# Retooling Libraries for the Data Challenge

**Dorothea Salo** examines how library systems and procedures need to change to accommodate research data.

## Abstract

Eager to prove their relevance among scholars leaving print behind, libraries have participated vocally in the last half-decade's conversation about digital research data. On the surface, libraries would seem to have much human and technological infrastructure ready-constructed to repurpose for data: digital library platforms and institutional repositories may appear fit for purpose. However, unless libraries understand the salient characteristics of research data, and how they do and do not fit with library processes and infrastructure, they run the risk of embarrassing missteps as they come to grips with the data challenge.

## Introduction

Whether managing research data is 'the new special collections,'[1] a new form of regular academic-library collection development, or a brand-new library specialty, the possibilities have excited a great deal of talk, planning, and educational opportunity in a profession seeking to expand its boundaries.

Faced with shrinking budgets and staffs, library administrators may well be tempted to repurpose existing technology infrastructure and staff to address the data curation challenge. Existing digital libraries and institutional repositories seem on the surface to be a natural fit for housing digital research data. Unfortunately, significant mismatches exist between research data and library digital warehouses, as well as the processes and procedures librarians typically use to fill those warehouses. Repurposing warehouses and staff for research data is therefore neither straightforward nor simple; in some cases, it may even prove impossible.

## Characteristics of Research Data

What do we know about research data? What are its salient characteristics with respect to stewardship?

### Size and Scope

Perhaps the commonest mental image of research data is terabytes of information pouring out of the merest twitch of the Large Hadron Collider Project. So-called 'Big Data' both captures the imagination of and creates sheer terror in the practical librarian or technologist. 'Small data,' however, may prove to be the bigger problem: data emerging from individual researchers and labs, especially those with little or no access to grants, or a hyperlocal research focus. Though each small-data producer produces only a trickle of data compared to the like of the Large Hadron Collider Project, the tens of thousands of small-data producers in aggregate may well produce as much data (or more, measured in bytes) as their Big Data counterparts [2]. Securely and reliably storing and auditing this amount of data is a serious challenge. The burgeoning 'small data' store means that institutions without local Big Data projects are by no means exempt from large-scale storage considerations.

Small data also represents a serious challenge in terms of human resources. Best practices instituted in a Big Data project reach all affected scientists quickly and completely; conversely, a small amount of expert intervention in such a project pays immense dividends. Because of the great numbers of individual scientists and labs producing small data, however, immensely more consultations and consultants are necessary to bring practices and the resulting data to an acceptable standard.

## Variability

Digital research data come in every imaginable shape and form. Even narrowing the universe of research data to 'image' yields everything from scans of historical glass negative photographs to digital microscope images of unicellular organisms taken hundreds at a time at varying depths of field so that the organism can be examined in three dimensions. The tools that researchers use naturally shape the resulting data. When the tool is proprietary, unfortunately, so may be the file format that it produced. When that tool does not include long-term data viability as a development goal, the data it produces are often neither interoperable nor preservable.

A major consequence of the diversity of forms and formats of digital research data is a concomitant diversity in desired interactions. The biologist with a 3-D stack of microscope images interacts very differently with those images than does a manuscript scholar trying to extract the underlying half-erased text from a palimpsest. These varying affordances must be respected by dissemination platforms if research data are to enjoy continued use.

One important set of interactions involves actual changes to data. Many sorts of research data are considerably less usable in their raw state than after they have had filters or algorithms or other processing performed on them. Others welcome correction, or are refined by comparison with other datasets. Two corollaries emerge: first, that planning and acting for data stewardship must take place throughout the research process, rather than being an add-on at the end; and second, that digital preservation systems designed to steward only final, unchanging materials can only fail faced with real-world datasets and data-use practices.

Finally, early experience with data-sharing has shown that it is all but impossible to forecast every conceivable use for a given dataset in advance. Systems that force end-users into unduly limited interactions with the data reduce the usefulness of those data.

