Academic Libraries as Data Quality Hubs[*]

Michael J. Giarlo
Penn State University
E-017 Paterno Library
University Park, PA 16802
michael@psu.edu

ABSTRACT

Academic libraries have a critical role to play as data quality hubs on campus, based on the need for increased data quality to sustain "e-science," and on academic libraries' record of providing curation and preservation services as part of their mission to provide enduring access to cultural heritage and to support scholarly communication. Scientific data is shown to be sufficiently at risk to demonstrate a clear niche for such services to be provided. Data quality measurements are defined, and digital curation processes are explained and mapped to these measurements in order to establish that academic libraries already have sufficient competencies "in-house" to provide data quality services. Opportunities for improvement and challenges are identified as areas that are fruitful for future research and exploration.

Categories and Subject Descriptors

E.0 [**Data**]: General; H.4 [**Information Systems Applications**]: Miscellaneous; H.3.7 [**Digital Libraries**]: General

Keywords

data quality, digital curation, digital preservation, academic libraries, stewardship, e-science, research data, trust

1. SCIENTIFIC DATA AT RISK

Data quality is a pressing, not to mention costly, issue in industry; a 2002 study [16] calculated that over $600 billion per year was spent on "data quality problems" [9]. At the same time, data quality issues have become an area of growing attention within academia and academic libraries [11, 6, 14, 12], as scientific practices evolve to exploit robust campus cyberinfrastructure and as funding agencies, such as the National Science Foundation and the National Institutes of Health, increasingly require data management plans to protect and amplify the impact of their investments.

As computing costs have dwindled, computer processing speed, network throughput, and storage capacity have grown, resulting in an explosion of scientific data. Experiments, in some disciplines more than others, are producing more data than their principal investigators and research assistants can handle [4]. Due to the wealth of data that is being produced, scientific practice is changing; the gathering of data for one experiment may drive dozens or hundreds of other experiments around the world [12].

Data is more abundant than ever before, and no less important, and yet it is at risk [14, 11]. "The survival of this data is in question since the data are not housed in long-lived institutions such as libraries. This situation threatens the underlying principles of scientific replicability since in many cases data cannot readily be collected again" [11]. There are numerous examples in the literature of analog data enabling scientific inquiry decades and longer past the date it was gathered [1]; how do we as a society, and particularly we within academia, not only preserve this wealth of data for future science but ensure it is of high quality?

1.1 Curatorial Practice and Challenges

Cultural heritage organizations such as libraries and archives have been stewards of society's cultural and scientific assets for millennia, providing public access to high-quality collections, and they remain so in the Internet age. Though the activities involved are different for analog assets, "[s]tewardship of digital resources involves both preservation and curation. Preservation entails standards-based, active management practices that guide data throughout the research life cycle, as well as ensure the long-term usability of these digital re-

[*] Prepared for the NSF III #1247471 workshop, **Curating for Quality: Ensuring Data Quality to Enable New Science**, in Arlington, VA, USA

[1] Ogburn [14] cites Stephen Jay Gould's "The Mismeasure of Man" in which we learn that "analysis and critique of cranial measurements in the 1800s, twin studies in the 1950s, and the rise of IQ testing were possible because the data were still available for scrutiny and replication"

sources. Curation involves ways of organizing, displaying, and repurposing preserved data" [6].

Digital preservation and digital curation, though relatively new practices, are widely treated in the literature [12, 8, 10, 14, 11, 17, 6]. Digital curation aims to make selected data accessible, usable, and useful throughout its lifecycle. Digital curation subsumes digital preservation; without viable data, which digital preservation enables, there's nothing to be curated [2].

An oft-cited mantra on the practice of digital curation is that "curation begins before creation [of the data]" [15]. And yet, "[b]y the time knowledge in digital form makes its way to a safe and sustainable repository [such as those provided by academic libraries], it may be unreadable, corrupted, erased, or otherwise impossible to recover and use. Scientific data files may be especially endangered due to their sheer size, computational elements, reliance on and integration with software, associated visualizations, few or competing standards, distributed ownership, dispersed storage, inaccessibility, lack of documented provenance, complex and dynamic nature, and the concomitant need for a specialized knowledge base — and experience — to handle data. Data also may be endangered by the practices of scholars who regard their data as having little value beyond the confines of a small group, a specific project, or a specified period" [14].

