Journal of Librarianship and Scholarly Communication

Volume 1 | Issue 3 eP1059

Academic Libraries as Data Quality Hubs

Michael J. Giarlo

Giarlo, MJ. (2013). Academic Libraries as Data Quality Hubs. *Journal of Librarianship and Scholarly Communication* 1(3):eP1059.
http://dx.doi.org/10.7710/2162-3309.1059

JLSC is a quarterly journal sponsored and published by Pacific University Library | ISSN 2162-3309 | http://jlsc-pub.org

Academic Libraries as Data Quality Hubs

Michael J. Giarlo *Digital Library Architect, Pennsylvania State University*

Abstract

Academic libraries have a critical role to play as data quality hubs on campus. There is an increased need to ensure data quality within 'e-science'. Given academic libraries' curation and preservation expertise, libraries are well suited to support the data quality process. Data quality measurements are discussed, including the fundamental elements of trust, authenticity, understandability, usability and integrity, and are applied to the Digital Curation Lifecycle model to demonstrate how these measures can be used to understand and evaluate data quality within the curatorial process. Opportunities for improvement and challenges are identified as areas that are fruitful for future research and exploration.

Implications for Practice:

- Managers and leaders within academic libraries with established data curation or digital repository services will learn how existing infrastructure could be extended to campus research data in order to make the problem of data quality more tractable.
- Library staff and faculty at libraries which lack the resources to establish data curation services will learn how they can get involved with data quality advancements on campus.
- Library staff and faculty will learn more about the problem of data quality, about indicators of data quality, about application of existing library proficiencies to data quality issues, and thus about marketing data quality services on campus.
- Digital curation practitioners and program managers will learn about exemplars and techniques of combining curation methodologies, such as incentives for post-hoc (e.g. crowdsourced) curation.
- Library administrators will be challenged to align research data curation rhetoric with resources.

Received: 10/5/2012 Accepted: 12/13/2012

INTRODUCTION

Academic libraries have a critical role to play as data quality hubs on campus, by providing data quality auditing and verification services for the research enterprise. In order to sustain 'e-science' or 'e-research', an emerging paradigm for scientific practice that relies upon data reuse, researchers need reassurance they are accessing high-quality data. But this very data is at risk for numerous reasons due to its size, rate of growth, heterogeneity, and lack of archival storage. Many academic libraries offer curation and preservation services as part of their mission to provide enduring access to cultural heritage and to support scholarly communication. Academic libraries have both the wherewithal and the mission to intervene in the critical area of data quality. Digital preservation and curation, core competencies of academic libraries, can be applied to support data quality services, by mapping curatorial practices to established data quality measurements. There are opportunities, both practical and aspirational, as well as challenges to data quality services. Academic libraries are ready for the challenge.

RESEARCH DATA AT RISK

Data quality is a pressing and costly concern. A 2002 study (Russom, 2006) calculated that over $600 billion per year was spent on "data quality problems" (Eckerson, 2002) in industry. Similarly in academia, data quality issues have received growing attention by academic libraries (ARL, 2006; Heidorn, 2011; Joint Information Systems Committee [JISC], 2004; Ogburn, 2010), 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.

The increased affordability of technology coupled with the concomitant increase in computer processing speed, network throughput, and storage capacity, has resulted in an explosion of research data. In fact, in some disciplines more data is produced than can be handled by principal investigators and research assistants (Adams, 2012). 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 (JISC, 2004).

Research data is more abundant, and more important, than ever before. Much of this data is deposited into disciplinary repositories and data banks that are funded by the U.S. government (Baker, 2012; Merali & Giles, 2005). Unfortunately, the government has prioritized funding of new services over maintenance of existing services, which jeopardizes the future of disciplinary data repositories (Merali & Giles, 2005).

The 2008-2012 global recession has exacerbated the funding crisis for such repositories (Baker, 2012) and has put the long-term stewardship of research data at risk (Ogburn, 2010). Cutting funding to government-supported disciplinary data repositories threatens the availability of research data, but creates an opportunity for organizations such as academic libraries to serve as stewards of this research data. "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" (Heidorn, 2011, p. 662).

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?

