Against the Grain

Volume 26 | Issue 6 Article 36

2014

Don's Conference Notes: Charleston Seminar: Introduction to Data Curation

Donald T. Hawkins
dthawkins@verizon.net

DOI: https://doi.org/10.7771/2380-176X.6971

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

Charleston Seminar: Introduction to Data Curation

The **2014 Charleston Conference** began with the inauguration of a new event: the **Charleston Seminar**, which is envisioned to occur annually from now on and become a series of seminars on topics of high current interest. This year's seminar, ***"Introduction to Data Curation,"*** was a 1-1/2 day event conducted by **Jonathan Crabtree** and **Christopher ("Cal") Lee**, both from the **University of North Carolina (UNC) at Chapel Hill**. **Crabtree** is Assistant Director for Archives and Information Technology at the **Odum Institute for Research in Social Science**[1] (which hosts the country's third-largest archive of computer-readable social science data), and **Lee** is Associate Professor at the **School of Information and Library Science**. Both are extremely well qualified to teach a seminar on data curation; see the **Charleston Conference Website**[2] for further biographical information.

Jonathan Crabtree

Cal Lee

The seminar was structured as a series of talks by the presenters, interspersed with audience interaction and exercises. The first day consisted mainly of the presentations summarized here.

History of Data Curation

Beginning in the 1950s, as organizations began to amass collections of digital data, the realization grew that those collections had long-term value and needed to be preserved. In many scientific fields, the focus on space exploration in the 1960s provided a significant impetus to data collection efforts, and a broadening awareness of the issues resulted in the development of standards and reference models. These efforts culminated in the late 1990s with a reference model for an open archival information system and involved researchers working in diverse disciplines. (Of course, librarians and archivists have been involved with data curation — though perhaps not with that label — as a routine part of their jobs for many years.) Today, digital curation activities can be found in a wide variety of applications, such as physical media properties, digital forensics and data recovery, social and physical science data archives, digital libraries, and medical information.

What is Digital Data Curation?

According to the **Digital Curation Centre**,[3] "Digital Curation" is the active management and preservation of digital resources for current and future generations of users.

The question this **Charleston Seminar** set out to answer was "What knowledge and competencies do professionals need in order to do digital curation work?" and the presentations reviewed a number of them.

The DigCCurr Project[4] at **UNC** has developed a course to prepare students to work in digital data curation, and has organized several international conferences, continuing education workshops, and the DigCCurr Professional Institute. A DigCCurr matrix of competencies needed to undertake a digital curation project as well as to organize a digital curation education curriculum is presented on the DigCCurr Website. It lists six major areas:

1. Mandates, values, and principles, including core reasons why the digital curation functions and skills should be carried out,
2. Functions and skills,
3. Professional, disciplinary, institutional, or organizational content,
4. Types of resources,
5. Prerequisite knowledge needed in order to get other things done, and
6. Steps (transition points) in the life of digital objects.

What Makes Data Different From Documents?

You may have noticed that the seminar focused on data curation, not document curation. According to Wikipedia,[5] "pieces of data are individual pieces of information" and "data as an abstract concept can be viewed as the lowest level of abstraction, from which information and then knowledge are derived." Data thus is any information that can be stored in digital form, such as text, numbers, images, video, software, etc., and it therefore includes documents. Curation practices for documents may vary, but once digitized, the issues relating to them are the same as those relating to data.

The **National Science Foundation (NSF)** has defined four categories of data:

1. Results of laboratory experiments,
2. Records of operations,
3. Observations from sensors or surveys, and
4. Computational simulations and algorithms.

Data can occur in many formats, in contrast to documents which generally contain only text and images.

The Data Curation Process

It is important to understand that digital data curation is an active and ongoing process, and understanding the research data cycle is critical to building relationships with data producers. Some objectives of data curation work are:

- Preserve research data,
- Enable possibility for secondary use,
- Understand the research context where data was created,
- Help next generation researchers discover the data,
- Help researchers understand their appropriate uses, and
- Understand collaboration points with research teams.

Data curators need to focus on quality issues, understand the difference that file formats make, and understand discipline-specific needs. Data curation does not mean just storing the data in a database. Most research outputs include both digital objects and data sets, and often digital objects depend on multiple files having a complex relationship to each other.

Challenges in the data curation process include a wide variation in data citation formats (in contrast to those generally employed for publications, which are fairly universal), missing data, proprietary software used to create and gather the data, and design of the research project. Faced with these sometimes daunting challenges, the data curator may be tempted to simply convert the data to text for storage, but that approach is not good curation practice because much of the representation information may be lost. Treating all formats, metadata representations, and non-textual data in the same way is very dangerous and must be avoided.