1.1.1 *Post-Hoc Curation Considered...*

As digital curation is a new practice, and is generally centered within cultural heritage organizations (rather than within the research enterprise), *post-hoc* curation is an unfortunate fact of life; researchers lack the incentive, the resources, the time, or the expertise to curate their own data [3], and so its curation falls to other parties after the data has been created, and often after it has been "archived." For especially massive data sets, furthermore, it is difficult even to imagine, *e.g.*, a research institute or academic department having sufficient resources to curate their own data at scale.

The practice of *post-hoc* curation (vs. "sheer curation," or curation by researchers at the time of creation) is less than ideal for a number of reasons.

First, one of the goals of curation is to enable the usefulness of a digital resource over time, and one of the tactics applied is to provide sufficient context for a resource such that future users can understand what an object is, where it came from, why it is significant, and how to use it. Context is often provided via documentation, descriptive metadata, or both [6, 11, 8, 12]. The creator(s) of the data, **not** its *post-hoc* curators, are best equipped to provide this context; to get a sense of this distinction, consider the difference between the tasks of cataloging your own book collection and cataloging a complete stranger's book collection.

Second, building on the prior reason, is that *post-hoc* curation happens some time after the data have been created, possibly a long enough time to lose track of important information; capturing the context around a data set is best done while the data is still fresh in its creator's mind, *i.e.*, before or during its creation. Documentation or metadata that is created by a party other than the data's creator, especially when performed after the responsible parties have moved on to other challenges, will suffer from this lack of context.

"This [*post-hoc* curation] activity is to provide representational information and description. This is particularly problematic for academic libraries, since the data being generated at research and teaching institutions are incredibly varied. Many representational schemes for the data and metadata will be required. No one individual will have all of the required skills. Data curators will need to collaborate closely with the data providers to understand the data" [11]. Whether researchers will have sufficient time, resources, and inclination to collaborate with academic libraries on the work of curating research data at scale is yet to be seen.

Finally, possibly the most limiting reason: there is a misalignment between the scale of the need for on-campus data curation and the level of commitment by academic libraries to address this need (as measured by the amount of resources allocated to this need vs. other needs). Data curation efforts are often understaffed and underresourced, with many academic libraries devoting one full-time equivalent employee, if that, to this role, to say nothing of the level of administrative and staff support for this role.

Academic libraries, institutional will and administrative support notwithstanding, are nonetheless uniquely positioned to tackle the problem of data quality in e-science by virtue of their record of effective stewardship, their commitment to providing access to high-quality data over the long-term, and their expertise in digital preservation and digital curation practices, as "[digital] curation is a process that can ensure the quality of data and its fitness for use" [8]. It is worth examining this claim in the context of a framework for measuring data quality.

2. MEASURING DATA QUALITY

There are a number of theoretical frameworks quantifying data quality measures already established, and Knight's 2005 paper compares a selection of a dozen "widely accepted [information quality] Frameworks collated from the last decade of [information science] research" [5]. Common features are identified for data quality (or information quality), such as that it is a concept with multiple dimensions, wherein the overall quality is a function of successive indicators. Another common feature of data quality frameworks is the grouping of quality indicators into categories, classes, or levels corresponding to, *e.g.*, semiotic levels, layers of intrinsicity and extrinsicity, and the subjectivity / objectivity spectrum.

The following framework is distilled from Knight's comparison of quality frameworks, and constitutes "a series of quality dimensions which represent a set of desirable characteristics for an information resource" [8]. The framework is then applied to the domain of research data quality as viewed from my perspective, that of a digital preservation technologist

[2] This characterization of digital curation and digital preservation is a mere gloss; more may be found, for instance, on the Digital Curation Centre's website: `http://www.dcc.ac.uk/digital-curation`.