CURATORIAL PRACTICE AND CHALLENGES

For millennia, libraries and archives stewarded society's cultural and scientific assets, providing public access to high-quality collections, and they continue to do so in the Internet age (ARL, 2006):

> Stewardship 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 resources. Curation involves ways of organizing, displaying, and repurposing preserved data. (p. 12)

[1] Ogburn (2010) cites Stephen Jay Gould's *The Mismeasure of Man*: "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" (p. 243).

Digital preservation and digital curation practices are addressed widely in the literature (ARL, 2006; Curry, Freitas, & O'Riain, 2010; Goble, Stevens, Hull, Wolstencroft, & Lopez, 2008; Heidorn, 2011; JISC, 2004; Ogburn, 2010; Williams, John, & Rowland, 2009). 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]" (Rusbridge, 2008). 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. Research 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. (Ogburn, 2010, p. 242)

Post-Hoc Curation Considered

As digital curation becomes a more common responsibility of cultural heritage organizations[3], *post-hoc* curation will become an unfortunate fact of life. Researchers lack the incentive, the resources, the time, or the expertise to curate their own data[4], and so its curation falls to other parties after the data has been created or 'archived'. For especially massive data sets, it is difficult to imagine a research institute or academic department having sufficient resources to curate their own data at scale. The practice of *post-hoc* curation, as opposed to "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, 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 (ARL, 2006; Curry et al., 2010; Heidorn, 2011; JISC, 2004). 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, *post-hoc* curation happens some time after the data have been created, possibly a long enough time to lose track of important information. But, capturing the context around a data set is best done while the data is still fresh in its creator's mind, ideally before or during its creation. Metadata and related documentation created by a third party will lack this valuable context, especially when performed after the creators have moved on to other challenges.

> 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. (Heidorn, 2011, p. 667)

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.

Possibly the most compelling reason against *post-hoc* curation is that while academic libraries have the expertise

[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] The work of discipline-specific repositories such as e.g., the Protein Data Bank, GenBank, the Biomedical Magnetic Resonance Data Bank, Dryad, and others are notable exceptions.

[4] Hereafter referred to as "sheer curation or curation at source" (Curry et al., 2010, p. 31).

to mitigate the risks of *post-hoc* curation, currently the resourcing levels within academic libraries are not sufficient nor proportional to the campus data curation needs. Data curation efforts are often understaffed and underresourced, given the relative newness as a library service offering, with many academic libraries devoting one full-time equivalent employee, if that.

Academic libraries, nonetheless, are 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" (Curry et al., 2010, p. 46). It is worth examining this claim in the context of a framework for measuring data quality.

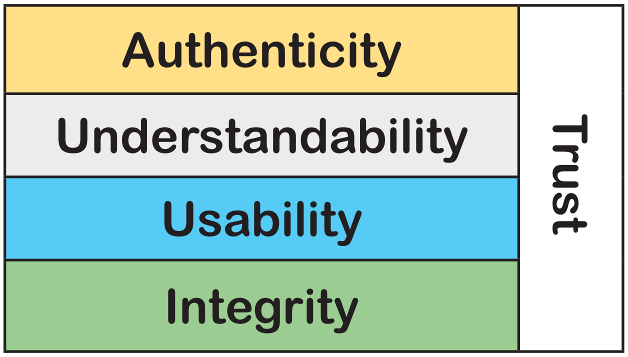
MEASURING DATA QUALITY

There are a number of "widely accepted [information quality] frameworks collated from the last decade of [information science] research" which can be used to assess data quality (Knight & Burn, 2005, p. 160). Data quality is a concept with multiple dimensions, wherein the overall quality is a function of successive indicators. These frameworks often group quality indicators into categories, classes, or levels corresponding to semiotic levels, layers of intrinsicality and extrinsicality, and the subjectivity / objectivity spectrum.

The following data quality framework (Figure 1) 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" (Curry et al., 2010, p.26). 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 acceptable use or application of the data, the subject discipline, the reputation of those responsible for the creation of the data, and the biases of the person who is evaluating the data[5].

Figure 1. Data Quality Framework



Authenticity: 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 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. In order to evaluate authenticity, the data must be understandable.

Understandability: Evaluation of the understandability of data requires that there be sufficient context, such as 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 is required for usability, understandability, authenticity, and thus overall quality. Integrity is subject to variation, or perturbation, and may have significant impact on other quality factors, depending on the extent of this perturbation. This perturbation can manifest in the file format or a location within the file.

