federated repositories of scientific literature, intelligent agents and ‘interoperation mechanisms among heterogeneous services’. Its website gives a description of digital libraries which puts their creation firmly in the hands of software engineers:

Digital libraries basically store materials in electronic format and manipulate large collections of those materials effectively. Research into digital libraries is research into network information systems, concentrating on how to develop the necessary infrastructure to effectively mass-manipulate the information on the Net. The key technological issues are how to search and display desired selections from and across large collections.

Bearing out this definition, the annual Joint Conference on Digital Libraries (JCDL), which has been running since 2001, is a collaboration between the Association for Computing Machinery (ACM) and the Institute of Electrical and Electronic Engineers (IEEE). However, in the years since it began, the conference has gradually become more accommodating to librarians and information scientists. For example, the JCDL 2004 conference catered for an audience which was a mix of librarians and computer scientists, with the latter group still probably outnumbering the former on the basis of the papers presented. Its themes plainly sought to encourage attendance from both communities, however, with titles such as ‘Educational aspects of digital libraries’, on the one hand, and ‘Mining and disambiguating names’ on the other.

What happened, then, during the few years which saw the first burst of energy associated with this new idea of ‘digital libraries’, was that research was done by computer scientists in order to provide solutions to the problem of putting research-quality digital content on the Web, with sufficient functionality to replace its normal format as print, or images, or laboratory instrumentation, together with some innovative new functionality never previously available. The collections of content which formed the testbeds for this research effort were, in effect, subject-based digital libraries. Once a collection of high quality content had been mounted on the Web – the maps of the Alexandria Digital Library, for example – they were there for all to use. In time, logically, it seemed that all knowledge domains could be represented by digital libraries, making institutional libraries redundant in the process.

Repository locus: institution vs discipline

When we talk about the ‘institutional repository’, we use ‘institution’ to refer the educational or research establishment which is the library’s parent body. Institutional repositories have emerged from universities, but are spreading into other types of educational organisation too, such as colleges and research institutes. However, research repositories were until quite recently based only around disciplines. The first and still best-known disciplinary repository to emerge was arXiv (www.arxiv.org), a repository of research papers in particle and high-energy physics, based originally at the Los Alamos National Laboratory. arXiv has been running for some 15 years, and is widely used by physicists across the world. Over time, it has expanded its coverage to embrace associated disciplines such as mathematics and computer science, and it has also changed its physical location, moving to Cornell University Library a few years ago.

While arXiv has been successful in capturing the market for deposited e-prints in these particular domains, it has been somewhat surprising to observe that it has not served as a model for many others. Economics has been partly successful, with EconPapers (working papers in economics – see www.econpapers.repec.org), and the literature of cognitive psychology is captured in e-print form in CogPrints (www.cogprints.org). In medicine, the PubMed Central service is somewhat different in that it provides digital copies of papers only after their publication in printed journals. As the e-print movement gathered pace in the late 1990s, promoted tirelessly by evangelists such as Professor Stevan Harnad of Southampton, who was inspired by the example of arXiv to call for it to be replicated across all subject disciplines, it became clear that methods of working by researchers in different disciplines were themselves sufficiently different that we could not cover the entire world of research by means of the physicists’ model.

There are significant differences in the ways in which academic and researchers work in different domains. For the purposes of managers of institutional repositories, the most significant relate to the place of peer review – the process of ‘refereeing’ by which research is validated by peer experts, or referees, and thereby permitted to enter the discourse and body of knowledge of a particular knowledge domain. What all domains hold in common is their need for peer review, if only to allow the researchers to point to citations in reputable journals in order to prove their credibility in the field, and to enhance their career prospects. Physicists, however, tend to be happy to have their research papers

circulated widely while the research described in them is still unvalidated, and, therefore, tentative. This form of paper, known in pre-digital times as the pre-print, is normally eventually replaced with a refereed version. Prior to the Web, scientists would send copies of pre-prints to fellow researchers working in institutions across the world, or distribute them at conferences, and thereby seek early feedback. Physicists, particularly in high-energy and particle physics, work at rapid speed, and are not content to wait for official peer review by journals to validate their ideas. In arXiv, many papers are submitted initially as pre-prints, and later replaced with 'post-prints' (refereed versions) which come complete with citation details to the published journal in which they appear. The substitution is not always applied, however. Some physicists are happy to let unrefereed papers remain in the repository, or to add refereed versions rather than replace one with the other.

In other disciplines, pre-prints are scarcely used (perhaps only to a small and very select group of peers), and research is carefully guarded until after refereeing, when the researcher, satisfied to have their work validated, will release it to the world in the form of journal publication. Not surprisingly, the fields of medicine and life sciences research behave in this way. The consequences of unverified medical hypotheses leaking into the public domain and creating hysterical stories in the press can obviously be very serious for a researcher. In these post-print oriented domains, repositories are still very useful, if only because the paper, once refereed, can then be placed immediately into a repository and made findable on the Web. Journal publication, even in electronic form, has an associated time-lag between acceptance and publication which can be many months, and researchers want their work to appear as early as possible.

It was in response to the growing awareness of the importance of arXiv that the institutional repository movement was created, and along with it the Open Archives Initiative (OAI), beginning with a conference in Santa Fe, New Mexico, in 1999. The rationale for this was that if academics working within disciplinary boundaries did not feel motivated to deposit copies of their pre- or post-refereed articles in such disciplinary archives as existed for them (if any), then their institutions could provide facilities to make the process easy, and might indeed eventually require the population of institutional repositories as a contractual obligation.

