Ricky Erway
Brian Lavoie

OCLC Research

The Economics of Data Integrity

Note: This is a pre-print version of a paper published in *Information Professionals 2050: Educational Possibilities and Pathways*. Please cite the published version; a suggested citation appears below. Correspondence about the article may be sent to erwayr@oclc.org.

Abstract
This brief paper is in response to a call for papers for the UNC/NSF Curating for Quality workshop. We describe the aspects of costs and sustainability of data curation as they pertain to data integrity.

© 2012 OCLC Online Computer Library, Inc.
6565 Kilgour Place, Dublin, Ohio 43017-3395 USA
http://www.oclc.org/

Reuse permitted consistent with terms of the Creative Commons Attribution--Noncommercial 3.0 license (CC BY-NC): http://creativecommons.org/licenses/by-nc/3.0/.

Suggested citation:
Erway, Ricky, and Brian Lavoie. 2012. "The Economics of Data Integrity." *Curating for Quality: Ensuring Data Quality to Enable New Science.* Pre-print available online at: http://www.oclc.org/research/publications/library/2012/library/erway-dataintegrity.pdf.

INTRODUCTION

It is difficult to consider sustainability of data quality without defining what data quality entails.

To data creators and to those who reuse the data, data quality may refer to ensuring accuracy of the data and supplementing the data with rich description, ancillary materials, context, or other enhancements that facilitate leveraging value from the data. These aspects are addressed in other workshop papers.

This paper addresses data quality in the traditional library or archives sense: ensuring data quality consists of making sure the data is uncorrupted, is what it purports to be, and that it persists and is accessible into the future. This includes technical aspects (are the processes adequate to preserve the data?), social aspects (can we trust that the processes will be followed reliably?), and economic aspects (is there adequate ongoing funding to preserve the data into the future?). In a library context, these issues generally fall within the scope of long-term digital preservation. This paper advocates for achieving sustainability through economies of scale made possible by collaboration.

The final report of the Blue Ribbon Task Force on Sustainable Preservation and Access identifies five conditions for sustainable digital preservation[1]:

- recognition of the benefits of preservation by decision makers;
- a process for selecting digital materials with long-term value;
- incentives for decision makers to preserve in the public interest;
- appropriate organization and governance of digital preservation activities; and
- mechanisms to secure an ongoing, efficient allocation of resources to digital preservation activities.

The focus of this paper is on this fifth condition, mechanisms to secure an ongoing, efficient allocation of resources to digital preservation activities.

"… ALLOCATION OF RESOURCES…"

In allocating resources to digital preservation, we are essentially creating mechanisms to "transfer funding and other resources from those who benefit from and are willing to pay for digital preservation, to those who are willing to provide preservation services."[2] Sometimes these stakeholders groups are one and the same; in more complicated

situations, they are distinct. In either case, the fundamental condition is clear: there must be recognition that allocating resources to ensure the long-term future of quality digital assets is a desirable, indeed necessary, activity. Without this recognition, the goal of maintaining long-term access to quality data is not achievable.

Allocating resources to digital preservation involves some key trade-offs that should be recognized upfront. One is that investing significant resources in curating high quality datasets could detract funds from producing new datasets via new research. Most funders are interested primarily in new research, creating a significant obstacle to preserving existing high quality datasets over the long term. The data management plans required by NSF and other grant-giving entities help to balance the two needs. One must also weigh the cost of preservation against the costs of replacing the data—and monitor when that balance tips in either direction. For example, for data produced through computer simulation models, it may be less costly to store the algorithms that produced the dataset than the data itself, thus preserving the option to re-create the data at a future time. Another possible trade-off is between sustainability and access. Generating a flow of funds to support long-term preservation may require charging a fee for access, which inevitably limits the scope for potential reuse to those able and willing to pay. Providing dark archive service is less expensive, but it has been shown that access has a positive effect on strengthening the incentive to preserve and also provides a monitoring function to alert the curator to changes to the data or to the need to migrate data to another format.

Another trade-off is the one between risk and reward. There are a variety of risks impacting the long-term sustainability of research data curation, from unexpected data loss to uncertainty about the value to future users. In allocating resources to the preservation of high-quality data, we are making a "bet" on realizing some future value from that investment (new scholarship, replicability of scientific findings, etc.). Only time will tell if efforts to preserve a particular dataset prove to be a wise allocation of resources. Decision-makers need to be prepared to revisit their preservation decisions frequently over time, and adjust their resource allocations accordingly.

