Big Data in the Campus Landscape

Curation

ECAR Working Group Paper | November 20, 2015

This paper is a publication of the ECAR Campus Cyberinfrastructure (ECAR-CCI) Working Group. It is the final paper in a four-part series focusing on Big Data in the Campus Landscape. ECAR working groups bring together higher education IT leaders to address core technology challenges. Individuals at EDUCAUSE member institutions are invited to collaborate on projects that advance emerging technologies important to colleges and universities. ECAR-CCI helps educational institutions develop institutional strategies and plan resource deployment and helps users harness and optimize the power and capabilities of these new integrated IT tools and systems for educational and research applications in higher education.

Complete Series List

1. Laying the Groundwork
2. Basic Infrastructure Support
3. Security and Privacy
4. Curation

Introduction

Data-curation challenges are not unique to big data, but the particular characteristics of large data imply both a sizeable investment of resources and some particular technical complexities and approaches. Data curation is the process of ensuring that data can be understood and reused by interested parties across disciplines, organizations, and the passage of time. Curating data also implies making choices about where to invest limited resources and understanding likely needs for the data, ranging from experimental reproducibility to genuine repurposing.[1] It subsumes preservation but goes beyond it. Stakeholders in the process include scholars, librarians, IT staff, funders, and policy makers.

As institutions develop services to support data-intensive scholarship and to meet funder requirements to facilitate the reuse of research data, one of the ideas that seems to be causing the greatest confusion is that of data curation, in part because various groups use it to cover such a wide range of disparate activities that are part of the data life cycle within scholarly work.[2] Current curation practices may need to be extended to accommodate research big data, but it is clear that fully realized data curation is not cheap and that much of the investment takes place early in the data life cycle. Indeed, in many situations, the argument for curation is speculative because we know so little about which data will see enough significant and impactful reuse—or about the nature of the reuse—to justify the cost of curation. Finally, it is important to recognize that relatively little of the research data produced appears currently to be curated, though the measures of this are complex and a bit speculative.[3] Numerous developments are coming together to drive expansion in the scale of data curation. Among them are research funder requirements for data management plans; shifting disciplinary norms about reproducibility and replicability of research and about data sharing more broadly; mandates (most notably from federal funders in the United States) to make research outcomes, including data, available to the public;[4] and most recently, increasing attention to monitoring compliance with these mandates. As the scale of the curation effort expands, it seems clear that there will be growing attention to controlling costs.

In this paper, we will examine some current data-curation activities, discuss why they need to be extended to research big data, and, most importantly, focus on why they are important and what purposes they are intended to serve. The emphasis here is not on technical details, specific disciplinary standards, or best practices but rather on the contributions and limitations of various roles and activities during the data life cycle, particularly as they apply to research big data.

Curation and Big Research Data

Over time, some data or metadata formats become obsolete and may be supplanted by others. Curatorial issues come to the fore when transformations from old to new are not invertible, when heuristics or renewed human intervention are needed to populate data elements in the new formats, or when issues such as migration from one lossy compression format to another are involved. Decisions need to be made about whether both old and new versions should be saved and what transformations should be made (e.g., only when data are being retrieved for reuse, as opposed to use in archival copies). If the data are very large, there are potentially serious economic and curatorial impacts of such choices.

Other activities are also called data curation. Some important ones, conducted primarily by scholars, are editorial (most often consensus based) and ongoing. These include aggregating and annotating data from multiple sources on a continuing basis to provide a maintained view of the "current best knowledge" or understanding of a phenomenon. Often these activities are carried out in the context of a complex, discipline-specific scientific information system rather than a simple repository. Examples include GenBank, the RCSB Protein Data Bank (PDB), OMIM, databases of papyrus fragments,[5] medieval manuscripts,[6] witness testimonies of survivors of the Holocaust and other genocides, and various archeology reconstructions. The historical antecedents here are specialized scholarly encyclopedias and dictionaries. As long as these important scientific information systems continue to be funded, the data they contain will be very well cared for and readily accessible to the relevant scholarly

Libraries and Data Description

Libraries have been dealing with the difficulty of describing content for centuries and will continue to be a key player in data curation. However, there are a few inherent difficulties regarding data description. First, for very broad and heterogeneous collections, high-precision description (as might be used by an expert in a particular subfield) is very difficult to offer. Second, descriptive terminology shifts and evolves over time with the growth of knowledge, so description needs to be maintained over time to facilitate discovery (while first-level data curation—simply describing the data within a static knowledge context—should not need as much maintenance).