The consequence of the OAI has been the appearance across the globe of many hundreds of institutional archives, alongside the disciplinary archives which continue to exist. This has presented a dilemma for some

researchers who – where the choice exists of both types of repository – have not been sure which one they should choose in which to deposit their paper. Repository managers have urged them not to worry, and to deposit in both, as the search services being developed can cope with redundancy and apply de-duplication of identical results in some cases. In this way they have brushed aside the legitimate confusion of researchers due to the urgency of the need they perceive to prove the open access publication paradigm change by means of capturing content on a large scale. Yet the confusion persists, and academics will rightly ask why, in a system designed to maximise efficiency, they should be asked to negotiate two separate submission interfaces in order to deposit a single paper. Submission interface design needs, therefore, to advance in order to cope with this dilemma, and this is one of the research and development challenges still facing the open access movement at the present time.

Repositories and digital libraries

Academic libraries today are increasingly involved with the digital library agenda represented at conferences such as JCDL because they see that there is a need to develop institutional digital libraries alongside subject-based digital libraries. The institutional library needs a presence on the Web – a place to describe its print and web-based services, and to bring together the content it makes available to its users. It needs to present its catalogue but also its other finding aids – to its collections of e-journals, its collections of digitised materials from its treasures, and other lists which are most usefully presented separately, such as electronic reserve texts or past exam papers. Institutional libraries also are growing the range of services they can offer via their website. Examples of these include interlibrary loan request – sometimes by electronic full-text delivery; requests to retrieve store items; book loan renewals and electronic reference support. In addition, library services need to be distributed out to other useful environments, such as student virtual learning environments and university portal sites. They need to be ‘skinned’ in various ways, and to be capable of being searched in an aggregated and in a user-defined sub-aggregated fashion. Some of the technology involved in providing these apparently obvious functional enhancements is astonishingly complex and difficult (such as federated searching across a heterogeneous commercially published database environment).

Digital libraries, then, belong both to knowledge domains and to institutions, in the same way as do repositories, which are constituent elements of each. Table 1.1 breaks down both digital libraries and repositories by institution and discipline. The libraries, on the left, depend more and more on the repositories, on the right, to provide them with the selections of collections they present as libraries, whether institutional or disciplinary.

There has been a great deal of experimentation and research into digital library developments across the globe in the past 15 years, and these days, as some of the experimentation and research begins to blossom into genuinely new and important services, academic libraries are employing a higher proportion of staff with IT experience and qualifications. This is particularly true in larger libraries. The consequence of this is that library managers need to understand the technologies of digital libraries, at least conceptually, and so plan for

Table 1.1 Examples of digital libraries and repositories

Digital libraries	Repositories
Disciplinary	
Alexandria Digital Library <i>www.alexandria.ucsb.edu</i>	arXiv <i>www.arxiv.org</i>
Perseus Project <i>www.perseus.tufts.edu/</i>	PubMed Central <i>www.pubmedcentral.nih.gov</i>
Digital Scriptorium <i>sunsite.berkeley.edu/Scriptorium/</i>	EconPapers <i>econpapers.repec.org/</i>
Center for Electronic Texts in the Humanities <i>www.ceth.rutgers.edu/</i>	CogPrints <i>cogprints.org</i>
Institutional	
California Digital Library <i>www.cdlib.org/</i>	Edinburgh Research Archive <i>www.era.lib.ed.ac.uk</i>
Illinois Digital Academic Library <i>www.idal.illinois.edu/</i>	DSpace at MIT <i>dspace.mit.edu/index.jsp</i>
	Nottingham Eprints <i>eprints.nottingham.ac.uk/</i>
	LSE Research Articles Online <i>eprints.lse.ac.uk/</i>

their future development along a trajectory which, for some, might see their goal as being transformed in their entirety into digital libraries, but for all must mean that over time their services are converted gradually to a basis in digital library technologies.

While for many this may seem like a threatening picture (James Thompson terrified librarians as far back as 1982 when he published a book called *The End of Libraries*), in fact libraries have always developed by importing technologies from elsewhere. Cataloguers and classifiers have relied for decades upon technologies developed by bibliographers, documentalists (an American term, referring to professionals who had ‘the delegated task of creating access for scholars to the topical contents of documents, especially of parts within printed documents and without limitation to particular collections’ (Buckland, 1997)), information scientists and, of course, more recently by computer scientists. These technologies have produced standard tools such as the Dewey Decimal Classification (DDC) system, Library of Congress Subject Headings (LCSH), Anglo-American Cataloguing Rules, 2nd edition (AACR2) and the Machine Readable Cataloguing (MARC) manual.

Perhaps the difference with the digital library technologies now being developed is that there is no automatic assumption that they will be handed to librarians to use, or at least, perhaps only to certain types of librarian – many of whom might not be expected to work inside campus library buildings. The digital library, in such a context, becomes something which takes some shape from the traditional library, and some shape from other sectors and disciplines. Several writers have likened the concept of the ‘digital library’ to that of the ‘horseless carriage’, a term which was a clumsy attempt to describe a new invention – automobility – by reference to what was familiar. Automobiles, of course, are much more than horseless carriages, as an enriched world of accessible content is much more than a digital library. Whether a new term emerges or not, it seems likely that digital libraries will be part of the bloodstream of knowledge at an earlier stage and in a more integrated way than libraries of printed objects ever were.