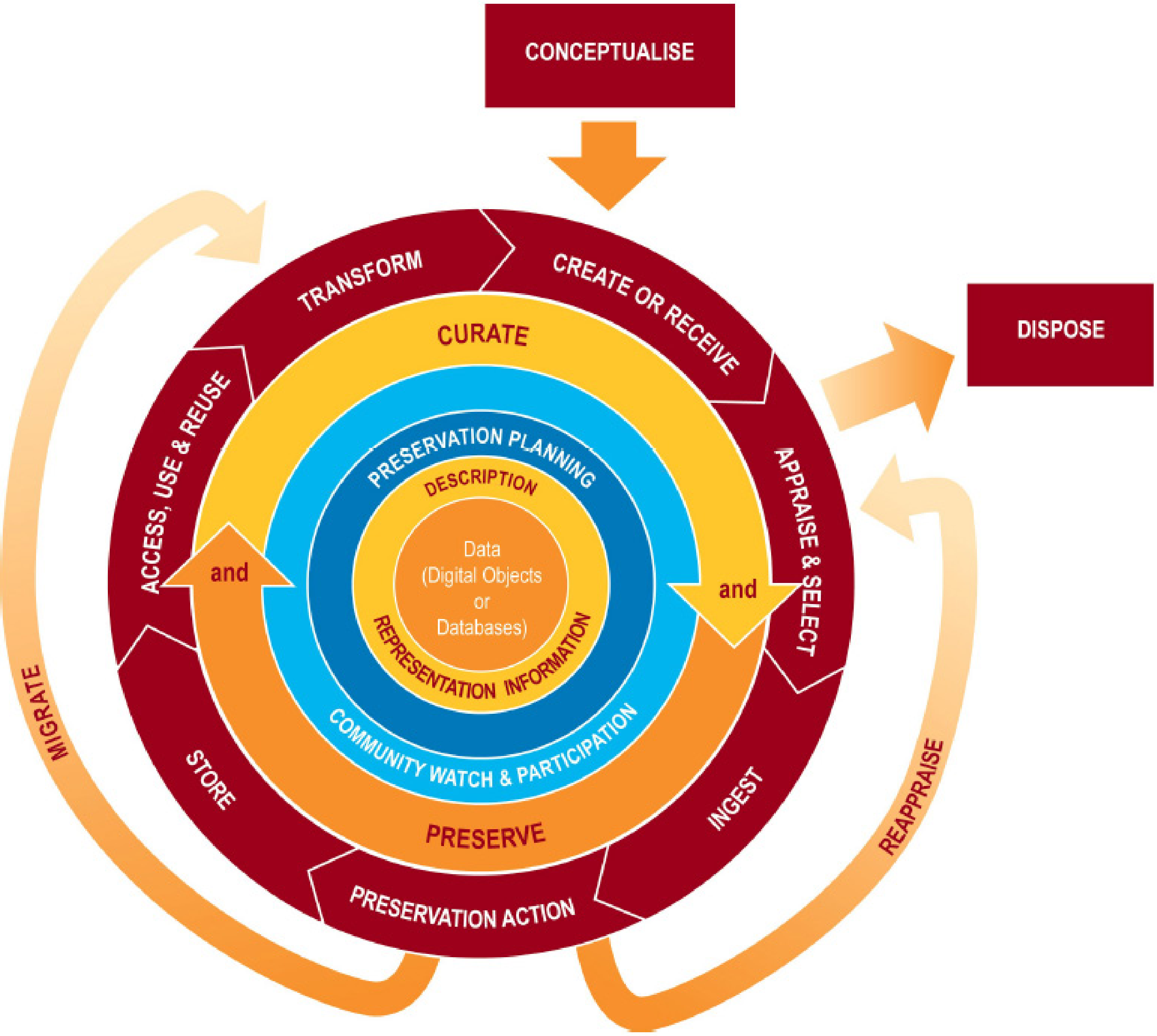
[5] Trust is a complex issue that though relevant is too far-reaching to be within the scope of this paper. It is nonetheless listed in the framework at the very top to establish that lower layers may be entirely discounted by an individual judging data quality if there are overriding trust issues. This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

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" (Wikipedia, 2012c). 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.

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 (Figure 2), from the author's perspective as a digital preservation technologist and practitioner of digital curation. The discussion will begin with the foundational element, integrity.