The relative roles of library staff and researchers are still an open question. Precise description of what is in a data set or database in order to enable reuse will often require deep knowledge of the research methods used to collect the data (the contexts and methodologies of observations, measurement, etc.) and a willingness to fully document this information. There are also more general descriptive activities needed to document the data's provenance and to help with discovery; these are more heavily based in information management and systems and less in specific disciplinary expertise. Ideally, teams will have research-level expertise in both the relevant discipline and the information sciences. In some areas (e.g., biomedical work) such dual expertise is not uncommon, but it's doubtful this will scale across the full range of academic disciplines.

communities. However, we have very limited experience with what should be done when funding for a system of this sort runs out, and hard choices will need to be made about what data are kept going forward, what level of curation (and access) will be provided for the data on an ongoing basis, who will accept responsibility for what data, who will foot the bill, and how the transition will be managed.[7]

Two Realms of Big Data

As should be clear, data curation is a challenge that applies to all kinds of research data, whether massive data sets or a few megabytes of spreadsheet data. It is worth briefly considering how big data may be different in shifting balances or introducing new considerations. To do so, it is useful to split the world of big data into two parts. The first is the domain of very-large-scale observational or experimental programs, such as the Large Hadron Collider at CERN and the various synoptic sky surveys like the Large Synoptic Survey Telescope (LSST), the National Ecological Observatory Network (NEON), and planetary missions. These are big, costly projects that need to integrate data management and curation into their fundamental planning and typically include dedicated, high-level IT and information management staff in their planning and operations.[8] Often a substantial part of the project budget is earmarked for managing the data that come out of the project. Failure to sufficiently consider data curation in these settings represents a substantial management failure. Thoughtful and sophisticated integration often exists between data (and metadata) generation strategies, community analysis of the data, and ongoing curation and preservation. In most cases, campus units will get involved in these very large programs only in an explicit support role; the project leaders will also lead in the development of data-curation strategies and will have the resources and access to expertise to do so effectively.

The second part of the world of big data involves individual principal investigators (PIs) or moderate-sized labs, where access to expertise both in IT generally and in data management and curation in particular may be quite limited and where there is much less direct community involvement and accountability (other than, for instance, review of a data-management plan as part of the evaluation of a grant proposal). As part of their work, such researchers can amass very large data sets (including, for example, databases of very-high-resolution images or video) and will need help from campus (or perhaps national or international disciplinary-oriented) services in designing and implementing curation (and preservation) strategies for these materials.

Considerable thought and analysis may be required to define what metadata and other documentation are needed for a given collection of data, yet this may have to be a relatively local set of decisions, as opposed to a broad and well-vetted community consensus typical for large-scale community projects. Automating the collection of metadata and/or the association of that metadata with data may not be easy to accomplish when using commercial (as opposed to custom-designed) lab equipment and related workflows. Additionally, the necessary metadata may be defined too late in the research project planning and design cycle to be fully incorporated into research workflows. Also important to note is that these smaller producers of large data may have limited IT data-management resources and will also need help with underlying data-preservation and security functions such as making and verifying copies and geographically distributed replicas, migrating or relocating copies, and the like.

Finally, curation decisions that expand the size of the data will have substantial financial consequences in a way that is not true for smaller data. For example, decisions to add a transcoded or reformatted copy of a very large data collection (consider video as a case in point) while still retaining the original—which would be the normal approach if the reformatting is not lossless and invertible—implies a substantial expense, both for the computational processing and for the subsequent additional storage.