Management of the trade-offs associated with preservation decision-making should be informed by the range of risks undertaken by service providers that could potentially impact the preservation process over time. Many times funding is available to establish a data archiving service, but is not available for ongoing operations. Likewise, funding may be included in a grant proposal to cover the costs of preparing a particular dataset for deposit in the archive, but not for its long-term care. Another challenge for service providers is that technology and preservation practices change over time, requiring acquisition of new hardware and software and reworking of processes. One of the best reasons to commit to preservation of a particular dataset is when the data is not reproducible. In these cases, the ramifications of failure are high.

CALCULATING COSTS

Before addressing where funding for data curation might come from, we need to have a sense of what the actual costs are. However each situation is unique and there are no set

answers. For example, economies of scale can have a dramatic effect on the per-unit cost of preservation.

The Keeping Research Data Safe (KRDS) framework[3] provides a way to calculate the costs of data curation for a specific situation. Costs can be broken down by the activities in which they are incurred; these costs can then be adjusted to reflect the particular conditions of a given digital archiving scenario (service adjustments), and appropriately distributed over time (economic adjustments). For example, preservation activities can differ in how long the data is to be maintained; the type of online, near-line, and offline storage used; security requirements of the data; frequency of refreshment; and the number of versions, editions, or copies to be kept. Service limitations can control costs, e.g., limiting the file formats accepted or insisting on a standard IPR statement. Moreover, costs can spread over time via inflation and depreciation.

Though the results are very individual, case studies using the KRDS cost framework have turned up some indicative findings. In general:

- The costs involved in the generation of the data during research, far outweigh the costs of archiving the data.
- Archiving costs are highest up front and become less significant over time. This is true for the archive overall (set-up costs are far higher than operational costs) and for each dataset (the ingest process is more costly than the ongoing maintenance costs).
- Use of off-the-shelf software and hardware solutions brings costs down significantly.
- Initial capital costs of storage media and systems are less than a third of the overall costs of ownership.
- Staff costs exceed those of any other component. In academic institutions, staff costs range from 50% - 90% of total costs. The degree to which processes are automated can have a significant impact.
- The number of depositors can affect costs. One or more middlemen, aggregating submissions and ingesting them in a standard manner, will mitigate the high costs associated with a large number of depositors.
- Changes in workload can have substantial effects on unit costs. In one case, when workload increased 600%, costs increased only 325%.
- Timing can be a factor. For example, addressing data migration early on is much cheaper than attempting to migrate from an already obsolete format.
- In some cases, the costs of deaccessioning a dataset exceed those of continuing to maintain it.

"… ONGOING AND EFFICIENT …"

Ensuring economically sustainable datasets (and their associated services) goes far beyond simply allocating resources. It also involves using those resources efficiently and leveraging collaboration to achieve economies of scale. Furthermore, preservation is an ongoing process, so the flow of funds to preservation activities must also be ongoing if long-term preservation objectives are to be achieved.

The requirement that the allocation of resources to digital preservation needs to be ongoing over time seems obvious, yet it is too frequently neglected in practice. It is easy to find examples of long-term preservation projects that are funded through short-term, one-off grants. When the funding runs out, the preservation activity must scramble to find another grant or other resources to keep the project running for a while longer; alternatively, the project simply ends. An example of such a situation is the UK Arts and Humanities Data Service, which had been funded by JISC .[4]

Just as preservation activities require a long-term view of the maintenance of, and access to, data assets, so too do they require a long-term view of their funding. Mechanisms that secure a reliable, ongoing flow of resources are the optimal way to fund long-term activities such as digital preservation.

Efficient use of available resources is another necessary aspect of sustainable digital preservation. Economies of scale argue for collaboration, leveraging fixed costs over a larger number of deposits. Data curators from the library, archive, and information technology sectors can ingest and preserve datasets. But ensuring the *quality* of data requires specific subject area expertise, due to varying needs of the disciplines. This is perhaps an even more significant opportunity for economies of scale. If every repository had to have a wide range of subject experts, the costs would be prohibitive.

In some university settings, data is curated in the department of origin. Here the benefit is the proximity of the researchers and others who understand and might use the data. In this case, however, the technical curation skills may be lacking.

A collaborative approach allows for a pool of subject specialists that serve a wide range of depositors and draw on a pool of people with experience in various technical aspects of data curation. An example is the Interuniversity Consortium for Political and Social Research (ICPSR),[5] which hosts research data files in the social sciences on behalf of 700 institutions. Staff with specialties in various fields work closely with researchers to prepare data for submission and ensure data integrity, while staff with technical skills are tasked with the preservation component.

An informal network of subject-based data repositories allows subject specialization at each repository. A related example is the array of disciplinary repositories for research preprints and published articles. Aggregating dataset deposits for a particular discipline not only allows for specialized help for ingest and ensuring data quality, but it also allows

for aggregation of users. A single set of functionality and support services can meet the needs of researchers in that particular field.