Figure 2. The DCC Curation Lifecycle Model (Higgins, 2008)



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[6], the effective practice of curation should provide data integrity.

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[7] 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" (Higgins, 2008, p. 138), 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.

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 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.

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 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 resulting from curation are thus critically important for users to make quality judgments. Data creators are *not* capable of independently ensuring data authenticity or trust in data; instead, it is the end user that will make that judgement.

Academic library services developed to support the data curation model could justifiably be marketed as data quality services on campus. The term "data quality hubs" is not substantially different from "data curation hubs" but rather frames the competencies of academic libraries in a way that applies to the emerging and yet critical area of e-science.

AREAS OF OPPORTUNITY

Curation Models

Given the issues with the practice of *post-hoc* curation highlighted earlier, it is worth examining alternative curation models. This is not to suggest that one model of curation is to be applied exclusively; a mix of *post-hoc* curation and curation-at-source models will likely be in place at most institutions. These curatorial models are not mutually exclusive and in fact it may be ideal to combine them, leveraging the researcher's deep domain knowledge and the professional curator's commitment, expertise, and tools to preserve data quality over time.

In order to fully adopt curation practices, researchers must be incentivized to integrate these practices into extant workflows. Curation, and thus data quality, will become an after-thought for researchers, unless the benefits of data curation are well-articulated, meaningful to the researcher, and the curatorial practices are assimilated into the researcher's workflows.

[6] Authenticity is evaluated higher up the stack.

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

Scaling Post-Hoc Curation

There are a number of successful community-based curation models, which may offer academic libraries a way to scale *post-hoc* curation while addressing deficiencies in the model. To wit, "[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" (Curry et al., 2010, p. 46).

The rise of the "citizen science" paradigm, demonstrated in the Galaxy Zoo and Zooniverse projects (Adams, 2012; Wikipedia, 2012a), suggests community crowdsourcing as a tactic that may be used to complement an institution's curation model. These initiatives leverage the 'wisdom of the crowd' in curating[8] massive data sets. 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 astronomical images (Adams, 2012).

There are numerous incentives in crowdsourcing this activity, such as access to broadly interesting and compellingly visualized data, competition, and the opportunity for a layperson with limited domain expertise to be involved in *bona fide* research and scientific discoveries. Consider "Hanny's *Voorwerp*" (Wikipedia, 2012b), an astronomical body discovered in Galaxy Zoo's data by an amateur astronomer. Because of the serendipitous discovery by an untrained curator, the Voorwerp is now being studied by professional astronomers. This is not a limited example, but one of many other collaborative or crowdsourced curation efforts highlighted in Curry's chapter on community data curation (Curry et al., 2010).

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 crowdsourced curation with verification by trained astronomers (Adams, 2012), who verify samples of curatorial work over time 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[9] 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 crowdsourcing as part of an institutional data curation strategy, each of which bears deeper 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 leverages the 'trust but verify' strategy, whereby a sampling of crowd-curated records is checked for quality (and corrected if need be), at scale.

Curry et al. (2010) 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 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 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 much-needed context to enable some level of understandability, usability, authenticity, and trust. Automated curation can thus help with data quality in

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

[9] http://www.google.com/recaptcha

a way that scales better than requiring intensive human curation of every data set.

CONCLUSION

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, or reframing existing curation and preservation services, around data quality.

Libraries that lack the resources to sustain a 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 (Williams et al., 2009). Such instruction could be especially helpful at institutions where the culture is that of extreme decentralization or sparse collaboration.

There is a tremendous teaching and outreach opportunity to further emphasize the value of curating for data quality for e-science, as "curation begins before creation." 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.

It Takes a Village

New "data science" programs such as the certificate program at the University of Washington (University of Washington Professional and Continuing Education [PCE], 2012) 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" (ARL, 2006, p. 12).

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 (JISC, 2004).

The provision of data quality services and extension of traditional stewardship to the realm of research data may not be feasible given the economic environment and existing commitments of academic libraries. Data quality presents an opportunity to offer new rationales to University administration for additional funding. The long-term stewardship risks associated with government-supported disciplinary data repositories (Baker, 2012; Merali & Giles, 2005) may help make the case that centrally-funded data services protect the University's investments in new research, increasing return on investment by ensuring for its long-term stewardship. As research data is typically owned by the institution itself in the United States (Clinical Tools, Inc. 2006; Erickson & Muskavitch, 2009), and not individual researchers, it is in the best interest of the institution to take a proactive role in safeguarding the data.