Data Curation Is More than Data Preservation

One of the greatest sources of confusion is the relationship between data curation and data preservation. At the most basic level, preservation is ensuring that collections of bits will be accurately carried into the future: This is the world of multiple, geographically distributed storage facilities; of fixity checks and error-correcting codes; and of copying bits from old media to new media on a regular basis. Large IT organizations have a good deal of expertise in this area, though they may not be as aggressive about repeated fixity checks as those responsible for ensuring preservation might like. Thus, although data preservation is challenging in itself—and particularly so for large data sets that stress storage and network resources—in this paper we assume that this challenge is recognized and will not pursue it further other than to note that the risk analysis and related economic tradeoffs are very different (and not well understood) for very large data sets that need to be retained for many decades; the cost of additional replicas is very high, and the threat and failure models and statistics over long periods of time are highly speculative.[9]

In contrast, at the most basic level, data curation can be viewed as ensuring that—among the information included in the bits carried forward—there is enough readily interpretable description and documentation (i.e., metadata) for someone in the future to understand what the underlying data are, how they are encoded, how they were collected, where they came from, who collected them and when, and how they should be interpreted. Some of this information may be textual; some parts may be highly structured and machine processable.[10] The descriptive and contextual information may also indicate constraints on use (for example, if the data set deals with information about human subjects).[11] Ensuring the ability to reuse data over time also requires careful consideration of how the data might depend on other materials—such as software, ontologies, instruments, and calibration—and, where necessary, how these dependencies can be managed. Note that over long periods of time the ability to reuse data depends much more on getting the curation documentation right; a researcher reusing a five-year-old data set has a good chance of being able to contact the data creator if the documentation is unclear, but a researcher trying to reuse data across half a century is most likely on his own.

The most commonplace view of this part of the curation process is that once documentation and analysis are complete—resulting in a deposit package of data, metadata, and perhaps other materials—and the package is placed into a preservation environment where bits are subsequently migrated and preserved, little else needs to be done. This attitude holds that data curators are mostly involved in preparing materials for deposit into preservation repositories or scientific information management systems. This perception is likely to change in the future as data curators undertake additional responsibilities, such as:

- Ongoing consideration of formats used to store data and metadata and decisions made about whether they should be migrated to newer formats while there exists a window in the software ecology in which older and newer formats coexist and migration tools are available
- Reassessment of external dependencies as these dependencies shift and are better understood
- Resource allocation and collection decisions about whether it remains important to retain and curate the data in question and whether stewardship and curation responsibilities should be reassigned from one organization to another[12]

Data Curation and Discovery

An important, second level of data curation deals with discovery. If you think of a data object that has been curated at the basic level just described, a researcher should have enough information to conduct a detailed assessment of the suitability of the data for various types of reuse and to actually use the data in those ways. But this supposes that the researcher already knows roughly what the data object is about and that it is likely to be of interest; at best, discovery might use relatively simple attributes, such as the name of the investigator that deposited it or the grant under which it was created.[13] When one considers distributed systems of repositories housing millions of data sets or data objects and imagines a long-term future where data sets may be discovered for reuse in contexts quite distant from those of their original creation, it is clear that, to maximize reuse, various forms of descriptive cataloging of the data objects will be necessary, as well as the development of various discipline-specific and cross-disciplinary search engines.[14] Note also that as underlying data become more richly connected to the research literature (through data-citation practices, for example, or new forms of publication), study and search of the research literature also becomes a discovery mechanism for underlying data.