Repositories of research papers

Repositories are simply databases, and what distinguishes institutional repositories is the idea that an internal database can serve more than an

administrative purpose, and can constitute a building block in a distributed international service which is a virtual database composed of a user-defined set of cooperating databases on the network. This is, therefore, an essentially subversive technology, capable of allowing grassroots publishing by non-publishers, and delivering a service with the same functionality and feel as large commercially published databases. The idea is not unique. The UK's Joint Information Systems Committee (JISC) funded a number of 'CLUMPS' projects as part of the Electronic Libraries Programme (eLib) in the late 1990s. These created federated databases of academic library catalogues. However, that was a more straightforward undertaking, as the library catalogues involved were already central university services, available to the outside world for searching over the Internet. The breakthrough was to make them searchable using the federated search technology of the z39.50 standard. With institutional repositories, however, the challenge is greater, as the databases must first be established locally, then opened to the Web, and then configured for harvesting in order to provide the shared database facility.

Institutional repositories are perhaps particularly applicable in the context of research publications, as they emanate from institutions, and with the right technology in place can be caught at source and built into services. An institutional repository can, therefore, serve as a publisher of research materials – peer-reviewed papers, e-prints, theses, reports, conference papers, working papers and other types of document.

Repositories of other objects

In its seminal analysis of trends in the library and information world published in 2003, the Online Computer Library Center (OCLC) reported on feedback obtained from surveying librarians and users about the relevance of libraries to digital content. One response highlights the need for librarians to address new custodial challenges:

Librarians are way too focused on published material: they should leave that to the Amazons and concentrate on the hard stuff.
(Online Computer Library Center, 2004)

By 'hard stuff' might be meant the types of content which are generated in academic institutions but which are not destined for publication. There are three reasonably well-known examples of such types of

content, though the potential clearly exists for many further types which may have more limited appeal as part of distributed services.

Learning objects

Whereas the development of repositories for research content represents a relatively intuitive migration from pre-digital to post-digital research publishing practices, with many of the same landmarks still visible (pre-prints, peer-review, journal titles and impact factors), when we move into the realm of teaching we find a much less coherent transition taking place. The development of teaching material in digital form embraces both institutionally-authored material – lecture notes, image collections, animated programs, assessments as revision aids – and externally-published material (typically textbooks). So, while it is conceivable that an institution could eventually capture all of its research outputs in an institutional repository, it is much less easy to see how its learning material could be captured so extensively. Copyright presents a major hurdle to this in itself, and institutions are not in as strong a position to seek copyright exemptions – even for textbooks to which they may have contributed – as they are with research papers written by their own academics.

Nonetheless, there is an obvious argument to be made for storing such teaching material as can be stored, in order that it can be reused by colleagues and future teachers, or modified and reused in new teaching contexts. ‘Learning objects’, however, are a heterogeneous group of materials (which is why such a clumsy, abstract name has been adopted for them). They vary enormously in format, in metadata requirements, and in size. Pulling them all together into a single repository presents considerable challenges. The advantages in doing so, however, are the same as those which apply to research outputs. It makes more efficient use of the institution’s resources; allows the digital content to be preserved over time; provides a comprehensive view of institutional product; supports high-quality searching; and permits interoperability with similar repositories across the Web, so contributing to a global service.

In the same way as applies with research outputs, learning object repositories can also be classified into disciplinary and institutional. In contrast to the situation with research materials, however, we find that institutional repositories of learning objects are relatively uncommon. It is rare as yet to find institutions which view the aggregated collection of their learning objects as having a useful ‘showcase’ value. In addition,

whereas research materials are likely to be quite widely read by other members of a disciplinary community across the world, the value in learning objects lies in their capacity to be re-used. For that reason, we find that disciplinary repositories of learning objects predominate, with a few cross-disciplinary services emerging, sometimes national in scope, such as the UK's JORUM repository, funded by JISC. The software platforms used for these repositories are not as standardised as are those for repositories of research materials, and are not all capable of being harvested via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). This imposes limits upon their value as part of a comprehensive virtual database of freely available learning materials.

Table 1.2 Examples of learning object repositories

Learning object repositories
Disciplinary
Several UK Higher Education Academy subject centres provide repositories of learning objects, including:
Bioscience: Imagebank www.bioscience.heacademy.ac.uk/imagebank/
Management and Accountancy: Learning and Teaching Resources www.business.heacademy.ac.uk/resources/landt/
Geography, Earth and Environmental Sciences: Resource Database www.tellus.ac.uk/
Health Sciences and Practice: Learning-Teaching Web-Resource www.health.heacademy.ac.uk/site/ltresource/index.php
Information and Computer Sciences: Learning Objects for Introductory Programming www.ics.heacademy.ac.uk/Resources/Learning_Objects/index.shtml
Languages, Linguistics and Area Studies: Materials Bank www.llas.ac.uk/resources/bankcontents.aspx
Materials Education: Database of Resources www.materials.ac.uk/resources/index.asp
Physical Sciences: Courseware www.physsci.heacademy.ac.uk/Resources/Courseware.aspx
Psychology: Resources www.psychology.heacademy.ac.uk/html/resources.asp

Table 1.2 Examples of learning object repositories (*cont'd*)

Learning object repositories
Disciplinary (<i>cont'd</i>)
Perseus Project <i>perseus.csad.ox.ac.uk</i>
Dr J's Illustrated Guide to the Classical World <i>liit.ilstu.edu/drjclassics/</i>
Images from History <i>www.hp.uab.edu/image_archive/index.html</i>
NORINA: The Norwegian Reference Centre for Laboratory Animal Science & Alternatives <i>oslovet.veths.no/NORINA/</i>
DERWeb: Dental Educational Resources on the Web <i>www.derweb.co.uk/</i>
Institutional
The University of Birmingham School of Dentistry Ecourse and CAL downloads <i>www.dentistry.bham.ac.uk/fordentists/caldownloads.asp</i>
University of Leicester School of Archaeological Studies, Departmental Image Collection <i>www.le.ac.uk/archaeology/image_collection/</i>
LORE: Learning Object Repository for Edinburgh University <i>www.lore.ed.ac.uk/</i>
Cross-disciplinary
JORUM <i>www.jorum.ac.uk</i>
SCRAN <i>www.scran.ac.uk</i>
Cooperative Learning Object Exchange (CLOE) <i>cloe.on.ca/</i>
HELIX Image Service for Higher Education <i>helix.dmu.ac.uk</i>
Multimedia Educational Resource for Learning and Online Teaching (MERLOT) <i>www.merlot.org</i>

