Tomorrow, and tomorrow, and tomorrow: poor players on the digital curation stage[1]

Chris Rusbridge,
Digital Curation Centre,
University of Edinburgh

"To-morrow, and to-morrow, and to-morrow,
Creeps in this petty pace from day to day,
To the last syllable of recorded time;
And all our yesterdays have lighted fools
The way to dusty death.
Out, out, brief candle!
Life's but a walking shadow; a poor player,
That struts and frets his hour upon the stage,
And then is heard no more: it is a tale
Told by an idiot, full of sound and fury,
Signifying nothing."

Shakespeare: Macbeth

This quotation in its true context is a cry of despair. But divorcing from the context, and reading somewhat metaphorically, the relevance of many of its resounding phrases to digital curation and preservation can be imagined. We look forward to the last syllable of *recorded* time, not just to the end of time. We look to our yesterdays, and perhaps the information that was lighted to a dusty death. We think about the poor players who strut and fret to make information available into the future, and then are heard no more (the silence of the librarian?). We see our tales reduced to sound and fury; not meaningful digital documents, but sequences of empty data points, signifying nothing.

Introduction

In this chapter, I will argue that there are non-obvious choices to be made about the "poor players" who manage data. In particular, the role of the librarian in this is not clear.

Reg Carr (Carr 2004)attempted to persuade his CURL colleagues at a meeting in Dublin that they should address the emerging importance of data collections head on. His efforts were accepted with enthusiasm by some, resisted by others. There were good reasons for both positions, but I argue that the latter in particular is a temporary phenomenon, strongly linked to budget constraints and to the current transitional phase of librarianship from "mostly physical" to "nearly all digital".

Paul Courant, economist and ex-provost of Michigan, pointed out to a JISC/NDIIPP meeting in May 2006 that, particularly in the context of library budgets and the need to curate data: "There's plenty of money for anything. There just isn't plenty of money for EVERYTHING!" Casting one's mind sufficiently far into the future, it is clear that the trend to digital is irreversible (unless the world as we know it crashes and burns), and hence in a comparatively short time, digital data as well as digital documents will become primary stuff for libraries, and the resources will adapt accordingly. Meanwhile, awkward decisions are needed, on whether to be pioneers, early adopters or late followers!

---

[1] Based on a Distinguished Visitor presentation with this title given at OCLC headquarters, Dublin Ohio on 26 October, 2006.

But whether they want it or not, what SHOULD be the role of librarians towards data? To answer this, we need to understand a bit more about data curation.

Data Curation

In his closing remarks to the 2005 International Digital Curation Conference in Bath, Cliff Lynch drew 3 views of digital curation from an extended audience-participative discussion:

- Curation as a finite process, with handover to preservation at its end point
- Curation as a whole life process, with evolving objects, and
- Curation as managing a growing, living collection.

Librarians are most comfortable with the last of these, currently in the physical world of books and physical information objects. Building a collection against a collection policy to meet the needs of a defined community; this is home territory for librarians (and archivists, and museum curators). It extends easily enough into the digital library world, as well.

Librarians have also grappled with the first of these concepts. The idea of working on a digital resource, preparing it for preservation, chimes well with scholarly traditions.

But the idea of the changing, evolving resource, is in some ways an uncomfortable one, and not just for librarians. Nevertheless, this is the basis for much of the emerging science of today. As an example of these issues, think of curating and preserving the UK's Ordnance Survey National Map Database, OS MasterMap (www.ordnancesurvey.co.uk/oswebsite/products/osmastermap/). Clearly a very different proposition from preserving the 1st to 7th editions of the map sheets, the most common response is to attempt an annual snapshot, although quite what this means and how it could be used (without large amounts of supporting proprietary software) is unclear.

For the Digital Curation Centre, curation is not simply preservation. It is "maintaining and adding value to a trusted body of digital information for current and future use". Emphasising the present as well as the future, we seek to ensure that those critical early steps are taken that will allow future preservation, and that current information will still be usable even as time extends indefinitely.

Lynch further pointed out that curation has a strong link with stewardship. It:

- Includes resource management
- Includes access and presentation
- Includes active care
- Involves long time, and thus
- Includes preservation.

Curation is clearly domain-dependent. There are significant domain-dependent issues relating to size, numbers of objects, complexity of objects, interventions needed, ethical and legal implications, policies, practices and incentives.