[3] Hereafter referred to as "sheer curation or curation at source" [8].

and practitioner of digital curation. It is not offered as a novel framework, nor a comprehensive one, but merely as a tool for understanding and evaluating the applicability of digital curation and preservation practices to the measure of data quality.

Trust
Evaluation of the extent to which data is trusted depends on a set of subjective factors, including whether the data is judged to be authentic, the uses to which the data is put, the subject discipline, the reputation of the party/ies responsible for the data, and the biases of the person who is evaluating the data [4].

Authenticity
Evaluation of the authenticity of data requires that data be understood. Authenticity in this context is a rough measure of the extent to which the data is judged to be "good science," answering questions pertaining to, *e.g.*, the reliability of the instruments used to gather the data; the soundness of underlying theoretical frameworks; the completeness, accuracy, and validity of the data; and ontological consistency within the data.

Understandability
Evaluation of the understandability of data requires that there be sufficient context (documentation, metadata, or provenance) describing the data, and that the data is usable.

Usability
Usability of data requires that data is discoverable and accessible; that data is in a usable file format; that the individual judging the data's quality has an appropriate tool to access the data; and that the data is of sufficient integrity to be rendered.

Integrity
Integrity of data assumes that the data can be proven to be identical, at the bit level, to some prior accepted or verified state. Data integrity may be required for usability, understandability, authenticity, trust, and thus overall quality, though this depends in part of the level of perturbation of integrity. Integrity changes will have varying effects depending on how significant the perturbation is, the file format, and where within the file the perturbation has occurred.

The relationship between the quality dimensions in this framework is analogous to that of the Semantic Web Layer Cake in that "each layer exploits and uses capabilities of the layers below" [1]. Viewed from the bottom up, this framework asserts that data integrity may be necessary but not sufficient for data quality; if the data lacks integrity, it may not be usable, and thus not understandable, authentic, or trustable — a very low measure of quality. On the other hand, unauthorized changes at the bit level may not effect the rendered data in any perceivable ways. Viewed from the top down, on the other hand, if an individual trusts a data set, she likely judges it to be of the highest quality even if it is not usable, understandable, or fixed in integrity.

3. APPLYING CURATION TO DATA QUALITY

Within the defined framework, how might the practice of curation help ensure data quality? Each of the indicators in this framework is evaluated within the context of the digital curation lifecycle [7].

3.1 Integrity

The curation lifecycle contains actions geared towards preservation of the digital asset, which includes bit-preservation via a number of possible tactics such as regular digital signature (or checksum) verification, replication, media refreshing, version management, and file-level backups. These tactics taken together should be sufficient to ensure that the data remains in the same state as originally processed. Assuming that the data was authentic to begin with [5], the effective practice of curation should provide data integrity.

3.2 Usability

Three of the seven sequential actions defined in the lifecycle model have a direct impact on the usability of data. First, the Create or Receive action [6] should include determination of an appropriate file format for the data, choosing a format that is judged to be widely accessible and preservable. The Access, Use, & Reuse action "[e]nsure[s] that data is accessible to both designated users and reusers, on a day-to-day basis", thus ensuring that the data is discoverable and made available to potential users of data. The Transform action, lastly, includes periodic evaluation of file formats and migration to new formats so data remain usable well after the original formats have been rendered obsolete.

3.3 Understandability

Context is provided for data, in order that users may understand the data, both in sequential actions within the curation lifecycle — those being Create or Receive and Preservation Action — and also within the full lifecycle action of Description and Representation Information. The generation, extraction, and application of metadata by machine agents and humans is thus a key part of the curation lifecycle, providing periodic management and addition of context to data. These actions make sure the data's purpose, impact, and provenance are established over the course of its lifecycle so that current and future users can make sense of data that they have discovered.