Corporate assets

Another role for institutional repositories is in the management of corporate assets. This is the territory of the archivist first and foremost, and embraces institutional records, including curricular descriptions

(calendar, prospectuses), examination results, annual reports from the institution and its subdepartments, and many other records. Clearly, much of this material is confidential or sensitive, and, therefore, not to be made available in the public domain. There is nevertheless a desire to make available, within a single campus, as much information of this sort as is consistent with devolving responsibility for its creation and maintenance to the most appropriate point of data entry. This is desirable for reasons of efficiency, as it reduces indirection. So, wherever possible, academics in their role as Directors of Studies, for example, should input information regarding the course choices of individual students directly into the database which acts as the source repository for the institution, rather than send details to a secretary or administrative officer to input. The development of web-based intranet environments is allowing more and more efficiency in operations involving corporate assets.

The archival function, which is often managed by the library, requires that these corporate assets – now in digital form – be subject to archival process. This means that they need to be appraised as to their future value for the institution, and then preserved for a defined period, which may in fact be an indefinite period of time. Because they are now digital, inevitably they contain the ability to be linked to associated documents which may be held in other institutional repositories potentially available on the Internet. Repositories of past exam papers, of course materials, or of prospectuses, are all examples. Some assets will only be available to authorised staff within an institution, and the most sensitive information will be available only to particular members of institutional staff with accorded privileges. Each category of material must take account of relevant legislation to protect the rights both of the data subject (e.g. the

Table 1.3 Examples of corporate assets in a linked repository environment

Preservation period	Public domain	Institutional domain	Protected domain
One-year or less	Prospectus	Student course records	Student financial records
Several years	Calendar	Student exam results	Staff salary records
Indefinite	Past exam papers	Student personal details	Staff appraisal records

Data Protection Act in the UK), and of the enquirer seeking access to corporate information for a valid reason (e.g. the Freedom of Information Act in the UK). Being digital, the assets themselves can be preserved, and can be interlinked in ways which make efficient use both of them and of the time of staff maintaining or accessing them. This requires that a lifecycle approach to assets be adopted by the institution, and at the present time few institutions are doing this optimally.

Granular content

Another challenge for the architects of systems based around repositories is presented by the granularity within objects. Documents are often compound objects, and so composed of more than a single file. A research paper may have colour photographs embedded within it, or an associated table of data. Learning objects, in particular, have a troublingly elastic definition. A learning object can be as large as a year-long course of study in a subject, or as small as a single image file. It is important that institutional repositories have the means to describe their objects both at the highest level of granularity – the document level – and at the lowest level for each constituent part which, for a variety of reasons, requires independent description.

Repository objects are, therefore, often hierarchies in themselves, sharing the character of archival records more than individual object catalogue records. Each constituent record of a document or object requires its own metadata and this has to show both the relationship of the part to the whole record, and also the rights which inhere in its referent. For example, a PhD thesis may include photographic images where the copyright is not owned by the author. The cataloguer of this thesis then has to try to identify the rights ownership. This is not an easy task, and it is often solved for repository managers by devolving it to the authors themselves. Authors are clearly better placed to establish the copyright in embedded objects than are cataloguers, and a workflow step to require this is regularly included.

What role do institutional repositories play?

While institutional digital libraries are making inroads into the consciousness of their users, it is nevertheless true that the march of digital content via the Web makes many of their services less vital than

they were, and even redundant to a growing proportion of users on campus. The concept of institutionality is an increasingly fragile one when we consider digital content and digital libraries, and we, therefore, must ask whether we should be developing institutional repositories at all. Are they an attempt to shut the stable door after the horse has bolted? All institutional digital library services face a tough battle in being accepted on campuses because alternative systems usually exist, and their shortcomings are not always obvious. Institutional repositories are not an intuitively necessary development in the minds of most academics. Few people yet feel they do not need a physical library on their campus doorstep, but many – particularly those experienced in using subject-based repositories such as arXiv – are surprised to hear librarians arguing for the creation of institutional repositories as new services.

Might it not be better for publishing agencies – content aggregators – to work on behalf of subject disciplines directly in the development both of repositories and broader digital libraries, with institutional digital libraries requiring only a minimal presence? After all, repository and digital library development to date has been more successful in the disciplinary than the institutional sphere, and has been driven directly by academics themselves. Libraries, by trying to create generalised institutional services, are confronted by the twin difficulties of acting as third-party agents between academics and their content and so being perceived as unnecessary, and of seeking to impose conditions upon academics in order to attract content, which may be resented.