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" (Ogburn, 2010, p. 244). 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.

AUTHOR'S NOTE

This paper was originally prepared for the NSF III #1247471 "Curating for Quality: Ensuring Data Quality to Enable New Science" workshop in Arlington, VA, USA. Workshop proceedings are available at http://datacuration.web.unc.edu.

REFERENCES

Adams, T. (2012, March). Galaxy Zoo and the new dawn of citizen science. The Guardian. Retrieved from http://www.guardian.co.uk/science/2012/mar/18/galaxy-zoo-crowdsourcing-citizen-scientists

Association of Research Libraries (Ed.). (2006). To stand the test of time: Long-term stewardship of digital data sets in science and engineering, Association of Research Libraries. Retrieved from http://www.arl.org/bm~doc/digdatarpt.pdf

Baker, M. (2012, September). Databases fight funding cuts. *Nature*, *489*, 19. http://dx.doi.org/10.1038/489019a

Clinical Tools, Inc. (2006). Guidelines for responsible data management in scientific research. Office of Research Integrity, U.S. Department of Health and Human Services. Retrieved from http://ori.hhs.gov/education/products/clinicaltools/data.pdf

Curry, E., Freitas, A., & O'Riain, S. (2010). The role of community-driven data curation for enterprises, In *Linking enterprise data* (pp. 25–47). Springer. Retrieved from http://3roundstones.com/led_book/led-curry-et-al.html

Eckerson, W. W. (2002, May). Data warehousing special report: Data quality and the bottom line. Application Development Trends. Retrieved from http://adtmag.com/articles/2002/05/01/data-warehousing-special-report-data-quality-and-the-bottom-line_633729392210484545.aspx

Erickson, S., & Muskavitch, K. M. (2009). Ownership of data — federal policies. Retrieved from http://ori.hhs.gov/education/products/rcradmin/topics/data/tutorial_2.shtml

Goble, C., Stevens, R., Hull, D., Wolstencroft, K., & Lopez, R. (2008, July). Data curation+ process curation=data integration + science. *Briefings in Bioinformatics*, *9*, 506–517. http://dx.doi.org/10.1093/bib/bbn034

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

Higgins, S. (2008). The DCC curation lifecycle model. International Journal of Digital Curation, 3, 134–140. http://dx.doi.org/10.2218/ijdc.v3i1.48

Joint Information Systems Committee. (2004, November). The data deluge. Retrieved from http://www.jisc.ac.uk/publications/briefingpapers/2004/pub_datadeluge.aspx

Knight, S., & Burn, J. (2005). Developing a framework for assessing information quality on the World Wide Web. *Informing Science*, *8*, 159–172. Retrieved from http://inform.nu/Articles/Vol8/v8p159-172Knig.pdf

Merali, Z., & Giles, J. (2005, June). Databases in peril. *Nature*, *435*, 1010–1011. http://dx.doi.org/10.1038/4351010a

Ogburn, J. L. (2010). The imperative for data curation. *portal: Libraries and the Academy*, *10*, 241–246. http://dx.doi.org/10.1353/pla.0.0100

Rusbridge, C. (2008). Project data life course. Retrieved from http://digitalcuration.blogspot.com/2008/11/project-data-life-course.html

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

University of Washington Professional and Continuing Education. (2012). Winter 2013 | Data science certificate. Retrieved from http://www.pce.uw.edu/certificates/data-science/web-winter-2013/

Wikipedia. (2012a). Galaxy Zoo. Retrieved from http://en.wikipedia.org/wiki/Galaxy_Zoo

Wikipedia. (2012b). Hanny's Voorwerp. Retrieved from http://en.wikipedia.org/wiki/Hanny's_Voorwerp

Wikipedia. (2012c). Semantic web stack. Retrieved from http://en.wikipedia.org/wiki/Semantic_Web_Stack

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

CORRESPONDING AUTHOR

Michael J. Giarlo
Digital Library Architect

Pennsylvania State University
E-003 Paterno Library
University Park, PA 16802-1807

leftwing@alumni.rutgers.edu