Against the Grain

Volume 27 | Issue 2 Article 32

2015

Don's Conference Notes: Data Infrastructure: The Importance of Qualityand Integrity — A CENDI/ NFAIS Workshop andCharleston Seminar — Being Earnest With Our Collections: Determining Key Challenges andBest Practices

Donald T. Hawkins
dthawkins@verizon.net

DOI: https://doi.org/10.7771/2380-176X.7056

Don's Conference Notes

by **Donald T. Hawkins** (Conference Blogger and Editor) <dthawkins@verizon.net>

Data Infrastructure: The Importance of Quality and Integrity — A CENDI/NFAIS Workshop

CENDI (the **Commerce, Energy, NASA, Defense Information Managers Group**, *http://www.cendi.gov*) and **NFAIS** (the **National Federation of Advanced Information Services**, *http://www.nfais.org*) held a joint workshop on data quality and integrity on November 20, 2014 at the beautiful headquarters of the **U.S. Patent and Trademark Office** in Alexandria, VA.

Keynote Address

The workshop was keynoted by **Marcia McNutt**, Editor-in-Chief of *Science Magazine*, who presented an excellent overview of the issues surrounding data quality and integrity and used three examples to demonstrate that public trust in science depends on integrity.

Marcia McNutt

1. The data must be trusted. For example, several studies of stress levels in animals gave differing results, so people said the data were unreliable. However, according to a recent article in *Nature,*[1] animals tend to suppress pain around men more than around women, but the investigators had not recorded whether the studies were done by men or women. There was nothing intrinsically wrong with the data, but the unknown variable gave the perception that the data were unreliable.

2. Experiments must be reproducible. In the "Miracle on the Hudson" plane crash in 2009, the pilot reported that he smelled "burning birds" as the engines shut down. This observation led to experiments in the U.S. and Japan on the limits of engines to tolerate bird strikes; however, the researchers in both countries could not reproduce each other's data. Further investigation on the methodology revealed that the U.S. investigators were using fresh test birds, and the Japanese were using frozen ones. Once both teams used the same type of birds, the data were reproducible.

3. Interpretations should be free from bias. Bias is one of the hardest things to avoid in data, and there are two types of bias: false positives and false negatives. How a question is phrased can introduce bias into the final results. And sometimes experiments cannot be repeated, such as those involving earthquakes, because the earth never repeats the same event in the same way.

Journals have an important role to play in promoting data quality, and they have an obligation to alert the scientific community when data is found to be not reproducible. Prestigious journals are setting standards for publication because scientists want to publish their results in them. In a joint editorial[2] in *Science* and *Nature* (only the third in history between these two journals), **McNutt** discussed actions that journals were taking to address reproducibility, the development of guidelines for publication of research, and requirements for authors to report their experimental parameters. Over 70 publishers have agreed to the guidelines.

Incentives for producing quality data and reproducible results are available to a number of organizations:

- Federal agencies have a responsibility to make reproducibility in research part of their funding guidelines and to instill a culture of scientific and data quality and integrity in their operations. The Department of the Interior is the first agency to issue a policy on data quality for its agencies.
- Universities train current and future researchers in the scientific method. One incentive for producing reproducible results is to reward those who do so.
- Similarly, scientific societies should consider honoring researchers who consistently produce reproducible results and adopt reproducibility guidelines for their publications.

It is clear that a team effort is needed in these incentives, but privacy issues may cause problems, especially in areas such as biomedicine where patient data is often used in studies. In such circumstances, the policy adopted by *Science* is a model. If an author cites privacy restrictions, the owners of the data must show that anyone who wishes to repeat the research can access the data under the same terms and restrictions as the original authors, thus eliminating any potential biases. And authors should be encouraged to deposit their data in a public repository with links to the data and any publications resulting from it.

U.S. Patent and Trademark Office, Alexandria, VA.