It is not yet clear whether institutional repositories will take root and flourish in the digital knowledge landscape. As an innovation, they are still at an early stage of diffusion. What is clear is that they are regarded as a strong and important new idea by many organisations which are concerned with the dissemination of research outputs. Their appeal lies in the idea of ‘groundedness’. Institutions are themselves the ground from which emerge the outputs of research – ideas, proposals, hypotheses, experiments, data and reported results. These outputs now share a common DNA in digital representation. It is this common base format which allows institutions to look more closely at their traditional way of managing research outputs – using print and microform – in order to discover whether there are new and more efficient modes of operation. ‘Research outputs’ traditionally are just that – research publications which are ‘put out’, given away to third parties for further processing. In such a process there is a loss of control, by the institution and the research funder, and with that loss of control come the problems

which libraries are well aware of, as are increasing numbers of academic staff and researchers: the loss of the alignment of the output with the aims of the research funder; and the partial loss of the output to research generally across the world, because publishers require payment for their efforts in dissemination. The outputs, now in the hands of publishers, have to be 'bought back'. Inevitably, this means that only some researchers will benefit. But if the outputs are of near-publication quality while still 'on the ground', because of their digital DNA, then what new opportunities are opened up?

In pre-digital times, when researchers wrote up their results for publication, they would have been posted, hand-written or in typescript, to a publisher – the only agent with the technology to present the finished paper in a pleasing form, and to reproduce it in multiples sufficient to meet the likely demand across the world, in their journals. Publishers also managed a third very important process – that of verification that the research was of a quality which made it valuable to other researchers. This is achieved by the system of peer review, and is critical to the advancement of knowledge, and, therefore, to the careers of researchers as they develop. If a piece of research is flawed or unoriginal, then the advancement of knowledge is stultified or even damaged, and at some future point this fact is likely to become obvious to other researchers, so that the researcher responsible is tarnished in the eyes of their peers – with obvious consequences for personal self-esteem and career development.

In the digital age, the presentation and reproduction functions do not require the intermediation of a publisher. This is what an institutional repository can do. In doing so, the institution is granted a capture function similar to the archival functions which have long existed – in pre-digital time also – for corporate records. Sending research papers to publishers immediately they have been written was a necessary process, but not an ideal one. If the overall work required is not made noticeably more arduous, how much better to record the outflow of the institution's research as it leaves the premises, stamping it at source with the institution's imprimatur, and asserting ownership rights over it – either for the institution or for the author themselves. In the words of Herbert Van de Sompel, 'Scholars deserve an innately digital scholarly communication system that is able to capture the digital scholarly record, make it accessible, and preserve it over time' (Van de Sompel et al., 2004). Van de Sompel's analysis is founded on a concern about data loss and the need to provide effective data curation, but it implies an emphasis on the role of the institution in the lifecycle nonetheless:

We feel this loss needs to be remedied in a future scholarly communication system by natively embedding the capability to record and expose such dynamics, relationships, and interactions in the scholarly communication infrastructure. Recording this body of information is synonymous to recording the evolution of scholarship at a fine granularity. This will allow tracing the origins of specific ideas to their roots, analyzing trends at a specific moment in time, and forecasting future research directions. (Ibid)

This new functionality is obviously desirable but was given little attention in the past because it was virtually impossible to administer, and there was no obvious benefit in any case. Institutional repositories now make the administration relatively simple, and the future benefits have come dramatically into focus in recent times. These benefits derive mainly from the extraordinary potential of repository networking which has been made possible through the development of the OAI-PMH protocol.

Herbert Van de Sompel, developer of OAI-PMH, has regularly described how the invention has the ability to serve the purposes of the academy – and the interested public – without sacrificing any of the tried and trusted elements of the research dissemination and publication process. He quotes the scholarly communication lifecycle model of Roosendaal and Guertz, with its five key components (Roosendaal and Guertz, 1997):

- *Registration*: allows claims of precedence for a scholarly finding.
- *Certification*: establishes the validity of a registered scholarly claim.
- *Awareness*: allows actors in the scholarly system to remain aware of new claims and findings.
- *Archiving*: preserves the scholarly record over time.
- *Rewarding*: rewards actors for their performance in the communication system based on metrics derived from that system.

In the traditional print world, registration and certification require publishers, and awareness and archiving are carried out by libraries. Rewarding is done by a variety of actors, both institutional (e.g. promotion by the university) and at national and international levels, through rewards such as increased funding for research, visiting professorships, and invitations to contribute to scholarly works and conferences. In what Van de Sompel elsewhere describes as a

‘decomposed scholarly communication system’ (Van de Sompel, 2000) involving repositories on the Web, there is no longer a need for separate agents responsible for each stage in the process. Instead, the repository, working in concert with other compliant repositories across the Web, becomes an ‘interoperable grid’ supplying in itself all of the elements of the system – registration, certification, awareness, archiving and rewarding.

Van de Sompel also presents librarians with some serious food for thought. The migration of the scholarly process onto the Web, with a central role for the institutional repository, raises questions about the continued role of the library as an agent for the purchase of published material:

It has become increasingly difficult for libraries to fulfil their fundamental role of safeguarding equity of access ... At the core of the problems that libraries are facing is the total dependency on information held upstream in the information chain. (Van de Sompel, 2000)

In other words, they are in danger of becoming redundant – in at least those of their functions which depend on content held elsewhere. But there is some good news for libraries if they can seize the initiative presented by institutional repositories and ensure that they run them on behalf of their organisations. Libraries are close to authors, and so in ‘a great position to fulfil the registration function i.e., obtain institutional material.’ They are also clearly well qualified to archive this material. They are ‘fast at embracing new technologies’, and full of very knowledgeable people. However, there are some dread warnings as well:

As organizations libraries are slow movers, hosted by slowly moving institutions. Libraries are slow to recognize the fact that a new technology may allow [or beg] for a new mode of operation. The information world runs on Internet time (Van de Sompel, 2000)

This slow speed of response might be fatal for libraries. They may have the technology at an early stage, but they generally do not use it to engender change in their host institution’s organisational practices, and so they run the risk of losing out to other players in the digital content marketplace. The greatest challenges of all for university libraries wishing to populate institutional repositories within their digital libraries

may, therefore, be outreach and liaison. These are not activities which are normally given high priority, and this must change if libraries are to claim a key role in the scholarly communication lifecycle.