The DCC takes a broad view of digital curation. Whilst not exclusively data-oriented, we predominantly focus on data resources for science and scholarship. We are concerned with:

- The sustainability of the resource.

- The creation or appraisal, selection, acquisition and ingest of the resource,
- Growth, development of and changes to the resource,
- Making the resource available ("publishing" it),
- Access management and other controls on the resource, and the ethical and legal basis of these controls,
- The ability to use, combine, re-combine, inter-operate, process, annotate, discuss and review the resource through time (some of which processes will in turn contribute to the development of the resource),
- Linkage, context and metadata relating to the resource,
- Maintaining authenticity, integrity, provenance and computational lineage information relating to the resource,
- Maintaining the meaning of the resource despite technology change and concept drift in the outside world,
- Preserving the resource, including preserving access to past states of a changing resource,
- De-selection and deliberate and/or accidental destruction of the resource.
- All of this, over potentially extended time periods, although timescales could also be comparatively short or medium term;
- Recognising the impacts of finite budgets and potential future policy changes, and
- Paying attention to the education, training and development of the people to support this.

A Curation example

As an example of the power of data-oriented science, take the case of an early test of the "National Virtual Observatory" concept in the US. Astronomers turn images of the sky at various wavelengths into databases of objects detected by analysing those images. All these objects are well described in spatial terms, and furnished with extensive contextual metadata derived in part from their origins in particular instruments. By combining and cross-searching databases derived from the Sloan Digital Sky Survey (SDSS) with those from the Two Micron Astronomical Sky Survey (TWOMASS), astronomers quickly made a discovery. The Johns Hopkins (Hopkins 2003)press release states:

> "Scientists working to create the NVO, an online portal for astronomical research unifying dozens of large astronomical databases, confirmed discovery of [a] new brown dwarf recently. The star emerged from a computerized search of information on millions of astronomical objects in two separate astronomical databases. Thanks to an NVO prototype, that search, formerly an endeavor requiring weeks or months of human attention, took approximately two minutes."

There are many more examples, from the Human Genome project to the satellite surveys that give us baselines for global warming, and beyond. It is a simple assertion that data are beginning to assume a role in the scientific and scholarly world similar to that of text. An article can tell you about a discovery; a database can let you test certain aspects of the theory or experimental process supporting that discovery. Verifiability is the basis of science.

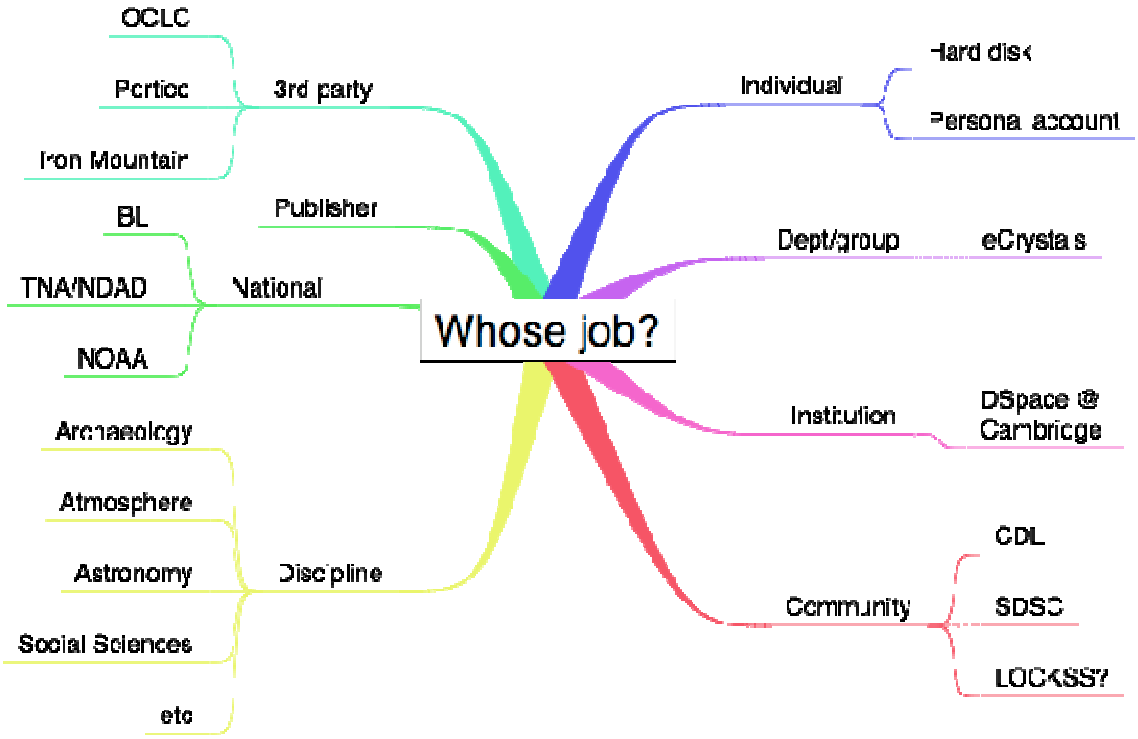
Who are the curation players?