Specializing in a narrow discipline can encourage compartmentalization. A benefit to the ICPSR approach is that it encourages cross-disciplinary research.

In some countries a national approach is taken, as with the DANS service in the Netherlands.[6] In the US, it is less likely that we would have a single national service, but a national *network* of data archives would help with discovery of relevant datasets for reuse and would facilitate multidisciplinary research. Additionally, a central infrastructure that provides support for locally-curated datasets might help in disciplines that aren't as well-funded as some of the big sciences.

No matter what approach is taken, preservation planners need to be cognizant of opportunities to lower the per-unit cost of preservation by spreading costs over higher volumes of preservation activity. Digital preservation is a shared problem, and shared problems often lend themselves to shared solutions through collaboration.

SUSTAINABLE BUSINESS MODELS

There are a number of ways to supplement start-up funding and to provide ongoing financial support. The report, *Lasting Impact: Sustainability of Disciplinary Repositories*,[7] identifies several different business models for disciplinary document repositories:

- Institutional support
- Use-based institutional contributions
- Support via consortium dues
- Distributed network of volunteers
- Federal government funding
- Decentralized arrangement
- Commercial "freemium" service (basic access is free; value-added services for fee)

The report notes that, in most cases, a combination of funding sources is used. These funding models can equally apply to data repositories. ICPSR, for example, is supported by member fees, use fees, and grants. [A more thorough listing of 155 repositories and their funding models is available from DataCite.[8]]

RECOMMENDATIONS

Because data curation involves many invested providers and beneficiaries—and many of those involved have both roles—there is much potential for addressing the challenges. The following are recommendations for optimizing for sustainability.

- Have a discipline-specific entity in between the researchers and the repository to help with setting policy regarding aspects such as selection criteria, retention periods, and transfers of stewardship.
- Use aggregators to work with depositors to normalize their submissions prior to ingest.
- Due to high degrees of change and uncertainty, agreements between content providers and repositories should include options to review, renew, refine, or terminate.
- Funders should consider providing ongoing support for trustworthy data archives and encourage automation developments that will decrease the number of manual processes.
- Institutions, funders, and publishers should impose and enforce meaningful mandates.
- The academy should recognize datasets as first-class scientific contributions in academic credentials to provide a personal incentive for researchers to prepare and submit their data for archiving.
- Because so much is in flux, we include this final counsel from the Blue Ribbon Task Force: "Hedging against uncertainties, postponing decisions when possible, recognizing that benefits, demand, and users will change, anticipating better information over time—these are the habits of mind that mark responsible digital stewardship and will help husband scarce resources while creating enough flexibility for bold moves and rescue of endangered assets when that becomes necessary."[9]

CONCLUSION

All academic institutions have or will have a need for some sort of data curation, but it is unrealistic to think that every institution will establish local data curation capacity. Due to the need for specialization in each subject, the need for a range of curation skills, the risks undertaken, and the economies of scale, it is unwise to attempt to replicate a broad range of data curation services, infrastructure, and expertise at every institution. Institutions so inclined might specialize in a particular field and offer services to all researchers in that discipline. Scholarly societies, government agencies, and commercial entities might take on similar roles. Consolidated solutions, where systems, infrastructure, and expertise can be spread over higher volumes of curation activity, offer lower per-unit costs. From the access perspective, specialized data repositories can focus on the needs particular to those who may want to reuse those datasets. Taking it up one more level, having broad discipline coverage, like ICPSR, or aggregated access to datasets at specialized repositories, facilitates interdisciplinary research. When these services are raised to the network level, expertise, economies, and benefits are shared.

REFERENCES

[1] Blue Ribbon Task Force on Sustainable Digital Preservation and Access. February 2010. *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information; Final Report of the Blue Ribbon Task Force on Sustainable Preservation and Access.* p. 12 http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.

[2] Lavoie, Brian F. 2012. Sustainable research data. In *Managing Research Data*, G. Pryor, Ed. Facet Publishing, London. p. 70.

[3] Beagrie, Neil., Julia Chruszcz, and Brian Lavoie. April 2008. *Keeping Research Data Safe: a cost model and guidance for UK universities, Final Report.* http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf.

[4] Arts and Humanities Data Service http://www.ahds.ac.uk/.

[5] Inter-university Consortium for Political and Social Research www.icpsr.umich.edu/.

[6] Data Archiving and Networked Services http://www.dans.knaw.nl/en.

[7] Erway, Ricky. 2012. Lasting Impact: Sustainability of Disciplinary Repositories. OCLC Research, Dublin, Ohio. http://www.oclc.org/research/publications/library/2012/2012-03.pdf.

[8] DataCite http://datacite.org/repolist.

[9] BRTF p. 47