Federal Policy Implications of Data Quality

Kevin Kirby, Enterprise Data Architect at the **Environmental Protection Agency (EPA)**, listed three recent legislative actions relating to data quality:

- ***The Data Act***, signed into law on May 9, 2014, is the nation's first legislative mandate for data transparency. It requires open, standardized data in federally funded research and publication of that data online.
- ***The Open Data Policy***, established in response to an Executive Order issued May 9, 2013, establishes data as an information resource and sets open and machine-readable data as the default for government information. One result of this policy has been a resurgence of interest in *data.gov*, the "home of the U.S. government's open data," which currently contains links to over 132,000 data sets.
- ***The Information Quality Act of 2002*** required the issuance of guidelines to Federal agencies ensuring the quality, objectivity, utility, and integrity of the information they disseminate.

In response, the **EPA** issued its own information quality guidelines (IQGs) *http://www.epa.gov/quality/informationguidelines/*, and is producing metadata records that describe data sets and provide links to them. It has also developed standards, controlled vocabularies, registries, and repositories for data elements. **Kirby** said that references and thesauri are very important in improving searches for data, and a data categorization scheme is still needed.

Daniel Morgan, Chief Data Officer at the **Department of Transportation (DOT)**, wondered if we are managing all of our assets properly. He noted that it is frequently difficult to standardize on definitions, but it is necessary; for example, the definition of a bridge is important in the National Bridge Inventory *(http://nationalbridges.com/)*. Data can become well regarded and trusted by capturing good metadata. Sometimes it is necessary to instill a culture within an agency's research community and implement a data management plan (which **DOT** has not done yet.)

Morgan said that we must reward people for sharing their data. Basic researchers need to interact with applied researchers, and we must help them to build good metadata. He suggested that the library community is a good place to turn for assistance in these areas.

Perspectives of Data Initiators, Funders, and Managers

Laura Biven, Senior Science and Technology Advisor at the **Department of Energy's (DOE's)** Office of Science, said that the Office supports about 22,000 scientists, graduate students, undergraduates, and engineers at over 300 institutions. It provides 47% of Federal support of basic research in the physical sciences and is also responsible for supporting over 28,000 users per year at the world's largest collection of scientific user facilities such as those at more than 30 National Laboratories and major universities. As a result, incoming data rates into computing sites are skyrocketing, and there is now an increased value in collecting data because of new analytic tools.

The Office of Science recently published its data management plan *(http://science.energy.gov/funding-opportunities/digital-data-management/)*, including principles and requirements, and its requirements will be included in all future solicitations for research funding. Other **DOE** offices will follow suit by October 1, 2015; by that time, there will be a single **DOE**-wide policy on data. Journal articles and accepted manuscripts from projects supported by **DOE** funding are now available on the **DOE's** Public Access Gateway for Energy and Science (PAGES) system *(http://www.osti.gov/pages/)*.

Dr. Isaac Kohane, Co-Director, **Center for Biomedical Informatics** at **Harvard Medical School**, focused on electronic medical records and said that one of the major challenges to reproducibility is getting the data in and

continued on page 60

out of a repository reliably. Even a small healthcare center can accumulate a large amount of data in a short time. **Harvard's** Shared Health Research Information Network (SHRINE, *http://catalyst.harvard.edu/services/shrine/*) is a repository of aggregated data on patients that can be used in medical research studies. **Kohane** said that the quality of the data is critical in such studies, and it is important to make data available and discoverable so that it can be used.

Melissa Cragin from the **National Science Foundation (NSF)** gave an update on public access plans for data (**NSF** does not conduct research; it only funds it). As a result of the Open Data Policy established in 2013, **NSF** has expanded its long-standing data sharing policy and is now requiring a two-page data management plan (DMP) as a supplement to all funding proposals. Publication and data management costs must be included as a direct charge in proposal budgets. Information in a DMP may include types of data, standards, access and sharing policies, provisions for re-use, and archiving plans. In a survey of DMPs, considerable variation was found in structure and content; using DMPs to understand trends is therefore a non-trivial effort. **Cragin** suggested that these issues for data need to be considered:

- Intersection of data management, public access, and preservation,
- Moving to a culture of sharing,
- Increasing our understanding of variations in the role of science "drivers," and
- Knowledge of and sustaining of the stakeholders.

Collaborative work is increasing, resulting in very large and complex data sets being produced, which is causing an increase in the need for access to tools for data sharing and publication. The traditional role of the single investigator with a team of graduate students is changing; big data is radically affecting the "long tail" of science.