Challenges in the Representation of Digital Information

It is well known that the context of information is never captured completely. Information professionals work to bridge the gaps in what is not captured by adding metadata to the information and developing environments for storing and replicating information.

continued on page 73

Don's Conference Notes
from page 72

Representations and interpretations of digital objects are complementary (although multiple interpretations are possible). Every digital object has physical, logical, and conceptual characteristics, and preservation makes the information represented by the objects useful. We must realize that there is no such thing as benign neglect of digital objects; they change and degrade over time, so preservation strategies are important. Extensive information on data preservation and archiving is available in the literature, and organizations such as the **Internet Archive**[6] have played a leading role in the development of such operations.

Data Management Plans and Data Curation Profiles

Data management plans are now required in proposals by funding agencies such as **NSF**. They contain information on how the data generated by a research project will be collected, processed, stored, and preserved, and cover issues such as:

- Access to the data,
- Sharing and re-use policies,
- Data standards and capture,
- Metadata,
- Storage and preservation of the data (including backup), and
- Security.

(For further details, watch for my report of the recent CENDI/NFAIS workshop on data quality in an upcoming issue of ***ATG***.[7])

A data curation profile describes the origin and lifecycle of data during a research project and is designed to capture the requirements for data as developed by the researchers. Using the profile, librarians and archivists can make decisions on how to archive and store it, based on scholars' needs and potential uses of the data. Reasons to develop data curation profiles include:

- To provide a guide for discussing data with researchers,
- To give insight into areas of attention in data management,
- To help assess information needs related to data collections,
- To give insight into differences between data in various disciplines,
- To help identify possible data services, and
- To create a starting point for curating a data set for archiving and preservation.

A data curation profile "toolkit" has been developed jointly by the **Purdue University Libraries** and the **Graduate School of Library and Information Science** at the **University of Illinois Urbana-Champaign** and is available for downloading.[8]

Metadata

Metadata, long created and used by information professionals, is commonly defined as "data about data," but it also can refer to data that facilitates the management and use of other data. It has been described as "the curator's best friend" and is essential in managing digital resources because it preserves their context, facilitates rights management, controls versions, and supports preservation. The *Framework of Guidance for Building Good Digital Collections*,[9] published by the **National Information Standards Organization (NISO)**, lists six principles applying to good metadata:

1. Ensure that it is appropriate to the materials in the collection,
2. Supports interoperability,
3. Uses standard controlled vocabulary terms,
4. States the conditions for use of the digital object,
5. Is authoritative and verifiable, and
6. Supports the long-term management of the collection.

The Dataverse Network

The Dataverse Network[10] is an open-source archiving software application that was developed at the **Institute for Quantitative Social Science** at **Harvard University**. According to its Website, its purpose is "to publish, share, reference, extract and analyze research data. It facilitates making data available to others, and allows one to replicate

continued on page 74

Don's Conference Notes
from page 73

work by others. Researchers, data authors, publishers, data distributors, and affiliated institutions all receive appropriate credit."

On the second day of the seminar, attendees performed an exercise to become familiar with the Dataverse Network and then each individual developed a workflow and prepared an action plan appropriate to his/her own environment.

Based on the attendee evaluations, this initial **Charleston Seminar** was a success. Attendees liked the mix of theoretical and practical information, despite the amount of material presented. Over 80% of them said they would attend another **Charleston Seminar** in the future. One comment summed it up well: "It ran very on-time. And they fit everything in! Very impressive."

Donald T. Hawkins *is an information industry freelance writer based in Pennsylvania. In addition to blogging and writing about conferences for* ***Against the Grain****, he blogs the* ***Computers in Libraries*** *and* ***Internet Librarian*** *conferences for* ***Information Today, Inc. (ITI)*** *and maintains the* ***Conference Calendar*** *on the* ***ITI*** *Website (****http://www.infotoday.com/calendar.asp****). He recently contributed a chapter to the book* ***Special Libraries: A Survival Guide*** *(ABC-Clio, 2013) and is the Editor of* ***Personal Archiving****, (Information Today, 2013). He holds a Ph.D. degree from the* ***University of California, Berkeley*** *and has worked in the online information industry for over 40 years.*

Endnotes

1. *http://www.irss.unc.edu/odum/contentSubpage.jsp?nodeid=657*
2. *http://www.katina.info/conference/conference-info/events/data-curation/*
3. *http://www.dcc.ac.uk/about/what/*
4. *http://ils.unc.edu/digccurr/digccurr-matrix.html*
5. *https://en.wikipedia.org/wiki/Data*
6. *http://www.archive.org*
7. **Donald T. Hawkins**, "Data Infrastructure: The Importance of Quality and Integrity — A CENDI/NFAIS Workshop," ***Against The Grain****, upcoming issue, 2015).*
8. *http://datacurationprofiles.org/*
9. *http://www.niso.org/publications/rp/framework3.pdf*
10. *http://thedata.org*