3.4 Authenticity and Trust

Authenticity and trust as dimensions of data quality are highly subjective. The curation process can document what instruments are used to generate data, but not how reliable

[4] Trust is a complex issue that though relevant is too far-reaching to be within the the scope of this position paper. It is nonetheless listed in the framework at the very top to establish that lower layers may be entirely discounted by an individual judging data quality if there are overriding trust issues. This topic is fertile for subsequent research

[5] Authenticity is evaluated higher up the stack.

[6] Again underscoring the mantra that "curation begins before creation"

a user judges those instruments to be; it can include metadata about the theoretical frameworks underlying the data, but not whether the frameworks are theoretically sound; it can clearly establish the parameters of the data, but it is up to the user to judge whether those are a complete or incomplete set of parameters. The context, provenance, and documentation provided by curation are thus critically important in arming users of data with the information they need to make quality judgments but are **not** capable of independently ensuring data authenticity or trust in data; that is entirely for the individual user to judge.

4. AREAS OF OPPORTUNITY

4.1 Curation Models

Given the issues with the practice of *post-hoc* curation raised above, it is worth examining alternative curation models. This is not to suggest that one model of curation is to be selected exclusively; a mix of *post-hoc* curation and curation-at-source models will likely be in place at most institutions.

The work required for doing curation at the source needs to be incentivized and integrated into the researcher's extant workflows. Unless there are clear and valuable incentives for researchers to spend time and thought on curatorial work, and unless curation can be made to fit into the way researchers currently work, curation will be an after-thought, and thus so will data quality.

These different curatorial models are not mutually exclusive and in fact it may be ideal to combine them, leveraging both the researcher's deep domain knowledge and the professional curator's commitment, expertise, and tools to preserve data quality over time.

4.1.1 Scaling Post-Hoc Curation

Curry has examined a number of successful community-based curation models, which may offer academic libraries a way to scale *post-hoc* curation and deal with the aforementioned deficiencies of this approach: "[d]ata curation teams have found it difficult to scale the traditional [*post-hoc* curation] approach and have tapped into community crowd-sourcing and automated and semi-automated curation algorithms" [8].

The rise of the "citizen science" paradigm, such as demonstrated in the Galaxy Zoo and Zooniverse projects [2, 4], suggests community crowd-sourcing as a tactic that may be used to complement an institution's curation model. These initiatives leverage the "wisdom of the crowd" in curating [7] massive data sets such as the astronomical image data in the original Galaxy Zoo project. Galaxy Zoo in particular has been wildly successful, attracting a user base numbering into the hundreds of thousands, who have worked together to classify hundreds of millions of records [4].

There are numerous incentives at play in crowdsourcing, such as access to broadly interesting and compellingly visualized data; competition; and a desire for the layperson to be involved with *bona fide* research with opportunities to make novel scientific discoveries despite limited domain expertise. Consider "Hanny's *Voorwerp* [3]," an astronomical body discovered in Galaxy Zoo's data set by an amateur astronomer. The *Voorwerp* is now being studied by more than one professional astronomer, studies that may never have happened if not for the serendipitous discovery of an untrained curator. There are numerous other collaborative or crowd-sourced curation efforts highlighted in Curry's chapter on community data curation [8].

Galaxy Zoo and other Zooniverse projects demonstrate aspects of a model that could be repurposed in academic libraries as libraries seek alternative models for research data curation that scale out.

As mentioned earlier, some combination of *post-hoc* curation and curation-at-source seems effective. The Galaxy Zoo project balances crowd-sourced curation with verification by trained astronomers [4], who verify samples of curatorial work over time, thus enabling network effects to take place — this form of training or correction is not unlike the balance between human correction and machine learning algorithms, or, *e.g.*, the reCAPTCHA [8] service. This sort of delegation of quality to the community is not unlike a principle found in the open source software world, which is that the more eyes are on a codebase, the more likely it is that defects will be found and corrected.

The challenges that face academic libraries in leveraging crowd-sourcing as part of an institutional data curation strategy, each of which bears more in-depth consideration or research, are finding or allocating sufficient resources to build tools; finding effective incentives to curate research data; building a community around the data that is large enough to realize the benefits of network effects; and coming up with a model that puts the "trust but verify" strategy, whereby a sampling of crowd-curated records is checked for quality (and corrected if need be), into effect at scale.