Principles that have guided **NSF's** funding activities have included recognition and support for peer review, collaboration among agencies, and encouraging support for existing archives. **NSF** is responding to current changes in scientific research and is developing a new plan for data management that follows these core principles:

- We will proceed incrementally.
- We will respect the diversity of sciences that NSF supports and the communities in which scientific research is conducted and scientists are educated.
- We will use automated techniques, when appropriate, to reduce investigator and administrative burdens while achieving accountability.

The plan is in the comment phase now; when approved, it will be posted on the **NSF** Website *(http://www.nsf.gov)* along with FAQs and guidance.

Disseminators and Service Providers

Jane Greenberg, Professor and Director of the Metadata Research Center *(http://cci.drexel.edu/mrc/)* at **Drexel University**, began by noting that data is only as good as its metadata. She is involved with **Drexel's** DRYAD project *(http://datadryad.org)*, a curated general-purpose repository that makes data discoverable, freely reusable, and citable. Researchers are repeatedly creating the same metadata by cutting and pasting it into templates; the motivation for DRYAD is to automate metadata generation and allow researchers to concentrate on the things that need human intervention. It is important to get scientists to think about owning their metadata.

When an article is accepted for publication, authors are asked to deposit it in the DRYAD repository and receive a Digital Object Identifier (DOI) for it. The email notifying the author of acceptance is parsed, and a form prepopulated with metadata is returned for completion. (Many researchers will not fill in blank forms because of the time involved, so prepopulating the form as much as possible increases the likelihood of a response.) The data set is stored in the system, and it may be published before the article, which many journal publishers do not approve of; nevertheless, about 50 of them have signed a Joint Data Archiving Policy (JDAP, *http://datadryad.org/pages/jdap*) and have become DRYAD partners. Once the data has been deposited, it and its metadata can be accessed and reused.

A recent report published by the Office of Economic Cooperation and Development (OECD)10 suggested these valuations of some common data elements:

- Market cap of Facebook per user: $40 - $300
- Revenues per record per user: $4-$7 per year for Facebook and Experian
- Market prices of personal data:
  $0.50 for street address; $2 for date of birth; $8 for Social Security number; $3 for driver's license number; and $35 for military record.[3]

Metadata is an asset and can be used, thus increasing the value of the initial investment in the data. Although it costs about $40 to produce a metadata record, many articles have a reuse rate of over 50%.

DRYAD is now receiving about 80 papers a week for deposit in its repository. It is based on **MIT's** DSpace technology *(http://www.dspace.org/)*; DOIs are generated by the DataCite system *(https://www.datacite.org/)*. DRYAD began with articles in evolutional biology and has now been extended to other subject areas. Some articles have been downloaded many times, which is one measure of DRYAD's success. DRYAD is governed by a 12-member board that sets policy and goals; a payment plan was launched in September 2014.

Bruce Wilson, Enterprise Architect at **Oak Ridge National Laboratory**, said that his job is to help scientists do their job. There are many reasons to enable access to federally funded research; **Wilson** asked how do we ensure data quality to facilitate this? We need to understand what is happening when researchers generate data and help them to automate the process (**Wilson** called this "data carpentry"). Because of today's tools, it is easy to generate huge quantities of data. We must focus on what users need; in common with several other speakers, **Wilson** said that "good enough" is not a bad policy. Here are his observations:

- Keep the end in focus: doing science.
- Make doing the right thing the easy thing.
  - Automation is often key.
  - Security and usability should not be mutually exclusive.
- Value standards, sustainability, and simplicity.
- Confidentiality is often over emphasized. Think integrity first, then long-term availability.
- Discovery is essential to availability. Metadata is hard and essential.

Science is a voyage of discovery, so we need to set objectives reasonably depending on how far ahead we can see. We must protect the confidentiality of some data, but how do we balance that with the need for access to public data? Many people are looking for data they can easily find with common search tools, but they miss the wealth that is available in areas that popular search engines cannot see. Many tools can expose data; the challenge is to deliver the information that scientists need to do their job at any time, anywhere, and on any device.