## Backlog

Libraries are not starting with a clean slate with research data, any more than they are with their own bibliographic data. Research practices have been partly or wholly digital long enough to have produced a substantial amount of data already. These data, particularly 'small data,' tend to be disorganised, poorly described if described at all, and in formats poorly suited to long-term reuse. Even more unfortunately, researchers have become accustomed to the processes that produce these sloppy data, which makes them liable to resist changing those processes to improve data viability.

To make matters yet worse, much research data that could benefit from being digital is still analogue, the laboratory notebook being the paradigm example. Digitising these resources is not straightforward; straight image scans might as well be analogue for all their digital reuse value, while re-keying is an unjustifiably enormous expense for materials that have a relatively low signal-to-noise ratio.

## Project Orientation

Particularly as science is ever more driven by grant cycles in the waning days of sustained funding, research data are managed by the project. The lack of continuity in this system vitiates incentives toward good data practices; why save data if the next project will be on a different theme with different collaborators? Institutional memory and tacit knowledge about good data practices tend not to accumulate, as collaborators scatter and procedures are worked up from scratch for each new project.

Tools, too, are chosen based on extremely short-term project considerations, magnifying the chance of poor choices from a longer-term stewardship perspective. Sustainability of tool output easily takes a back seat to whiz-bang features. Once projects are finished, usually marked by the publication of articles or reports, intermediate work products such as research data are either deleted altogether or swept unorganised and undescribed into dusty digital closets.

### Non-standard Data and Data Formats

The early days of almost any new venture are marked by tremendous experimentation and frequent blind alleys. Though necessary for progress, this phenomenon is terrible if the goal is any sort of standard result. The diversity of research data necessarily implies that complete standardisation is impossible; that more variation exists than is strictly necessary, however, is undeniable.

A few disciplines have created data standards, usually because of a strong centralised data repository that imposes those standards on researchers wishing to contribute data. The Inter-University Consortium for Political and Social Research [3], with its extensive data-standard documentation and strict standards, is a fine example. Standardisation may also be an emergent quality of collaboration, as evidenced by the standards promulgated by the International Virtual Observatory Alliance [4].

In most disciplines, however, and certainly those where research is individualised or hyperlocal, incentives to create, much less follow, data standards are minimal or non-existent. The resulting sloppy Tower of Babel vitiates both reuse and long-term stewardship.

## Characteristics of Digital Libraries

What do we know about digital libraries? How well do their technical and organisational infrastructures map to research data stewardship?

### Curated

Because digitisation to library standards is expensive, materials in digital libraries are chosen and handled with enormous care. High standards prevail in digitisation quality, in associated metadata, and in presentation. Only materials deemed important enough to warrant such care are digitised at all.

These mindsets and processes are much too labour-intensive to transfer to the enormous backlog of existing research data. They are also at sea faced with the sloppiness of researchers' data practices; not a few librarians will simply assume that data cannot be worth curating if researchers themselves take so little care of them! Digital librarians also rebel at the idea of researchers' sloppy digital data existing alongside their own beautifully curated materials.

If existing digital library processes and procedures cannot hold up under the deluge, libraries will have to choose the datasets they lavish effort on, much as they chose materials to digitise. Are they prepared to alienate researchers whose datasets are not chosen? Will they truly become estimators of data quality across the breadth of data-producing disciplines in the institution? If not, will they rely solely on technical criteria such as file format, regardless of data quality or importance? Whatever criteria are chosen, will those criteria conflict with institution-wide mandates such as the preservation of theses and dissertations with their accompanying materials? [5]

## Taylorist Production Processes

Because digitisation is expensive, most established digital libraries digitise as efficiently and cost-effectively as they can manage. Where this does not mean outsourcing (which it often does), it means exactly the sort of rote, minimum-effort, minimum-judgment workflows known as 'Taylorist' after American manufacturing efficiency expert Frederick Taylor [6]. The variability of research data defeats Taylorist processes utterly. Such processes simply cannot keep up with the professional judgment and technical skill required when most new projects involve new file formats and metadata standards and require individual massaging for ingest and preservation.