Metadata

There are two main components in the construction of digital libraries. One is technologies, several of which have been spawned by the digital library community itself. The other is metadata. Metadata has long been the ‘bread and butter’ of libraries. ‘Data about data’, as it is commonly described, metadata creation involves the production of records which act as proxies to the holdings of libraries, allowing those holdings to be discovered, whether the searcher is looking for a known item, or an item on a particular subject, or by a particular creator. For hundreds of years, cataloguing (still the favoured term by many librarians, rather than ‘metadata creation’ which can sound like an attempt to give a core library activity a pseudo-scientific digital age identity) was a local activity, designed to make usable the particular collections held by particular libraries in particular places. With the scale increase in volume of holdings typically held by libraries in the twentieth century, however, it soon became obvious that much of this labour could be shared so that the same items were not being separately catalogued in many different places. Subject indexing had already been standardised through the widespread adoption of the Dewey Decimal and Library of Congress schemes, among others, and cataloguing likewise became a standardised activity for an increasing proportion of the total, as libraries saw the benefits in sharing their efforts. Catalogue records, including standardised subject codes and headings, were initially posted out to subscribing libraries on cards, but were to become distributed over wide area networks from the early 1980s onwards, as the first generation of computer-based catalogues appeared in libraries.

The MACHine Readable Cataloguing standard (MARC) evolved as an international standard, and is still massively used across the world for the cataloguing of books and journals according to a very detailed *schema* (defined set of record elements). MARC can be used for the description of items other than books and journals, but it falters in its capacity to describe the requirements of digital objects, and has been supplanted in digital repositories by other, more appropriate standards – none of which yet has the universality of use of MARC.

Designers of databases or repositories of digital content have had to give very high prominence to the importance of metadata in their undertakings, however reluctantly. Although traditionally a professional library activity (which means that cataloguers require a degree in library science, or a degree in another subject together with a postgraduate library science qualification), cataloguing is generally not considered to be at the glamorous end of information work, and yet metadata is the key to unlocking the digital content which institutions – via their libraries – are so keen to make available. It is ironic that, just as the combination of shared cataloguing and Internet connectivity was in the process of diminishing the relative importance of cataloguing as a locally-based activity, and deprofessionalising it in some libraries, the digital order has revealed content which is much more complex to describe than print. This is due to the fact that digital documents are illusory objects, presentations or performances composed of many layers of technology.

The document that one reads on a computer screen is assembled from a stack of machine protocols, a particular operating system, and applications software. The work may be stored in a single file, or be composed of several interrelated files. Each file will have a particular file format. Each of these components has a generation number. Besides that, an authoritative copy of the document may be stored in a digital store somewhere (or should be), and it will have a number of rights which exceed by some distance the relatively simple copyright with which the print world was familiar. There may be the separate rights of one or several creators, as well as the rights enjoyed by the reader of the document, and rights belonging to the institution which hosts the copy a reader may happen to be using. In short, for digital content, the description of the object – who created it, what it is called, what it is about, where it was published – is only one of many dimensions of metadata which need to be recorded in order for identification to take place.

The first generation of institutional repositories used the Qualified Dublin Core metadata schema to describe the content of their objects. Qualified Dublin Core is still widely used, mainly because it is specified by the OAI-PMH as a ‘lowest common denominator’ format well suited to supporting harvesting into a commonly structured repository, and, therefore, supporting discovery interoperability. Qualified Dublin Core utilises a 15-element record. A typical QDC record for an institutional repository item is shown in Table 1.4. This is the record for a biological sciences paper held in the Edinburgh Research Archive.

Table 1.4 QDC record from the Edinburgh Research Archive

DC Field	Value	Language
contributor.author	Byrne, Mary E	–
contributor.author	Barley, Ross	–
contributor.author	Curtis, Mark	–
contributor.author	Arroyo, Juana Maria	–
contributor.author	Dunham, Maitreya	–
contributor.author	Hudson, Andrew	–
contributor.author	Martienssen, Robert A	–
coverage.spatial	5	en
date.accessioned	2005-02-08T17:10:54Z	–
date.available	2005-02-08T17:10:54Z	–
date.issued	2000-12-21	–
identifier.citation	Byrne ME, Barley R, Curtis M, Arroyo JM, Dunham M, Hudson A,	en
	Martienssen RA, NATURE, 408 (6815): 967-971 DEC 21 2000	
identifier.uri	www.nature.com	–
identifier.uri	http://hdl.handle.net/1842/687	–
description.abstract	Meristem function in plants requires both the maintenance of stem cells and the specification of founder cells from which lateral organs arise. Lateral organs are patterned along proximodistal, dorsoventral and mediolateral axes (1,2). Here we show that the Arabidopsis mutant asymmetric leaves1 (as1) disrupts this process. AS1 encodes a myb domain protein, closely related to PHANTASTICA in Antirrhinum and ROUGH SHEATH2 in maize, both of which negatively regulate knotted-class homeobox genes. AS1 negatively regulates the homeobox genes KNAT1 and KNAT2 and is, in turn, negatively regulated by the meristematic homeobox gene SHOOT MERISTEMLESS. This genetic pathway defines a mechanism for differentiating between stem cells and organ founder cells within the shoot apical meristem and demonstrates that genes expressed in organ primordia interact with meristematic genes to regulate shoot morphogenesis	en

