REORGANIZING LIBRARIES FOR THE DATA CHALLENGE

Tirth Das
OIC, Librarian, Dr. Raheja Library, Central Arid Zone Research Institute, Jodhpur, Rajasthan, India, Email- tdas@cazri.res.in

Shivdan Singh Rajput
Senior Research Fellow, Dr. Raheja Library, Central Arid Zone Research Institute, Jodhpur, Rajasthan, India, Email: - sdclib0@gmail.com

Vimal Kishore Purohit
Technical Officer, Department of Computer science, Central Arid Zone Research Institute, Jodhpur, Rajasthan, India

Abstract

This article "Reorganizing Libraries for the Data Challenge" takes a broad view of the evolution of collecting data in a network environment and suggests some future directions based on various simple models. The authors look at the changing dynamics of print collections, at the greater engagement with research and learning behaviours, and at trends in scholarly communication. The goal is to provide context within which libraries can discuss changing patterns of investment across collection categories. The authors argue that the network is reconfiguring not only individual research libraries but also the whole library system, as reduced transaction costs facilitate the unbundling of functions and their consolidation in network platforms and with other external service providers. This article derives from a review of key data challenge confronted by reorganizing libraries that are actively investing in online collections and services. Con- ducted in the first instance to help refine the programmatic goals of the digital library federation (dlf), it took account of the digital library developments, successes, needs, and challenges perceived by profession.

Introduction

Whether managing research data is the new special collections, 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 decreasing budgets and staffs, library administrators may well be interested to repurpose existing technology infrastructure and staff to address the data curtain 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.

Data curation indicates management activities required to maintain research data long-term such that it is available for reuse and preservation. In science, data curation may indicate the process of extraction of important information from scientific texts, such 3S research articles by experts. To be converted into an electronic format. According to the University of Illinois' Graduate School of Library and Information Science, "Data curation is the active and on-going management of data through its life cycle of interest and usefulness to scholarship, science, and education. Data curation activities enable data discovery and retrieval, maintain its quality, add value, and provide for re-use over time, and this new field includes authentication, archiving, management, preservation, retrieval, and representation."

Characteristics of Research Data

Size and. Scope of Data

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 utter terror in the practical librarian or technologist. Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

'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 hyper local research focus. Though each small-data producer produces only a trickle of data compared to Big Data, the tens of thousands of small-data producers in aggregate may well produce as much data as their Big Data counterparts. Securely and consistently storing and auditing this amount of data is a severe challenge. The increasing 'small data' store means that institutions without local Big Data projects are by no means excused 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 an acted 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, hugely more consultations and consultants are necessary to bring practices and the resulting data to an adequate standard.

Inconsistency

Digital research data come in every presumable 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 changing 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, in appropriate 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 practical nor supportable.

A major consequence of the diversity of forms and formats of digital research data is an associated 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 (A manuscript page from a scroll or book from which the text has been scraped or washed off and which can be used again). These varying affordances (A quality of an object, or an environment, which allows an individual to perform an action) must be respected by spreading 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 fewer usable in their raw state than after they have had filters or algorithms or other processing performed on them. Others welcome correction, or are distinguished by comparison with other datasets. Two repercussions emerge: first, that planning and acting for data stewardship (The careful and responsible management of something entrusted to one's care) must take place throughout the research process, rather than being an add- on at the end; and second, that digital conservation systems designed to steward only final, unchanging materials can only fail faced with real-world datasets and data-use practices.

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

Bottleneck

Libraries are not starting with a clean schedule 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 an extensive amount of data already. These data, particularly 'small data,' tend to be disorganised, poorly defined if described at all, and in formats poorly suitable to long-term reprocess. Even more lnappropriatev researchers have become accustomed to the processes that produce these disordered data, which makes them liable to resist changing those processes to improve data practicability.

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

Project Direction

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 destroys incentives toward good data practices; why save data if the next project will be on a different theme with different collaborators? Institutional memory and silent 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 acne 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 removed unorganised and undescribed into dirty digital closets.

Unusual Data and Data Formats

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

A few disciplines have created data standards, usually because of a strong centralised data warehouse that imposes those standards on researchers wishing to contribute data. In most disciplines, however, and certainly those where research is individualised or hyper local, incentives to create, much less follow, data standards are minimal or non-existent. The resulting messy Tower of Babel spoils both reuse and long-term stewardship.