Libraries employing Taylorist processes tend to specialise in certain content types and digitisation processes; this maximises return on investment in a given workflow. A library with deep scanning expertise probably does not have equivalent expertise in text encoding. Digital library platforms specialise alongside, for clear and obvious reasons. Both specialised processes and specialised platforms fail when faced with highly heterogeneous, not to say sloppy, research data.

Some, though not all, data can be shoehorned into a digital library not optimised for them, but only at the cost of the affordances surrounding them [7]. Consider data over which a complex Web-based interaction environment has been built. The data can be removed from the environment for preservation, but only at the cost of loss of the specialised interactions that make the data valuable to begin with. If the dataset can be browsed via the Web interface, a static Web snapshot becomes possible, but it too will lack sophisticated interaction. If the digital library takes on the not inconsiderable job of recreating the entire environment, it is committing to rewriting interaction code over and over again indefinitely as computing environments change.

Taylorist digital library processes presume a near-total control over the data creation environment, which is impossible to accomplish for research data. Such processes also presume well-established content and metadata standards, which often do not exist.

Finally, Taylorist workflows determine an organisational structure where library professionals are the supervisors and project managers. They plan and oversee, but do not carry out projects. The actual digitisation and metadata work is done by para-professionals. The resulting bifurcation in technical skill and related knowledge leaves too little practical knowledge at the top of the organization to ensure that the library professionals are automatically capable of working effectively with either researchers and their data.

## *Ad hoc* Production Processes

Not all digital libraries manage to establish Taylorist platforms and workflows. Smaller libraries, when they digitise materials at all, do so on a project-centered, *ad hoc* basis. This shares all the data pitfalls of project-based research processes. These libraries often depend on consortial or vendor-supplied technology platforms, sharply limiting the amount of digital expertise built within the library organisation.

### Unitary Organisations

Digital libraries tend to be self-contained organisational units, silos in both library and institutional contexts. Their public service staffing is minimal, limited to soliciting projects and perhaps marketing to end-users of digitised content. Many such units rarely interact directly with research faculty, though a few do solicit projects from them.

Such an organisational model cannot scale up to interacting with an entire campus full of researchers. It is also focused on being solely the endpoint for data; given the growing consensus that data curation must be addressed throughout the data lifecycle [8], this staffing model can have only a limited impact on solving the data stewardship problem.

## Characteristics of Institutional Repositories

Unlike digital libraries, institutional repositories were nominally created to accept all kinds of digital content or data. In practice, however, they were clearly optimised for standard research publications; data with different affordances and intended use fit only poorly and with difficulty. Other technical and organisational problems impede collection of research data by institutional repositories as well.

### Institutionally Bounded

The word 'institutional' in 'institutional repository' is no accident; it derives from the practice of certain journal publishers forbidding deposit of any version of a published article into disciplinary repositories such as arXiv or the Social Science Research Network, but permitting deposit into institutional repositories [9]. The catch, of course, is that any repository venturing beyond its own institution's borders risks losing its safe harbour.

Unfortunately, this sharp boundary limits how effectively institutional repositories can address research data problems. One classic backlog problem turns up when researchers leave an institution, leaving their Web presence and research data behind them. Since they are no longer affiliated with the institution, can the institutional repository intervene? Cross-institutional collaborations present similar difficulty; their data are liable to fall through the cracks because no institution's repository can comfortably take responsibility for them. Now that research no longer stops at institutional borders, institution-focused solutions will often prove inadequate.