Megan Force, Digital Research Analyst for Physical Science, **Thomson Reuters**, described **Thomson's** Data Citation Index (DCI, *http://wokinfo.com/products_tools/multidisciplinary/dci/*), which is part of the Web of Science. The DCI provides citations to data sets and can merge them into the metadata for an article. A recent study found that many researchers are not receiving adequate credit for their digital scholarship, so they are reluctant to share it, and many data repositories do not have clear standards or mechanisms to promote sharing.

The DCI was developed in response to researchers' problems in finding and sharing data. So far, 220 repositories are indexed; at its launch in 2012 it contained over 4 million data records. The DCI is cross-disciplinary and searches can be conducted across disciplines. Criteria for including a repository in the DCI include:

- Editorial content that is desirable to the research community,
- Persistence and stability, with a steady flow of new information (or at least an assurance that someone is in charge of the repository),
- Thoroughness and detail of descriptive information, and
- Links from the data to the research literature.

Formal citations to data sets are often difficult to find because they are buried in the text of articles or are cited in bibliographies. Efforts are underway to capture these citations and add them to the DCI.

Following this session, attendees and speakers were asked for a wish list for data producers. The following were mentioned:

- Better attribution of authorship of data,
- Support for the data carpentry movement, which is a vehicle for culture change, and
- The ability to show evidence that current research is moving science forward.

continued on page 61

Don's Conference Notes
from page 60

The audience noted that many of the points discussed in this session are related to the incentive structure of science and how scientists get credit for their work. They must perceive benefits of making their data available. Persistent identifiers are essential for data because some data sets may be in more than one repository. For credit purposes, only one identifier is needed, but whenever the data is changed, a new identifier must be used for each version of the data set. If a data set has been created from several others, all of the contributing data sets be cited.

Endnotes

1. "Olfactory exposure to males, including men, causes stress and related analgesia in rodents," *Nature Methods* 11, 629–632 (2014).
2. "Journals Unite For Reproducibility," *Science*, 346 (6210): 679 (November 7, 2014).
3. OECD (2013), "Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value," *OECD Digital Economy Papers,* No. 220, OECD Publishing. *http://dx.doi.org/10.1787/5k486qtxldmq-en*

Charleston Seminar — Being Earnest With Our Collections: Determining Key Challenges and Best Practices

Charleston Conferences have traditionally concluded with a "rump session," where a few hardy attendees gathered for a free-flowing discussion on topics of interest. During the past few years, a desire for a more structured session grew, and it came to fruition this year in the form of a new **Charleston Seminar** entitled **"Being Earnest With Our Collections: Determining Key Challenges and Best Practices."** The seminar attracted significant interest and drew about 50 attendees, who listened to four presentations on topics of current interest.

eBooks: Key Challenges, Future Possibilities

Michael Levine-Clark, Associate Dean for Scholarly Communications and Collections Services, **University of Denver**, and **Rebecca Seger**, Director, Institutional Sales, **Oxford University Press USA**, began by identifying the following key challenges facing today's eBook market.

1) Developing sustainable, flexible, and predictable business models has become difficult because today's trends affect all the players. Budget crises occur regularly in libraries, causing publishers' revenues to become even more unpredictable than they have been in the past. In the academic library market, demands for short-term loans and multiple eBook access models (subscriptions, purchases, or demand-driven acquisitions (DDAs)) have arisen. All of these forces are challenging, resulting in little predictability and sustainability in the eBook market.

2) In order to preserve their content, eBook producers and aggregators must consider which hosting platform will provide them with sustainability and long-term access to their products. Leased eBooks and those available through DDA are subject to these concerns. Every book that a publisher produces is a market risk and has continuing fixed costs. (Thus, long-term scholarly publishing in some disciplines is changing; for example, **Wiley** has ceased publishing physics books.) Publishers and aggregators have become the "library shelves" for eBooks, and they are experiencing pressure to impose hosting fees for content that may or may not be purchased. Perhaps a dual hosting model would be viable, with aggregators providing access across a range of publishers and managing discovery, and publishers implementing post-purchase access. It is important to ensure that all published scholarly monographs are preserved in a trusted repository such as **Portico** *(http://www.portico.org/digital-preservation/)* or **LOCKSS** *(http://www.lockss.org/)*.