Characteristics of Digital libraries

Curated

Because digitisation to library standards is expensive, materials in digital libraries are chosen and handled with enormous care. High standards succeed in digitisation quality, in associated metadata (A set of data that describes and gives information about other data), and in presentation. Only materials deemed important enough to warrant such care are digitised at all.

These mind sets 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 researcher's data practices; not a few librarians will simply assume that data cannot be worth curating if researcher themselves take so little care of them. Digital librarians also rebel at the

idea of researchers' untidy 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 whatever criteria are chosen, will those criteria conflict with institution-wide mandates such as the preservation of theses and dissertations with their accompanying materials.

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, it means exactly the sort of habit, minimum-effort, and minimum-judgment workflows known as 'Taylorist' after American manufacturing efficiency expert Frederick Taylor. The variability of research data defeats Taylorist processes totally. 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 skill 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 shoe horned (Force into an inadequate space) into a digital library not optimised for them, but only at the cost of the affordances surrounding them. Consider data over which a composite 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 negligible job of restoring 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.

Finality, Tavlorist 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 junction 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 both 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-cantered, ad hoc basis. This shares all the data pitfalls of project-based research processes. These libraries often depend on associations 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, storage tower 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 end point for data; given the growing consensus that data curation must be addressed throughout the data lifecycle.

Characteristics of Institutional Warehouses (Repositories)

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

Institutionally Confined

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 arrive or the Social Science Research Network, but permitting deposit into institutional repositories. The catch, of course, is that any repository submitting beyond its own institution's borders risks losing its safe port.

Inappropriate 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. 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 res u Its will often prove inadequate.

Optimised for Articles

Various 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, absolute materials. Deposit processes in many institutional repositories adopt 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 amend 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 restriction 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 platforms for research publications. Unfortunately, their probable user base, those academic staff who already enjoy on line services and social interactions, are exposed to much more polished storage and service offerings from the likes of Flickr, Slide Share, and Google Docs. These tools also tend to be well-personalised to the content they seek, a much more difficult scheme for an institutional repository claiming to be all things to all types of data. In accepting everything, institutional repositories offer appropriate affordances-image light boxes, page-turners, manipulation and remix tools-for almost nothing.

Inadequate Staffing

Few institutional repositories are fully inserted within their libraries, much less their institutions. Still distressingly common is the 'maverick manager' staffing model based on the misconception that academic staff would freely 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 annoying 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 frightening challenge; most repository staff members, not themselves librarians, do Taylorist search, capture, and description of published work.

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 Designs

End-to-end, soup-ta-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 data store, and no matter how heterogeneous it's content above the bits-and-bytes level. Still, most high-level technologies need to be flexible in order to incorporate 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 rose 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 accessible 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. 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 micro services' stack at the California Digital Library. Taking its indication from the UNIX philosophy of chaining small, discrete tools to manage complex processes, this sys-tem 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.

APls, Plugins, and Mods

What makes flexibility technologically viable, given that the small programmer complement in most libraries does not allow custom programming for every imaginable dataset or interface is the ease with which a data repository can be made to interact with the outside technology world. This means application programming interfaces (APls) 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 APls.

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 occur 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 submit

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 (computing jargon for "code, data, or software of poor quality") every so often; datasets should also be evaluated periodically for oldness 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 setting 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 mind-set of building data and metadata with the rapid intent of easy sharing and remixing. Libraries can no longer stick desperately to decaying, secret, 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-clarity as a first-level goal, even

when actual standardisation is out of reach due to data equality 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. Historically, libraries have had a great deal of trouble sharing software code and communicating about technology-related solutions. In the data jurisdiction, 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. Moreover, libraries poor in technological capacity will be left behind entirely if libraries that build solutions do not share them. This possibility is especially fearsome for small science, many of whose experts do not work at major research institutions.

Staffing and Funding Models

Although it is early days yet, patterns can be separated in the experiences of libraries taking the dive into research-data work.

They generally begin by charting 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 adjustments in scalability, sustainability, and breadth of disciplinary coverage. Targeted involvements stress overburdened library staff less, but leave serious gaps in campus exposure. 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 employed specifically to help with data management or with data management one of several duties. Libraries hoping to fund a data curaticn programme with grant earmarks should take special note of this possibility, as it may severely cut the number and wealth of grantees available to work with the library.