### Optimised for Articles

Some institutional repositories say that they open their doors wide for any sort of useful digital material. The promise is partial at best; most repository software can only accept final, immutable materials. Again, this is an (arguably premature) optimisation for the favoured use case of a finished, final scholarly article or book chapter. For data, permitting only the immutable is unacceptable, as explained above; much of the value of data is precisely its mutability in the face of new evidence or new processes.

Deposit processes in many institutional repositories assume a limited number of files to deposit, such that they can be described and uploaded one at a time by a human being. Applying this manual process to datasets is like trying to empty the ocean with an eyedropper. The SWORD protocol holds potential to ameliorate this problem, but the protocol has not yet made its way into researcher or even library tools or processes.

Most repositories rely on Dublin Core metadata, largely because the OAI-PMH metadata exchange standard asserts unqualified Dublin Core as a minimum interoperability layer. Few repositories venture beyond qualified Dublin Core. Those who do, or wish to, find that much repository software can only manage key-value pairs. Now that many, if not most, metadata and exchange standards for research data use XML or RDF as a base, this limitation seriously vitiates repositories' ability to manage datasets.

### Cookie-cutter Look and Feel

Institutional repositories have been designed, insofar as they were designed at all, as institutional showcases for research publications. Their visual appearance tends to be university-corporate and sterile, if not library-amateurish owing to difficulty of customisation.

Unfortunately, their likeliest userbase, those academic staff who already enjoy online services and social interactions, are exposed to much more polished storage and service offerings from the likes of Flickr, SlideShare, and Google Docs. These tools also tend to be well-tailored to the content they seek, a much more difficult proposition for an institutional repository claiming to be all things to all types of data. In accepting everything, institutional repositories offer appropriate affordances—image lightboxes, page-turners, manipulation and remix tools—for almost nothing.

Worse still, most institutional repositories are self-contained silos, offering next to no way to access items programmatically via APIs, much less a mod- or plugin-friendly architecture. Not only do content-specific or discipline-specific tools not exist inside institutional repositories, such tools cannot even be built atop or alongside them!

### Inadequate Staffing

Few institutional repositories are fully embedded within their libraries, much less their institutions. Still distressingly common is the 'maverick manager' staffing model [10] based on the misconception that academic staff would willingly and en masse provide effort in the form of self-archiving. If one staff member cannot capture the institution's intellectual output, how can one staff member expand the repository's mission to capture research data, given the many and vexing difficulties caused by their variability and the tremendous amount of hand-holding and reformatting necessary to massage those data into acceptable form for sharing and archival?

Even those repositories with a somewhat larger staff will find research data a daunting challenge; most repository staff members, not themselves librarians, do Taylorist search, capture, and description of published work. Insofar as working with research data is anything but Taylorist, their skills will not transfer well.

## Ways Forward

Many of the mismatches between library technical infrastructures and the needs of researchers and their data can be resolved, given sufficient drive and resources.

### Flexible Storage and Metadata Architectures

End-to-end, soup-to-nuts silos, as many digital library and repository software packages are, cannot possibly meet the data challenge appropriately. Some low-level functionality is the same for all digital materials, to be sure; no more than one checksum/audit solution should be needed for any datastore, no matter how heterogeneous its content above the bits-and-bytes level. Still, most higher-level technologies need to be flexible in order to encompass the broadest possible variety of data and interactions.

Universities relying on vendor-hosted solutions such as Ex Libris's DigiTool or BePress face a special problem: they do not control the technology underlying their repository, and as the history of the integrated library system demonstrates, asking vendors (especially vendors who sense that their clients are locked into their platform) to bestir themselves to create new functionality is often a losing battle.

## De-coupling Ingest, Storage and Use

Ingest, storage, and end-user interfaces should be as loosely coupled as possible. Ideally, the same storage pool should be available to as many ingest mechanisms as researchers and their technology staff can dream up, and the items within should be usable within as many reuse, remix, and re-evaluation environments as the Web can produce.