Collection Management Matters — Frienemies: Vendor Tech Support

Column Editor: **Glenda Alvin** (Associate Professor, Assistant Director for Collection Management and Administration, Head, Acquisitions and Serials, Brown-Daniel Library, Tennessee State University, 3500 John A. Merritt Blvd., Nashville, TN 37209; Phone: 615-963-5230; Fax: 615-963-1368) <galvin@tnstate.edu>

Two of the many responsibilities that I juggle are being the administrator for both our link resolver and facilitating access to our online journals. To have both of these services function effectively, I have to communicate with the vendors' technical support departments on a regular basis. When these people are responsive and genuinely care about making the product perform as advertised, things can be resolved fairly quickly and satisfactorily. However, if the support department does not really know what a link resolver does or understand why your access to the journal results in an error screen, it can lead to a long, drawn-out, frustrating, and sometimes futile effort.

Our former Dean was forward thinking and loved library innovation and technology, so consequently, when we migrated to **Innovative Interfaces (III)** in 2005, we purchased a couple of products that looked wonderful in the demos, but no one had the will or the skills to implement them once they were ours. One of these was our link resolver. We knew what it did, but even after our Webinar, we were clueless as to how to make it work. Both the Webmaster and computer specialist, who back then doubled as the systems person, would not take it on. Not wanting to waste money and seeing its potential for helping students link to full-text articles, non-techie me decided to make an attempt to implement it. After I had some initial success, with heavy support from the **III** HelpDesk and the WebBridge Listserv, I decided to keep going and install the link resolver in every database that was open URL-compliant. Thus began my love-hate relationship with vendor tech support.

Some tech support departments are very helpful and will even go to the extent of using a guest login, so they can have the same user experience you are describing to replicate the error. Technical support at two of my major vendors were very helpful when I was implementing WebBridge, and they even checked back with me to see if I was satisfied with the solution. "Jerry" at a third aggregator's site shared advice about copy/pasting the URL into Notepad and how to get rid of white space. If it was not an issue on his end, he made helpful suggestions about how I could remedy the situation on my end and encouraged me to call him back with the results. But he moved on, and the folks that followed were not as helpful. For instance, I found a page on their support site that had the open URLs for one of their subsidiary products. Tried as I may, I could not get any of them to work. I contacted technical support and was told that open URL linking for that product was not supported. When I sent a screenshot from their support Website that displayed the (erroneous) open URLs for the subsidiary databases, the tech told me that she would check with the product manager. After sending follow-up inquiries for a month, I received an email from the same rep that said the open URLs were not supported for the product — virtually the same wording as her first response. The page with the errant URLs disappeared from the vendor's support site.

Even more aggravating are the vendors who hire technical support personnel who do not have sufficient experience with open URL linking. I had problems getting the link resolver to work in one database of a large periodical vendor. When I contacted the **III** HelpDesk, they said that the problem was with the database vendor. After much back and forth, I was finally put in touch with a senior tech support supervisor who did not understand what the problem was, although I kept sending screenshots with explanations. When I found myself sending email with definitions of open URL linking and explaining how it worked, I realized that if I had to explain it to her on that level, there was no way she was going to be able to help me. In desperation, I went back to **III** and explained that the vendor was incapable of solving the problem, and they resolved the issue for me. This same vendor listed the WebBridge link twice on each citation and could not remove it. Even today, they cannot just have the link resolver show on abstracts only. It offers "all or nothing," so the link resolver button has to appear on every article citation or not at all.

Over the years I have learned some tell-tale signs of when to know whether or not I am dealing with someone who can actually solve the problem once it lands in their lap:

a) They give you bad advice about what to do to solve the problems, without testing their solutions themselves and when those fail, then

b) They don't respond to your email about what progress they are making with solving the issue, until,

c) They tell you to check the link resolver listserv and the wiki to see if you can solve the problem yourself — as if you have not done that already! Many a time my hands have been poised over the keyboard preparing to write a nice-nasty note saying, in effect, "You did not ask me, but I have already done that!" Then I figured what good would it do? They obviously cannot help, so I move on to the next option.

My experience with an article delivery service taught me that things can always get worse. After being assured that they had a WebBridge expert to help me implement the service, I received a corrupted coverage load and a manual written by another **III** library system's department. I got it up and running except in one important database with heavy usage. I offered a guest login, which they ignored, and every solution they sent was

continued on page 75