Table 1.4 QDC record from the Edinburgh Research Archive (*cont'd*)

DC Field	Value	Language
format.extent	369139 bytes	–
format.mimetype	application/pdf	–
language.iso	en	–
publisher	Nature Publishing Group	en
subject	Asymmetric	en
subject	leaves	en
subject	mediates	en
subject	leaf	en
subject	patterning	en
subject	stem cell	en
subject	function	en
subject	Arabidopsis	en
title	Asymmetric leaves1 mediates leaf patterning and stem cell function in Arabidopsis	en
type	Research Paper	en

Gradually, however, the multidimensionality of digital objects is bringing new metadata schemas into play, which provide for the ‘packaging’ of metadata from a variety of schemas suited to the different dimensions of objects – descriptive metadata, technical metadata, rights metadata and other dimensions in some cases. There are a few such schemas now in implementation, including METS, MPEG-21 DIDL and SCORM.¹ They are commonly referred to as *complex object formats*. The frontrunner among them at the present time for institutional repositories of research materials is METS, the Metadata Encoding and Transmission Standard, jointly developed by OCLC and the Library of Congress.

The METS standard is structured into seven sections, as follows:

1. *The header*: metadata describing the document.
2. *Descriptive*: this section may point to external descriptive metadata (such as a MARC record), or contain internally embedded descriptive metadata, or both.
3. *Administrative metadata*: this section provides information describing how the files were created and stored, intellectual property rights, etc.

4. *The file*: lists all files containing content which comprise the digital object.
5. *The structural map*: this outlines a hierarchical structure for the object, and links elements to content files and related metadata.
6. *The structural links*: this records the existence of hyperlinks between nodes in the hierarchy outlined in the structural map.
7. *The behaviour*: this can be used to associate executable behaviours with content in the METS object.

For learning objects, the IEEE Learning Object Metadata (LOM) standard tends to be used as the source schema. LOM implements the IMS Content Packaging standard, and so performs a similar task to METS in supporting the cataloguing of compound objects but within a pedagogical context. The LOM standard, for example, has nine categories (General, Lifecycle, Meta-Metadata, Technical, Educational, Rights, Relation, Annotation and Classification). Educational and Annotation are clearly categories which are quite specific to learning objects.

Cataloguing for the digital library requires skilled practitioners, and this may create a tension in libraries as traditional library cataloguing in the past few decades has increasingly been taken up by computer-based shared cataloguing systems which have reduced the burden of original cataloguing, and turned the bulk of print cataloguing activity into a relatively routine operation. By contrast, a metadata editor needs to know several different metadata schemas, and to apply them – or, often, to interpret them in a standardised way – to a heterogeneous range of digital object formats. The dominance of the age of MARC is over, when a single schema served for the description of any book or journal, and also for a few other things besides – realia (concrete objects housed in libraries, such as toys, or exhibition objects), and early websites among them. Now the metadata editor or creator has to be an artisan, with a variety of tools in their workshop, appropriate to the digital object in hand.

This adjustment is difficult for libraries because it requires a changing of organisational shape, and a reclustered of professional posts around the activities of the digital library. Few libraries have yet made the switch to this mode of hospitality to digital data. One reason for this is that it costs extra money to do so, as printed items are still flowing into our libraries in numbers at least as great as they ever did. The second reason is that there is no established workflow for the capture of records for digital materials. Libraries physically need to unpack printed items, and they ensure that the items are catalogued and classified before they leave the back-of-house to

take their place on the shelves. This is essential in order to make a library out of a chaos of materials. In the digital realm, however, the materials which are added to a library repository are in many cases already findable on the Web. They need to be made more findable, and consistent with each other – not simply because librarians like order and consistency, but because knowledge machinery (such as the OAI-PMH) relies upon standard ways of description in order to generate meaningful indexes.

Repositories and research impact

Researchers are rewarded for their work not financially but through its impact. They want their research to be read, consumed and understood. They want their peers to comment on it, credit it and add to or extend it. Naturally, they want to receive credit for adding to human knowledge of the world; equally naturally, they want to help make the world a better place.

The conventional method of research dissemination via publication in journals is much more limited in its possible impact (through market forces) than is the new method of publication of the same research in open access repositories. Studies have already shown that open access research papers are read more widely, and, therefore, cited more frequently, than papers which are not housed in repositories. The consequence of this is that they have greater impact.

The Institute of Scientific Information (ISI) has produced impact rankings for scholarly journals for many years, based upon its series of citation indexes, now web-based and known as *Web of Knowledge* (*wok.mimas.ac.uk*). Impact factors are based upon the average number of times that papers in a given journal title are cited by other papers – a fair measure of their research impact, though not without some distortions, as ISI itself points out in its regular publication which presents impact rankings, the *Journal of Citation Reports*, where its online help text states:

You should not depend solely on citation data in your journal evaluations. Citation data are not meant to replace informed peer review. Careful attention should be paid to the many conditions that can influence citation rates such as language, journal history and format, publication schedule, and subject specialty.