Of the three main open-source institutional repository platforms, only Fedora Commons comes close to fulfilling this requirement. DSpace is a classic silo, and EPrints requires multiple software instances to accommodate differing interface needs. The trade-off, of course, is that Fedora Commons by itself does not offer end-to-end solutions, though projects such as Hydra [11] and Islandora [12] are beginning to fill the gaps. The key is that a Fedora repository running Islandora need not accept and disseminate materials only through Islandora's connection to a Drupal content-management system; any number of other linkages can be arranged behind the scenes.

Another fruitful approach is the 'curation microservices' stack at the California Digital Library [13]. Taking its cue from the UNIX philosophy of chaining small, discrete tools to manage complex processes, this system builds and deploys small, discrete, interoperable tools to manage separable segments of the data-curation problem. As individual tools 'wear out' or become obsolete, they can be redeveloped or replaced without breaking the rest of the system.

## APIs, Plugins, Mods

What makes flexibility technologically feasible, given that the small programmer complement in most libraries does not allow custom programming for every imaginable dataset or interaction, is the ease with which a data repository can be made to interact with the outside technology world. This means application programming interfaces (APIs) as well as plugin- and modification-friendly architectures.

Once again, Fedora Commons is the clear leader in open-source repository packages, boasting clearly documented and comprehensive APIs. DSpace users should be pleased that the platform is moving onto a Fedora base; the move should allow them to keep their existing workflows while vastly increasing their flexibility to build new ones.

## Versioning and De-accessioning

The ideal data repository leverages researcher inertia. The earlier in the research process data professionals and proper data management systems appear, the more likely it is that data emerge from research in appropriate form for sharing, reuse, and long-term preservation.

Therefore, versioning, change tracking, and rollback are vital elements of a good data repository. This is trickier than it sounds; change tracking is easy on a wiki, difficult in an XML file, perhaps impossible in a system based on proprietary instruments. Without this capacity, however, repositories are reduced to begging researchers for final versions once more, and researchers will have to exert themselves to comply.

Inertia suggests that a flexible storage repository intended for research data will be put to other uses, research-related and non-research-related. Over time, a great deal of junk is liable to build up, interfering with discovery and consuming storage space unnecessarily. Policies and technology infrastructure must permit the de-accessioning and removal of obvious cruft every so often; datasets should also be evaluated periodically for obsolescence both technological and intellectual.

## Standards and Interoperability

Data and metadata standards do not exist to meet many research data needs. Although interoperability-conscious approaches may reduce the cost of data interchange in a highly heterogeneous technology environment, additional standardisation is welcome and will reduce costs further. Would-be data curators need to remain aware of standards activities, both inside and outside large national and international standards bodies. Whenever possible, librarians should lend their metadata and digital preservation expertise to scientific standardisation activities.

Linked data deserves special mention here, not so much for its technical details as for the mindset of building data and metadata with the express intent of easy sharing and remixing. Libraries can no longer cling desperately to decrepit, arcane, inward-focused standards such as MARC, not if the ultimate goal is to be part of a great global sea of data. Instead, all descriptive efforts must have easy human- and machine-comprehensibility as a first-level goal, even when actual standardisation is out of reach due to data homogeneity or lack of appropriate standards.

## Code Sharing

The danger of flexibility, especially in the absence of standards, is recreating the Tower of Babel, mentioned above. Ten different clever ways of representing a page-scanned book are not nearly as valuable as one clever way applied by ten different book-scanning projects.

Historically, libraries have had a great deal of trouble sharing software code and communicating about technology-related solutions [14]. In the data realm, this is not acceptable, if indeed it ever was. Even collectively libraries barely have the technology and human resources to meet the research data challenge; how can it be done if libraries waste effort redundantly solving problems in parallel? Moreover, libraries poor in technological capacity will be left behind entirely if libraries that build solutions do not share them. This possibility is especially frightening for small science, many of whose practitioners do not work at major research institutions.