Conclusion

There never has been and probably never will be enough resources to properly curate all research data to the highest standards; even if such resources existed, it would not be worth doing because much research data are not (and never will be) reused for various reasons or can be readily reproduced. Unfortunately, the overlapping communities of scholars, data curators, and data archivists are only at the beginning of a real understanding of which data are—and which are not—likely to be reused to advance scholarship. As a consequence, these communities are still learning how resources for data preservation and curation should be allocated. Further, there is a tension in that data curation and preservation compete for resources (including finite researcher attention, as well as more fungible financial resources) with the pursuit of new scholarship. Funding agencies are reluctant to invest in data curation and preservation—particularly as an ongoing investment, as opposed to a line item in a research grant—because this investment shifts funds away from the underwriting of new research, which is their primary mission. It is interesting to think about how the organization of various research communities and the funding structures for those communities will shape this challenge. For example, as part of the National Institutes of Health, constituent agencies such as the National Library of Medicine are given specific missions (and budgets) to manage the scholarly record; the NSF, in contrast, has no parallel organization, and each discipline supported by NSF struggles with these issues in a much more ad hoc fashion. Other complications include long time horizons and the limits to the transfer of data across centuries—realistically, reuse over such time horizons may be quite difficult. As noted above, some of the major challenges include the following:

- **The relative costs of preservation versus curation:** To a great extent, curation is a speculative investment: You invest to make data usable across time but only get a return if those data are actually used. Moreover, with the exception of large-scale community-oriented projects (sky surveys, sequencing, etc.), we often do not have a good sense of what data will be used and how often. As a result, there's great interest in trying to reduce the costs of curation to a minimum, at least for large classes of data. Approaches such as standardizing or templating classes of observational or experimental data captures will help a great deal. Preservation, by contrast, is largely a pay-as-you-go model.

- **The value of metadata:** In cases where the data are derived from observations, reliable information (metadata) about the instruments used, their calibration, and their accuracy will be needed. In a big data context, this is extremely important. Where instruments generate vast amounts of data over long periods of time, automatically capturing such context and attaching it to the data facilitates curation. For very large initiatives, this is now often part of the explicit planning for the design of instruments; there is, however, a sizeable “middle range” of commercial instrumentation used in research that does not seem to do this particularly well, representing a promising target for future work. In cases where the data are derived from computations, reliable information about the computations (including the applications program, software libraries, and underlying operating system, ideally with the ability to rerun the computation) is essential. Historically this has been an overwhelming challenge. Recent progress in virtualization is making this increasingly tractable, however, though many challenges persist (including how to integrate proprietary software and the use of network-based services as part of applications), making it difficult to define the “boundaries” of a computation that is being captured and documented.

Being able to understand a data set means understanding something of the conceptual scientific context under which it was developed. Major scientific paradigm shifts will certainly create problems here, something that has not yet been well studied.

Author

Clifford A. Lynch
Executive Director
Coalition for Networked Information (CNI)

Acknowledgements

In addition, the following ECAR Campus Cyberinfrastructure (ECAR-CCI) members contributed to this paper:

Guy T. Almes (Co-Chair)
Director, Academy for Advanced Telecommunications and Learning Technologies
Texas A&M University

Celeste Anderson
Director, External Networking Group
University of Southern California

Curtis W. Hillegas (Co-Chair)
Associate CIO, Research Computing
Princeton University

Timothy Lance
President
NYSERNet

Karen A. Wetzel (Staff Liaison)
Program Manager
EDUCAUSE

Ralph J. Zottola
CTO, Strategy, Research and Communications
University of Massachusetts Central Office

For more information about this paper and about ECAR working groups in general, contact Karen A. Wetzel, EDUCAUSE Program Manager, at ECARWG@educause.edu.

Citation for This Work

Lynch, Clifford A. *Big Data in the Campus Landscape: Curation*. ECAR working group paper. Louisville, CO: ECAR, November 20, 2015. Available from http://www.educause.edu/ecar.

Notes

1. Another useful definition of data curation comes from the Council on Library and Information Resources, which points to a definition from the University of Illinois Graduate School of Library and Information Science and references Sayeed Choudhury's "stack model" for data management employed by the Johns Hopkins University and the model's components—storage, archiving, preservation, and curation.

2. For more about the data life cycle, see Michael Fary and Kim Owen, *Developing an Institutional Research Data Management Plan Service*, ECAR working group paper, January 8, 2013.