Curry [8] has identified a number of social and technical best practices around community curation, which may be useful in addressing these challenges: early and sustained stakeholder involvement; outreach beyond the existing community via multiple channels including both emerging social media and more traditional channels such as newsletters and mass email; connection of curation activities to tangible payoffs; an appropriate and clear governance model; community-standard data representations; balance between automated and human curation with the latter always overriding the former; and recording and displaying provenance events to provide additional context to crowd curators and users.

In addition to human curation, whether via trained curators or citizen curators in "the crowd," there is a growing number of increasingly sophisticated tools for automated curation which could be used as a less costly and more timely tier of curation (until such time as a human curator has time to curate a data set). Tools for automated curation such as for subject classification, part-of-speech tagging, semantic entity extraction, and characterization can provide

[7] Or, at least, classifying, cataloging, and otherwise annotating these data sets, even if it not inclusive of all activities within the curation lifecycle.

[8] `http://www.google.com/recaptcha`

much-needed context to enable some level of understandability, usability, authenticity, and trust. Automated curation can thus help with data quality in a way that scales in a less resource-constrained way than requiring intensive human curation of every data set.

4.2 Academic Libraries as Data Quality Hubs

Academic libraries have an opportunity to serve as data quality hubs on campus, extending their established digital curation and preservation services to the research enterprise, doing for e-science what libraries have a wealth of experience doing for other areas of scholarly communication. With the scramble to establish data management support services in the wake of the NSF's data management plan requirement, the timing is opportune to take advantage of the new and reinforced connections between libraries and researchers by offering new services around data quality.

Libraries that lack the resources to sustain a new university service around data quality, or libraries on campuses where other organizations (such as central IT) might be better resourced or positioned to provide such services, may play a less active but equally vital role. Libraries are in large part the centers of campus, where so much of the institution's research, publishing, and instruction come together. Librarians that serve as liaisons to academic departments and research institutes provide a crucial connection that libraries could use for outreach and marketing in the area of data quality services; though the libraries may not provide data quality services themselves, they may serve a consultative role, pointing at relevant services on campus and abroad, helping to "knit" them together for the research enterprise.

Libraries can also offer assistance in the form of instruction, not radically different from existing information literacy programs, particularly around practical tools and processes pertaining to personal digital curation [17]. Such instruction could be especially helpful at institutions where the culture is that of extreme decentralization or sparse collaboration.

There is a tremendous opportunity as well to offer workshops and otherwise emphasize the value of curation in providing data quality for e-science, and also to publicize the "curation begins before creation" mantra. The sooner libraries can insert themselves into the research process, the better the data quality situation will be on campus. Libraries need to figure out how to "hack" academic culture and scientific practice in such a way that curatorial skills are considered required within the new scientific process.

4.2.1 Helping Others to Help Us Help Others

New "data science" programs such as the certificate program at the University of Washington [13] give the author hope that there is some movement in this area. The focus on data gathering, analysis, and visualization is an important start; quality and curation, however, are noticeably absent. A more complete degree program in data science would effectively combine these topics with those within data curation and retention, pulling together domain-specific knowledge, scientific methodology, computer science techniques, and best practices from the information science, information technology, and cultural heritage realms to ensure effective management of data quality over time. The onus is on cultural heritage institutions such as academic libraries to make this happen, a daunting and enormous challenge to be realistic. It falls to us to make a convincing value-added argument regarding curation and preservation of data to researchers. Funding agencies like the NSF and NIH can help with this by continuing to require substantial data management plans, as can academic research offices and subject disciplines and institutes; forging or strengthening partnerships with these departments would be strategic for libraries on campus. This recommendation echoes one of the findings of the 2006 Association of Research Libraries report on data stewardship, namely that "[a] change in both the culture of federal funding agencies and of the research enterprise regarding digital data stewardship is necessary if the programs and initiatives that support the long-term preservation, curation, and stewardship of digital data are to be successful" [6].