Given this critical importance, how can we assure the continued curation of data? Or in the context of this article, what is the librarian's role in data curation? Whose job is it anyway? Who are the 'poor players' in data curation?

Perhaps not complete, here is a classification of data curators:

- Individuals, using their hard disks, or perhaps networked drives
- Departments or groups
- Institutions, perhaps in the shape of their libraries
- Communities of institutions, either formal (as consortia), or informal (as in the case of the LOCKSS[2] system)
- Disciplines
- Publishers
- National services, perhaps national libraries or archives, or national data services, and/or
- Other 3rd party services.

Figure n.1: Some examples of the classification



Individuals

James M. Caruthers, a professor of chemical engineering at Purdue University has claimed 'Small Science will produce 2-3 times more data than Big Science, but is much more at risk' (Carlson 2006). Big Science results from large collaborative exercises; the sharing implied results in better-defined data formats and access protocols, and often formal support for data sharing in proposals. But Small Science, in the labs of individual Principal Investigators usually results in data managed by their Research Assistants, or even PhD students. The data are often on individual or at best shared drives. They will often not even be adequately backed up. The individuals concerned are intimately involved in the scientific work; they know so much that they do not feel a need to write down: they know *too* much, and are too busy, to

[2] Lots Of Copies Keep Stuff Safe, a distributed service founded at Stanford University.

create good metadata or documentation. At best some time after the PI has moved attention on to a successor project, at worst when a staff member leaves and the accounts are deactivated and then deleted, these data will simply disappear; they have no tomorrow. This case is both most common and most worrying!

Groups or departments

In many cases, group or department curation efforts will be of similar standard, with similar risks, to the work of individuals. However, there are some beacons shining out. Take for example, the eCrystals data resource (http://ecrystals.chem.soton.ac.uk/) curated by the National Crystallography Service at Southampton. Informed by the eBank projects (www.ukoln.ac.uk/projects/ebank-uk/), and lately the Repository for the Laboratory (R4L) project (http://r4l.eprints.org/about.html), they are attempting to capture automatically both process-oriented but essentially private data, and also publicly available crystal structures resulting from their analyses. In R4L, they aim to capture key metadata as part of their workflow. For example, health and safety considerations require them to plan their work out well in advance; this information provides a useful source of data, as does capture of environmental information, and even the staff present in the lab at any time. Their final results are made available in the industry-standard CIF format. They are supported by their library, and are trying to extend their repository to form a federation. Even so, it is not yet clear how they relate to many other significant activities in the field taking slightly different approaches, including the American ReciprocalNet effort, and the French Crystallography Online Database (COD), or even IUCr mentioned below. Nevertheless, with the level of domain knowledge and service commitment displayed, the medium term tomorrow of this collection seems assured.

Institutional and Library

While the eCrystals repository is identifiably separate from the Southampton Institutional Repository and linked to the Crystallography group, the DSpace @ Cambridge repository is clearly institutional, a collaboration between the Library and Computing Service at Cambridge University. It contains many collections, including Archaeology, Manuscripts, Learning objects etc. But the largest collection by far, with some 250,000 digital objects deposited so far, is the World-Wide Molecular Matrix (http://www.dspace.cam.ac.uk/handle/1810/724) including structures of small molecules encoded in Chemical Markup Language (CML). This is definitely an institutional repository, but this part is definitely a chemical collection. The Library applies no chemical skills in curating this collection, relying exclusively on the considerable enthusiasm of its depositor, Peter Murray Rust. The collection is isolated from other chemical collections, and the repository provides no (non-generic) services that are particularly relevant to chemistry.