This methodology is not an appropriate way to measure the impact of open access research papers, however, as it is based on journal titles

rather than individual papers. While a growing number of open access journal titles now exist, by which is meant that they are published free on the Web with the costs of publication met by authors paying to publish, these are generally a poor measure of comparison because they are newly-founded journals competing against existing, established titles, and, therefore, almost inevitably producing less impact. Harnad and Brody point out this limitation, and suggest a way in which impact can be measured for papers deposited in open access repositories:

To get a realistic estimate of the effect of OA on impact, it is not enough to compare only the 2% of ISI journals that are OA journals with the 98% that are not, to find that they are equal in impact (for this may well be comparing apples with oranges, even if you equate for subject matter). (Harnad and Brody, 2004)

What further needs to be compared is:

- (1) the citation impact of the much higher percentage (perhaps as high as 20–40% according to Swan & Brown's (2004) sample) of articles from the 98% non-OA journals that have been made OA by their authors (by self-archiving them)
with
- (2) the citation impact of articles from those very same journals and issues that have not been made OA by their authors. (Ibid.)

Building on Steve Lawrence's seminal paper, 'Online or invisible' (Lawrence, 2001), Harnad and Brody's analysis of the physics literature for 2001 revealed that the ratio of open access article to non-open access article citations varied from 2.5:1 to 5.8:1. They are now extending the analysis to other disciplines. Kristin Antelman uses the same evidence in order to draw a significant conclusion for libraries. If they learn more about the working methods of the researchers in their institutions, they can provide a strong impetus to the adoption of open access repositories by researchers:

Librarians must be able to draw on a sophisticated understanding of the scholarly communication practices of individual disciplines even as they are rapidly evolving, including scholars' use of prepublication research material not traditionally part of the domain of libraries in a print environment. If we choose to

implement institutional repositories, we also must be able to persuade faculty, many of whom are for a variety of reasons quite reluctant, to contribute their prime research output. Data showing that freely available articles in their discipline are more likely to be cited is powerful evidence of the value of repositories as well as other open-access channels. (Antelman, 2004)

Antelman studied the relevant impact of open and restricted access papers in four disciplines – philosophy, political science, electronic and electrical engineering, and mathematics. She found that, while mathematics had the highest overall proportion of papers available on open access (69 per cent), the discipline in which the comparison between open and restricted access shows the greatest difference in impact measured by citations was in fact political science, which only had 29 per cent of its papers available on open access.

Scientists and social scientists are becoming more and more comfortable with reading articles in online form as a preference. The evidence from the arts and humanities does not yet bear this out, but Antelman believes that the behaviour of researchers in those fields will also change once a critical mass of papers is available in open access repositories. Lawrence points out the part that is played by convenience in the higher impact of open access papers. If papers are easy to get hold of, by being fully available from a usable online source, then they are more likely to be cited, particularly by researchers in a hurry. Libraries need to provide speedy access as part of their service (Ranganathan's Fourth Law), but they also observe a duty to ensure that the material they provide represents a balanced provision. It is ironic that one of the complaints made about some of the largest commercial publishers is that they have manipulated the convenience factor in order to serve their own commercial advantage, by putting their own journal articles within such easy reach of academics that they can benefit from the growing profile of their own titles whose impact factors are thereby boosted. Jean-Claude Guédon points this out in his seminal 2002 work *In Oldenburg's Long Shadow: Librarians, Research Scientists, Publishers, and the Control of Scientific Publishing*:

If, through the manipulation of the number of articles in a given database, a publisher manages to affect the rate of use of its own articles, it also stands to reason that this publisher is able to affect the citation rate of its articles. If this situation leads just one Ohio scientist to cite one more Elsevier article in one of his/her articles, this affects the impact factor of the journal where the article

appears. Of course, with one citation, the effect is too small to be detected, but imagine now that event repeated an untold number of times in Ohio and across other similarly structured consortia. It will lead to increasing the number of citations to Elsevier articles. As a result, the impact factor of Elsevier journals should begin to go up. As a consequence, these journals begin to attract more authors; but then, with a greater choice of authors, the quality should go up. In effect, a kind of quality pump has been successfully primed and it begins to propel the journal up the pecking order ladder among the core journals. (Guédon, 2002)

What holds true for a commercial publisher with a vast number of full-text articles available to subscribing institutions could equally hold true for a large, interoperating, worldwide network of open access full-text institutional repositories. As Antelman remarks, the comprehensiveness of this network is likely to have a beneficial consequence for the quality of research:

One may speculate that when articles are only a mouse click away, 'bad' author behaviors that have been described in the citation analysis literature will be less common. One example is citation bias, where authors reference only journals they can access. (Antelman, 2004)

The provision which librarians wish is one whose underlying motive is aligned to the motives of authors in publishing their research in the first place, and the 'quality pump' maintained by libraries managing institutional repositories can then serve the advancement of knowledge rather than publisher profit.

Conclusion

A growing proportion of the research community has discovered the utility of the Web for the dissemination of their research outputs, and has now been using it – for many years in the case of some disciplines. The approach has been somewhat haphazard, however, as scholars are neither publishers nor librarians. The library community, increasingly focused on a digital library agenda, has understood the need to intervene in order to ensure that the material being disseminated is managed

successfully through proper description, indexing and storage for long-term preservation. The approach which has now proved its value and begun to gain ground for research outputs is also now being used for other types of material which are generated within institutions.

The marriage of research generation by academics, with output management by librarians, has created a new form of publication, with open values, which presents a growing challenge to the commercial publishers which have controlled research publication for many decades. Commercial publishers operate on the assumption of a profit motive both for themselves and for their authors. As this motive is absent in the case of academics seeking the publication of their research, it may be that with this new form of publication in repositories owned and run from within the academy, research publication has finally found its most appropriate form.

Note

1. For definitions, see MPEG-21 (2003) 'Information Technology, Multimedia Framework, Part 2: Digital Item Declaration', ISO/IEC 21000-2:2003; Advanced Distributed Learning (2003) 'The Sharable Content Object Reference Model (SCORM) – Version 1.3 – WD'.