3) Resource sharing in the print world is commonly done via interlibrary loan (ILL) — a core value. Some librarians have suggested implementing ILL for eBooks as well, but does that make sense? When a print book is out on ILL, access to it at the owning library is unavailable; how can that be implemented for eBooks without causing confusion to users? **Levine-Clark** and **Seger** said that we should work with publishers to establish a model that allows immediate access to everything with faster delivery to users, but is cheaper than ILL. Replacing ILL with short-term loans is a positive development; perhaps owning libraries could ask publishers if they could pay for usage when the eBooks in their collections are actually used, or else borrowing libraries could pay a "DDA fee." Or perhaps content from short-term loans could be embargoed until the publisher's production costs have been recouped. Whatever model eventually emerges, it is critical to ensure that eBooks are more portable and accessible to users, not less.

4) Now that many textbooks used in academic courses are available as eBooks, how should libraries handle them? Libraries traditionally do not purchase textbooks for their collections. Publishers are concerned about loss of revenue when textbooks that would have been purchased by many students on a campus become available electronically. Libraries want books in their collections regardless of their use in a class, but publishers want to replicate the course reserve shelf without undermining their market. Course adoption often sustains unprofitable monograph publishing; it will be important to develop models that will be workable for all parties but that will not add to a library's costs. The book rental market is also changing the economics of textbook publishing.

5) What is the future of the scholarly monograph and how can both libraries and end users be accommodated in an age of electronic publishing? Is monograph publishing sustainable in an environment of shrinking budgets and, thus, shrinking purchases? Can this form of scholarship thrive in a digital world? One possibility is for a hybrid purchasing model in which the library buys the book text, and the users (students) pay for added functionality such as searching, the ability to make notes, etc. It is important for libraries to work with publishers to find solutions, and there should be more communication between them. It is also important to recognize that print still matters in an eBook environment.

Mapping a Cloud Strategy and Transitioning From Legacy Systems

Robert MacDonald, Associate Dean for Library Technologies, **Indiana University**, said that cloud usage is booming. He quoted a recently published RightScale "State of the Cloud" Report *(http://www.rightscale.com/lp/2014-state-of-the-cloud-report)* which reported that 94% of today's businesses are using cloud storage. The next major trend will be an increase in public cloud usage; in the last year alone, global spending on public cloud services has dramatically increased, from $47 billion to $170 billion. Many enterprises are taking a hybrid approach to cloud services, using both their own servers as well as public services such as **Amazon's** Web Services. (As the market leader, it is four times the size of every other competitor.)

Libraries must decide when or if they should move their data to the cloud. Key decision points include: Where does my data actually reside? How do I control it? How do I get it into the system and back out? They also need to consider from a cost or service perspective when would be the right time to migrate. There may not be any urgency, and because costs are currently decreasing, it might be prudent to wait. A new type of cloud service, business processes as a service (BPaaS), has recently emerged, in which a user can configure a cloud-based system from parts of several other services. Such an environment gives users more control of a cloud ecosystem, but it may require technical support from people with system administration skills.

Moving to a cloud-based service is a large and potentially transformational change for libraries. **Jill Grogg**, Electronic Resources Coordinator, **University of Alabama Libraries**, discussed the importance of considering the human element of change, noting that if something is not terrifying, it is not truly change! She said that implementing a change with the magnitude of a move from legacy to next-generation systems necessitates a serious self-reflection and a thorough understanding of communication. We communicate every day with many people: our bosses, co-workers, and our children. Understanding communication means understanding negotiation. Analyze and interpret noise, both literal and metaphorical. Change is unsettling, and it makes people anxious. It is important to deal with questions showing anxiety at the time they come up, then act at the appropriate time, provide good feedback, and move decisively.

Alternative Serial Distribution Systems For Libraries

Jonathan Harwell, Head of Collections and Services, **Rollins College**, and **James Bunnelle**, Acquisitions and Collection Development Librarian, **Lewis & Clark College**, said that we need to focus our attention on creat-

continued on page 62