3. See, for example, the recent study Kevin B. Read, Jerry R. Sheehan, Michael F. Huerta, Lou S. Knecht, James G. Mork, Betsy L. Humphreys, and the NIH Big Data Annotator Group, "Sizing the Problem of Improving Discovery and Access to NIH-Funded Data: A Preliminary Study," *PLoS One*, July 24, 2015: DOI: 10.1371/journal.pone.0132735, and The Stewardship Gap Project.

4. For more information, see the NIH Data Sharing Policy and the NSF ENG Data Management Plan Requirements. See also Overview of OSTP Responses, which outlines the responses of federal agencies to the 2013 OSTP memo, as described in the blog Expanding Public Access to the Results of Federally Funded Research. Finally, the Association of Research Libraries maintains a list of current public access mandates by federal funders and provides some analysis of the various agency policies—see Access to Federally Funded Research.

5. For instance, see the Tebtunis Papyri Database, hosted by the Bancroft Library at the University of California, Berkeley, or the Duke Papyrus database.

6 See, for example, the Johns Hopkins University *Roman de la Rose* Digital Library.

7. A few case studies can be mentioned that describe what happens after the funding ends. After the funding for the Sloan Digital Sky Survey ran out, a few libraries, notably the library at Johns Hopkins University, took ongoing responsibility for preserving the data, working in conjunction with a consortium of astronomers. Another case study was the closing of the U.K. Arts and Humanities Data Service (AHDS), a central repository for data sets produced by work funded through the U.K. Arts and Humanities Funding Council. Here the responsibility for substantial amounts of data stored in the AHDS was taken on by the various universities that hosted the investigators who produced it. But it is also notable that a substantial number of biomedical databases have gone away in recent years. *Nucleic Acids Research* publishes an annual issue surveying developments in databases for molecular biology and has been documenting this die-off of databases. Finally, note that it's important to be specific about different situations: There is a big difference between the defunding of a curated collection of data integrating results from many researchers and/or many instruments and the case in which an *individual* data set in an institutional or disciplinary repository runs out of support funding in a cost-recovery regime.

8. Fary and Owen, *Developing an Institutional Research Data Management Plan Service*. See also Andrew Cox, "RDM: Who Does What?" *RDMInsight*, August 14, 2015.

9. This challenge is being focused on in the new ECAR working group on Data Protection: A Fundamental Shift (expected completion summer 2016).

10. Note that this is not a dichotomy. XML-based metadata can be both textual and also highly structured and machine processable.

11. These reuse, retention, and access constraints can be very complex, particularly when institutional review boards get involved. Consider cases that range from a data set that needs to be kept until all the people in it have been dead for 30 years (common for medical trials) to a data set that must be embargoed until everybody mentioned in it has been dead for 30 years (a common constraint on archival materials). If a data set involving human subjects needs to be repurposed, typically the original investigator and his IRB need to get involved in approving the repurpose and reuse; if the investigator has died, for example, it's easy to end up with what are effectively orphaned data sets with no clear succession of responsibility.

12. While the options and best practices here are not well understood, one can imagine a number of decisions that result in keeping data but putting it on life support in various ways, either by reducing the number of copies retained or postponing various types of format migration and dependency analysis, thereby trading lower costs of maintaining the data for what would probably be much higher costs if a researcher tried to reuse it in the future. Also note that appraisal decisions are not simply based on the intrinsic value of a given data set but rather occur in a competitive environment where limited resources need to be allocated for the ongoing curation (and preservation) of a vast array of data resources in various disciplines. The disciplinary, national, and international mechanisms for conducting these kinds of large-scale reappraisals simply do not exist today and are going to have to be developed in coming decades.

13. Even discovery by investigator name or funding source is very difficult, as one needs to deal with variations in names and ways of describing funding sources. Programs such as ORCID and FundRef are developing unambiguous identifiers that allow better discovery, but this means that these identifiers need to be associated with the data set as part of the initial curation process.

14. See, for example, the work of the NSF DataONE project at What is DataONE?