Institutional repositories boasting significant involvement by subject connections or bibliographers are best situated to take on data-related tasks. The potent combination of a technically proficient repository manager with a discipline-savvy association 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. Unique expertise in metadata, digital preservation, public service, and technology translation will serve researchers well 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. http://www.slideshare.net/lljohnston/leslie-johnston-bigdata-at-libraries
2. http://technet.microsoft.com/enus/magazine/hh987046.asx
3. Carnegie Foundation for the Advancement of Teaching. "Carnegie Classifications. "
4. http://classifications.carnegiefoundation.org/methodology/basic.php
5. Gomm, Roger, Martyn Hammersley, and Peter Foster, eds. Case Study Method. SAGE Publications Ltd., 2009. doi: 10.4135/9780857024367.
6. Persaud, Nadini. "Pilot Study." In Encyclopedia of Research Design, edited by Neil J. Salkind, 1033-34. Thousand Oaks, CA: SAGE Publications, Inc., 2010. doi: 10.4135/9781412961288.n312
7. Astroff .Roberta J. "Reorganizing for the Distributed Library"http://www.ala.org/acrl/sites/ala.org.acrl/files/content/conferences/confsandpreconfs/2013/papers/Astroff_Reorganizing.pdf
8. Geleijnse, Hans "Reorganizing the Library for the Future" http://arno.uvt.nl/show.cgi?fid=80688
9. Anderson, R. (2008), "Future proofing the library: Strategies for acquisitions, cataloging, and collection development,"The Serials Librarian, Vol. 55 No. 4,pp. 560-567.Availableat http://dx.doi.org/10.1080/03615260802399908
10. Association of Research Libraries (2010), "Envisioning research library futures: A scenario thinking project," available at:http://www.arl.org/rtl/plan/scenarios/usersguide/index.shtml(accessed 31 October 2011).
11. Barclay, D.A. (2007), "Creating an academic library for the twenty-first century," New Directions for Higher Education, No. 139, pp. 103-114
12. Becker, L. K. W. (2006), "Globalisation and Internationalisation: Models and Patterns of Change for Australian Academic Librarians," Australian Academic &Research Libraries, Vol. 37 No. 4, pp. 282-298.
13. Cohen, P. (2010), "Fending off digital decay, bit by bit, "The New York Times, 16 March, pp. C1. Available at http://www.nytimes.com/2010/03/16/books/16archive.html?_r=1

14. Darnton, R. (2008), "The library in the new age, "The New York Review of Books, 12 June. Available at http://www.nybooks.com/articles/archives/2008/jun/12/the-Library-in-the-new-age/19
15. Fitch, D.K., Thomason, J., and Wells, E. C. (1993), "Turning the library upside down: Reorganization using total quality management principles," Journal of Academic Librarianship, Vol. 19 No. 5, pp. 294-299
16. Franklin, B. (2009), "Aligning library strategy and structure with the campus academic plan: A case study, "Journal of Library Administration, Vol. 49, pp. 495-505.
17. Hardy, L. (2010), "The future of libraries, "American School Board Journal, Vol. 197 No. 1, pp. 22-26.
18. Haynes, E. (2010), "The class of 2012: How will we meet their needs and expectations?"Library Media Connection, Vol. 28 No. 4, pp. 10-11
19. .Kellogg, C. (2009),"What will the library of the future look like?"Los Angeles Times, 11 February. Available athttp://latimesblogs.latimes.com/jacketcopy/2009/02/what-will-the-l.html
20. Kolowich, S. (2009),"Libraries of the future," Inside Higher Ed, 24 September, available at http://www.insidehighered.com/news/2009/09/24/libraries.
21. Pritchard, S. M. (2008), "Deconstructing the library: Reconceptualising collections, spaces, and services", Journal of Library Administration, Vol. 48 No. 2, pp. 219-233
22. Schonfeld, R. C. and Housewright, R. (2010), "Faculty survey 2009: Key strategic insights for libraries, publishers, and societies, "available at: http://www.ithaka.org/ithaka-s-r/research/faculty-surveys-2000-2009/Faculty%20Study%202009.pdf
23. Staley, D. J. and Malenfant, K. (2010), "Futures thinking for academic librarians: Higher education in 2025," available at: http://www.ala.org/ala/mgrps/divs/acrl/issues/value/futures2025.pd