## Staffing and Funding Models

Although it is early days yet, patterns can be discerned in the experiences of libraries taking the plunge into research-data work. They generally begin by surveying local academic staff about their data and data-management practices. Having decided (inevitably) that help with data management is a genuine campus need, libraries then approach campus academic leaders for buy-in. They then launch pilot projects in one or more of several forms: building a repository for a specific kind of data, a discipline-agnostic consulting service (often intended to sustain itself via grant earmarks), or targeted involvement in specific research projects.

Any staffing and funding approach will face trade-offs in scalability, sustainability, and breadth of disciplinary coverage. Targeted interventions stress overburdened library staff less, but leave serious gaps in campus coverage. Grant-funded services may well be financially sustainable, but they threaten to leave unfunded disciplines without aid. Consulting services may be able to provide a base level of service to the entire campus, but that base level may be very low indeed (especially if disciplinary expertise elsewhere in the libraries is not available to the consultants), and financial sustainability is a serious concern.

Another likely outcome, particularly in wealthy Big Data projects, will be the embedded librarian, either hired specifically to help with data management or with data management one of several duties. Libraries hoping to fund a data curation programme with grant earmarks should take special note of this possibility, as it may drastically cut the number and wealth of grantees available to work with the library.

Institutional repositories boasting significant involvement by subject liaisons or bibliographers are best situated to take on data-related responsibilities. The potent combination of a technically adept repository manager with a discipline-savvy liaison can make headway on a substantial range of data problems. Maverick managers, however, will likely have to be satisfied doing the best consulting job they can, given their abject poverty of resource.

## Conclusion

None of the challenges presented herein should discourage librarians from engaging with the research data challenge. Our unique expertise in metadata, digital preservation, public service, and technology translation will serve researchers well, as will our sturdy common sense and the domain expertise of our subject librarians.

However, unless we proceed with clear understanding of researchers and their data, as well as our own systems and habits, we will simply trip over ourselves. Research data are too important, and our role in curating them at present too insecure, to allow that to happen.

## References

1. Attributed to Sayeed Choudhury by Palmer et al. "Center for Informatics Research in Science and Scholarship" 2007 http://groups.lis.illinois.edu/guest_lectures/showcase10/palmer.ppt , last visited 7 July 2010.

2. Heidorn, P. Bryan. Shedding light on the dark data in the long tail of science. Library Trends 57:2, Fall 2008, pp. 280-299.

3. Inter-University Consortium for Political and Social Research http://www.icpsr.umich.edu/icpsrweb/ICPSR/

4. International Virtual Observatory Alliance http://www.ivoa.net/

5. A collections or preservation policy that disallows Web sites would not be able to collect Web-based theses, e.g. http://exploringthehyper.net/ or http://digital.library.wisc.edu/1793/32316

6. "Taylorism." Encyclopædia Britannica Online. 29 July 2010 http://www.britannica.com/EBchecked/topic/1387100/Taylorism

7. Ben O'Steen discusses this problem with regard to databases in a post entitled "Handling tabular data" on his "Less Talk, More Code" weblog http://oxfordrepo.blogspot.com/2009/02/handling-tabular-data.html

8. See the DCC Curation Lifecycle Model http://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf

9. See, for example, Springer Verlag's policy http://www.sherpa.ac.uk/romeo/search.php?id=74

10. Salo, Dorothea. "Innkeeper at the Roach Motel." Library Trends 57:2, Fall 2008, pp. 98-123.

11. Hydra https://wiki.duraspace.org/display/hydra/The+Hydra+Project

12. Islandora http://islandora.ca/

13. Abrams, Stephen, John Kunze, and David Loy. "An Emergent Micro-Services Approach to Digital Curation Infrastructure." International Journal of Digital Curation 5:1, 2010, pp. 172-185.

14. Askey, Dale. "We Love Open Source Software. No, You Can't Have Our Code." code4lib journal 5, 2008. http://journal.code4lib.org/articles/527