There are some who hold that institutional repositories such as this have no valuable place in the data world, arguing that domain science knowledge is essential to adequately curate science data (Lawrence 2005). However, institutional repositories do have two major advantages over discipline-based repositories: their institutional resource base is driven by their institution's continuing interest in disclosing its research, and their association with major institutional collecting organisations like the library tends to give them a stronger tomorrow. In this case, Cambridge has made a strong commitment to its repository, so the future of the repository and the collection in its current form is reasonably assured.

Comparatively few other libraries can claim any significant data holdings in their institutional repositories. The OpenDOAR service (www.opendoar.org/) listed 5 in the UK at a recent visit. The library role then is not yet nationally significant, and there is little sign of curation

repositories appearing on an institutional basis anywhere other than in libraries. Given the issues about Small Science, however, perhaps librarians should be looking at increasing their involvement.

Community Services

If one institution can do reasonably well, can a community service supporting many institutions do better? One interesting example is the California Digital Library. Set up by the Regents of the University of California, and located in the Office of the President, CDL provides digital services to the constituent university libraries. These services include the UC Libraries Digital Preservation Repository Service (www.cdlib.org/inside/projects/preservation/dpr/). It does appear to be from a document rather than data tradition, and like UK libraries, has a passive role in relation to the collections preserved; either the individual libraries or more likely the research groups and staff they serve (two or more steps removed) provide the curation skills, and CDL provides preservation. Nevertheless, their tomorrow is reasonably well assured.

LOCKSS (www.lockss.org) provides a completely different example of a community service. In this case the community is much more like an Open Source community: a self-selected group of collectors using open software on cheap commodity computing boxes gathers web-like objects (for which the collectors have the required rights) into a cache, continually checks their integrity against other boxes, and makes them available to their community should the original disappear[3]. LOCKSS is also rooted in a document tradition (libraries collecting eJournals), but is being increasingly applied in other contexts. However, there is intrinsically little domain knowledge in a system such as LOCKSS. Nevertheless, it is potentially a very valuable model because of its high redundancy, low cost, high reliability and high attack resistance; these are properties that it is difficult to replicate in larger scale systems. Consequently, a LOCKSS system of peers configured to capture data could also have a strong tomorrow.

Disciplines

Of the examples above, only the group example had the active involvement of domain scientists in the curation of their data. The "doubters" of institutional repositories claim that discipline-based repositories have the major advantage of that active involvement. As the National Science Board report on Long-Lived Digital Data Collections (NSB 2005) suggests, they also act in "community-proxy" roles, particularly when it comes to defining data and metadata standards. Here are a few examples:

- Archaeology in the UK is served by AHDS Archaeology (formerly ADS). Staffed by archaeologist curators, they understand complex issues such as the legal opportunities and requirements provided for archaeological finds discovered during building and civil engineering development processes. They have a strong relationship with their community and their peers, being located within an academic archaeology department. As an example, see their digital resource on Roman Amphorae (Keay 2005), AHDS Archaeology does appear to relate solely to the UK (in their immediate stakeholder group, rather than archaeological scope), and internationally the scene appears rather

[3] This picture is slightly complicated by the LOCKSS Alliance, a subscription-based membership organisation whose fees sustain the Stanford LOCKSS team for software development and the activities including publisher negotiations associated with acquiring the rights and technical capabilities to preserve new content.

fragmented. Their "tomorrow" is supported by a combination of funding sources, including deposit fees and research council grants.