4.2.2 Our Challenge

Are academic libraries adequately prepared for this role? A new suite of data quality services on campus may require not insignificant re-skilling and re-education of the workforce, and may also require some reorganization and redefinition of positions [12].

I agree strongly with Ogburn, who argues that "funding and planning for the care and retention of data must be built into the front end, not the back end, of the research process. Data files must be attended to while they are compiled and analyzed in order to keep them available for a reasonable life span. This will require librarians to be conversant with the language and methods of science, at the table for campus cyberinfrastructure planning, and working with researchers at the beginning stages of grant planning" [14].

Academic libraries **need** to be conversant with the language and methods of science and to be involved with advances in campus cyberinfrastructure. We have the expertise and the challenge of data quality is well within the traditional mission of libraries. The time has come for academic libraries to serve as data quality hubs on campus to enable a new generation of scientific discovery and inquiry for the good of our society.

5. REFERENCES

[1] http://en.wikipedia.org/wiki/Semantic_Web_Stack.

[2] http://en.wikipedia.org/wiki/Galaxy_Zoo.

[3] http://en.wikipedia.org/wiki/Hanny{'}s_Voorwerp.

[4] T. Adams. Galaxy zoo and the new dawn of citizen science. *The Guardian*, March 2012. http://www.guardian.co.uk/science/2012/mar/18/galaxy-zoo-crowdsourcing-citizen-scientists.

[5] S. Knight and J. Burn. Developing a framework for assessing information quality on the world wide web. *Informing Science*, 8:159–172, 2005. http://inform.nu/Articles/Vol8/v8p159-172Knig.pdf.

[6] Association of Research Libraries. *To Stand the Test of Time: Long-Term Stewardship of Digital Data Sets in Science and Engineering.* Association of Research Libraries, 2006.

http://www.arl.org/bm~doc/digdatarpt.pdf.

[7] Digital Curation Centre. DCC curation lifecycle model. http://www.dcc.ac.uk/resources/curation-lifecycle-model.

[8] E. Curry, A. Freitas, and S. O'Riain. *The Role of Community-Driven Data Curation for Enterprises*, pages 25–47. Springer, 2010. http://3roundstones.com/led_book/led-curry-et-al.html.

[9] W. W. Eckerson. Data warehousing special report: Data quality and the bottom line. *Application Development Trends*, May 2002. http://adtmag.com/articles/2002/05/01/data-warehousing-special-report-data-quality-and-the-bottom-line_633729392210484545.aspx.

[10] C. Goble, R. Stevens, D. Hull, K. Wolstencroft, and R. Lopez. Data curation + process curation=data integration + science. *Briefings in Bioinformatics*, 9:506–517, July 2008. http://bib.oxfordjournals.org/content/9/6/506.full.

[11] P. B. Heidorn. The emerging role of libraries in data curation and e-science. *Journal of Library Administration*, 51:662–672, October 2011. http://dx.doi.org/10.1080/01930826.2011.601269.

[12] JISC. The data deluge. http://www.jisc.ac.uk/publications/briefingpapers/2004/pub_datadeluge.aspx, November 2004.

[13] U. of Washington Professional and C. Education. Winter 2013 | data science certificate. http://www.pce.uw.edu/certificates/data-science/web-winter-2013/, 2012.

[14] J. L. Ogburn. The imperative for data curation. *portal: Libraries and the Academy*, 10:241–246, 2010. http://muse.jhu.edu/journals/pla/summary/v010/10.2.ogburn.html.

[15] C. Rusbridge. Project data life course. http://digitalcuration.blogspot.com/2008/11/project-data-life-course.html, 2008.

[16] P. Russom. Liability and leverage - a case for data quality. *Information Management*, August 2006. http://www.information-management.com/issues/20060801/1060128-1.html.

[17] P. Williams, J. L. John, and I. Rowland. The personal curation of digital objects: A lifecycle approach. *Aslib Proceedings*, 61:340–363, 2009. http://dx.doi.org/10.1108/00012530910973767.