- As mentioned above, Astronomy is an example of Big Science that is organising itself around systems of Virtual Observatories. This is part of a major international effort. Astronomy requires very expensive, shared large facilities (it is definitely Big Science), and is used to collaborating internationally, and to sharing data. The VOs are well integrated into their community, who understand that they are essential to generate certain types of new scientific knowledge. Because they can clearly be seen as another Large Facility (a telescope into the past, perhaps), their tomorrow is well assured by community commitment.
- Atmospheric Science, like most of the environmental sciences, clearly understands the value of past observations (which cannot be repeated), and hence the value of curating today's and tomorrow's observations. The Director of the British Atmospheric Data Centre (BADC), which is funded by NERC, is a strong believer in the necessity of having domain scientists as curators (Lawrence 2005); he also acts on his belief in exercising a strong community-proxy role. Internationally, atmospheric science seems well served with repositories, but perhaps they are more fragmented than one might expect (although the NERC Data Grid is trying to unify a few of them). Their tomorrow is mostly dependent on grant funding, but with a strong commitment to the need to support such activities from its funder.
- High Energy Physics is another example of Big Science (indeed, flexing its muscles as Biggest Science!). The Large Hadron Collider is building tiers of data stores in many different countries to handle the floods of data that will emerge once it becomes operational.
- Pharmacology is interesting. In particular, the International Union of Pharmacology (IUPHAR) has a database of pharmacological receptors. It is attempting to build academic credit for the contributors to this database, and as such is taking steps to introduce quite fine-grained data citations (Buneman 2006). Funding of the database is extremely limited, which certainly acts against an assured tomorrow, particularly if curation requires significant database investment.
- The Social Sciences, both in the UK and the US, have long and mature histories of data curation. Both ESDS (www.esds.ac.uk) in the UK and ICPSR (www.icpsr.umich.edu/) in the US are staffed by Social Science curators; they are alert to opportunities, able to appraise material offered, and have a strong relationship to their disciplines, where acceptance for deposit can be seen as a badge of merit. In the case of IPCSR, with their broad mix of funding streams, tomorrow is fairly well assured. In the case of ESDS, with more limited funding streams (primarily ESRC and JISC) there may appear to be more risk. However, ESDS is certainly viewed as one of the jewels in the ESRC portfolio, so their tomorrow is also pretty strong.

These examples, some more successful than others, show that discipline-based curation services can work, and do have advantages. However, disciplines are hard to define, and fracture almost as soon as defined. The successful examples above do not represent the full breadth of the discipline base; in fact they are exceptions rather than the rule. ESRC can see a need for one curation service covering all of the economic and social sciences in the UK, but NERC funds 7 just within the environmental sciences. It is not clear that anyone in the UK funds any curation service covering engineering data, despite the obvious long-lived compliance requirements. There are around 800 databases world-wide of relevance to nucleic medicine (Bateman 2006), of which maybe 100 are supported by the European Bio-Informatics Institute. It is a very patchy picture, and one where directors of discipline-oriented curation services are perpetually chasing funding, and live in fear of those dreaded words: "policy change".

Publishers

Sometimes publishers have close connections with their disciplines. While some publishers are distrusted as rapacious, whose possible moves towards data collection would be seen as yet more attempts to gain exclusive rights for profit, others can be seen as having a strong and trusted role. One such is the International Union of Crystallographers (IUCr), which publishes Acta Crystallographica in various parts. IUCr, working with their community, defined the Crystallographic Information Framework (CIF, www.iucr.org/iucr-top/cif/) that allows crystallographic information to be shared, and has made deposit of validated structure information a pre-requisite for publication of articles or structures in their journals. They provide services that allow CIF files to be checked for various quality parameters. They are also pioneers in the use of Digital Object Identifiers for data objects. Their combination publishing and membership business model is probably secure for tomorrow, although all publishing business models are under threat right now.

While publisher mandates for deposit can be powerful drivers (as can funder mandates such as the Wellcome Trust's) it is clear that many publishers would be viewed with extreme suspicion if they tried the same approach. Perhaps the key here is IUCr's close identification with its discipline.

National bodies

What about national bodies? The British Library is undertaking a serious development programme for its Digital Object Management system, motivated by its upcoming non-print legal deposit powers and responsibilities. However, it is oriented towards "cultural heritage", broadly interpreted, and does not claim much data or science domain expertise. While the BL will no doubt accept data (eg the National Mapping Database referred to above), they are not natural data curators.

The National Archive has set up the National Digital Archive of Datasets (NDAD, www.ndad.nationalarchives.gov.uk) at the University of London Computing Centre, to be a specialist archive for government datasets. NDAD understand the complex government regulations, dynamics and requirements, and are technology specialists, understanding databases very well. Although some of their datasets have significant science value, NDAD is not staffed by domain scientists, and they remain subject generalists. Tomorrow, there is every likelihood that the operation will pass back in-house to TNA, who will, however, have a long-term interest in sustaining it as part of its statutory duties.

In the US there is a variety of national bodies with a discipline science responsibility. NASA and NOAA are two examples making serious data available for public and scientific use. In these organisations, domain scientists do curate the data, and sometimes with massive budgets. However, they tend to be subject to the current political context, which can lead to continuity problems (policy change again!), and are often subject to "un-funded mandates" (legal requirements to carry out responsibilities without the means to do so). The political context places these organisations continually in some jeopardy.

Third parties

What about 3rd party organisations? The first two worth mentioning could also be classed as community-based, perhaps. OCLC runs a digital preservation service (www.oclc.org/digitalarchive/about/), on a demand-driven basis, agnostic as to content. Its tomorrow is based on belief in a business case; it is unlikely to be paying its way at this stage. Portico (www.portico.org) is a preservation service set up (essentially) by Mellon, with subscription funding from universities and publishers, to preserve eJournals. It too has no data

or domain science expertise, and is highly dependent on those publisher preservation rights agreements. The funding mix and the power of Mellon (which cannot afford to see this venture fail) probably means its tomorrow is secure.

Finally, we should think about the role of real for-profit 3rd parties, such as Iron Mountain (www.ironmountain.com/digital/erecords/archives.asp). Records management IS a curation problem, and any company that can make a successful business from electronic records management is very likely to seek to branch out into other forms of curation. They may have no science expertise, but they may have self-belief, ambition and large reserves that will allow them to buy in the skills they need to secure a market. Their tomorrow, however, may be very dependent on the viability (quarter by quarter) of their business plans[4], competition, take-over, the stock market, interest and exchange rates… We should be concerned at their forthcoming roles, but only in the sense of taking opportunities aware of the risks.

Moving things to the network level

Lorcan Dempsey (Dempsey 2006) often talks about "moving things to the network level". It is clear from the above that institutions have some fundamental sustainability advantages, but lack the critical mass of domain science involvement in curation, or fragment it when they can sustain it. Disciplines do exist at the network level, and have huge advantages for data curation in being able to direct domain expertise to the curation task. But sustainability is always an issue for disciplines (and many network level services), and many if not most disciplines have never even got to the point where sustainability has to be confronted!

Can we combine the institution and the discipline to achieve network effects with institution components? The much-touted Web 2.0 effects are achieved by cunning combinations of mass appeal, highly scalable centralised services, and some "power of crowds" synergies from the participation of many individuals. It is difficult to see how this will work in the academic sector, at least at scales that will attract venture capital (although there are a few examples, such as Connotea, www.connotea.org). However, perhaps there is some way of putting together disciplinary segments of institutional repositories to achieve network-level effects? It is not clear how (or if) this can be done, but we should be trying!

Conclusions

At the beginning, I asked what should be the role of librarians in data curation. There is as yet no clear answer, and certainly no simple answer. But for now, librarians SHOULD be continuing to take data ever more seriously, thinking about the relationship between publications and the data on which they are based, and working with their discipline colleagues where opportunities arise. Capturing ANY valuable data is never a wasted opportunity.

Yesterday, Reg Carr took leadership positions in CURL, RLG and JISC, supporting efforts such as these. Tomorrow, we hope his successors will be as visionary.

Bibliography

Bateman, A. (2006). "EDITORIAL." Nucl. Acids Res. **34**(suppl_1): D1-.

[4] For example, at time of writing it appeared that an email curation service, Cryoserver, may have gone into liquidation, with uncertain effects on its clients: Williams, C. (2006). Email archiver melts away. Channel Register.

Buneman, P. (2006). How to cite curated databases and how to make them citable. Proceedings of the Conference on Scientific and Statistical Database Management.

Carlson, S. (2006). Lost in a Sea of Science Data: Librarians are called in to archive huge amounts of information, but cultural and financial barriers stand in the way. The Chronicle of Higher Education. **52:** 35.

Carr, R. (2004, March xx). "The Challenge of e-Science for Research Libraries." from http://www.bodley.ox.ac.uk/librarian/escience/escience.htm

Dempsey, L. (2006). "Libraries and the Long Tail: Some Thoughts about Libraries in a Network Age." D-Lib Magazine **12**(4).

Hopkins. (2003, March xx). "Virtual Observatory Prototype Produces Surprise Discovery." Headlines @ Hopkins, from http://www.jhu.edu/news_info/news/home03/mar03/nvo.html

Keay, S. (2005). Roman Amphorae: a digital resource, AHDS Archaeology.

Lawrence, B. (2005, March xx). "Function Creep and Institutional Repositories." Bryan's Blog, from http://home.badc.rl.ac.uk/lawrence/blog/2005/03/31/function_creep_and_institutional_repositories

NSB (2005). Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century. National Science Board Report. N. S. Foundation.

Williams, C. (2006). Email archiver melts away. Channel Register.