# The data life cycle applied to our own data

**Abigail Goben, MLS; Rebecca Raszewski, MS, AHIP**

See end of article for authors' affiliations.

DOI: http://dx.doi.org/10.3163/1536-5050.103.1.008

Increased demand for data-driven decision making is driving the need for librarians to be facile with the data life cycle. This case study follows the migration of reference desk statistics from handwritten to digital format. This shift presented two opportunities: first, the availability of a nonsensitive data set to improve the librarians' understanding of data-management and statistical analysis skills, and second, the use of analytics to directly inform staffing decisions and departmental strategic goals. By working through each step of the data life cycle, library faculty explored data gathering, storage, sharing, and analysis questions.

## INTRODUCTION

In 2003, the National Institutes of Health mandated that researchers provide data-sharing plans for grant applications requesting more than $500,000 [1]. This mandate, combined with the requirements of the 2011 National Science Foundation's (NSF's) Data Management Plan and other emerging restrictions for funding [2], has contributed to a greater interest in data sets, as well as their sharing and reuse in the health sciences. Librarians could benefit significantly from being directly exposed to the stages that data move through, from creation until deletion or destruction, a process commonly called the data life cycle. Understanding the data life cycle is essential both in their own work and as they collaborate with researchers. Firsthand experience assists not only in developing solutions for the potential barriers in gathering and analyzing data, but also in curating data sets and discovering relevant external data sets. Librarians should also recognize that they themselves generate many valuable data sets as part of their everyday workflow. This case study provides a review of a process developed around a commonly available data set: reference desk statistics. This process can be easily employed to provide librarians with a self-directed opportunity to enhance their data-management, data-curation, and data-analysis skills.

## CONTEXT

The Library of the Health Sciences (LHS-C) at the University of Illinois at Chicago (UIC) is one of the largest health sciences library systems in the United States. LHS-C has been subject to many of the same circumstances affecting libraries throughout the nation. In fall 2011, resource redistribution had reduced support staff availability in the information services department by 50%. To compensate for this loss, the information services faculty was charged by the department head to identify potential solutions. This led to the goals of modernizing desk statistics collection and reconsidering the information services staffing model. One important change was that all future data collection would be performed electronically. With retroactive digitization of records, both historical research and trend analysis are possible.

At the same time, the information services faculty began receiving increased requests for research data-management assistance. Various tutorials and continuing education had been pursued and reviewed [3–5]. At the time, the available material focused primarily on creating a data-management plan and did not provide librarians with the opportunity to explore their own data as a self-educational tool. It became clear that a working knowledge of the data life cycle from start to finish was essential, and the recently transitioned desk metrics provided a nonsensitive data set to work through the data life cycle model.

## LITERATURE REVIEW

Reference desk data are often the subject of library research. A 2007 study conducted at LHS-C examined both quantitative and qualitative data collected between 1990 and 2005 [6]. This was followed in 2010 by Barrett's examination of reference desk statistics from 1990–2009 at the Crawford Library of the Health Sciences–Rockford, a regional campus of UIC [7]. In both studies, data indicated that reference desk interactions seemed to be on the decline. Other research includes a 2009 paper from McMaster University Library incorporating evidence-based practice into an operational review of the library. Analysis of data collected through a form capturing reference desk statistics in conjunction with a sheet for observational tracking resulted in minimizing dedicated librarian staffing of the reference desk, emphasizing drop-in consultations for complex questions, and improving support staff training [8]. More recently, a 2011 article by Carter and Ambrosi described one methodology for tracking reference desk statistics via tools available from Google; however, the authors did not cover the topic of managing the data set after collection [9].

Recent library research has highlighted the importance of data management. A 2013 systematic review of the emerging roles in health sciences librarianship from 1990–2012 explicitly identified data management librarians [10]. A 2012 article on translational researchers' perceptions of data maintenance presented the library as having a role in areas such as repository management, training in searching databases, and metadata description and discovery [11]. Further, Carlson's 2013 paper identified barriers and opportunities for librarian education in this area, particularly suggesting that levels of engagement with data remained stagnant despite workshop attendance. Barriers to the respondents'

Supplemental Figure 1 is available with the online version of this journal.

**Table 1**
Life cycle stages and identified questions

| Stages | Questions |
|---|---|
| Identifying | ■ What data are available?<br>■ What is the current audience for these data?<br>■ What potential future audiences exist for these data?<br>■ Is this an isolated data set or could it be combined with other sets? |
| Digitizing | ■ Are the data in digital format?<br>■ If no, what would it take to digitize the data?<br>■ Are the data in a stable digital format that can be preserved? |
| Cleaning | ■ How many people have touched or will touch the data?<br>■ What rules have been created to ensure consistent data standardization?<br>■ What tools am I using to standardize the data? |
| Describing | ■ Is there a README.txt file outlining the project?<br>■ Is there a standard ontology applicable to this data set?<br>■ What information would others need to use the data? |
| Storing and preserving | ■ What access is needed to work with the data now?<br>■ Who needs access now?<br>■ What are the best storage options for the future?<br>■ What is the intended duration of preservation? |
| Sharing | ■ Are there any privacy concerns about these data?<br>■ Who is the owner of this data set?<br>■ What institutional policies apply to these data?<br>■ How can sharing rights be maximized? |
| Analyzing | ■ What analysis tools are available?<br>■ What are the limitations of the data set? |

engagement with data curation included organizational support, staffing, and time [12]. Finally, Marshall et al. provided a case study of the data-management process for librarians. While the article provided excellent guidelines, it neither fully explored each data life cycle stage nor discussed library-generated data sets [13].

## METHODS AND MATERIALS

There are several possible templates for describing the data life cycle. Popular examples include those from the Digital Curation Center, the NSF-funded geology project DataONE, and the California Digital Library [14–16]. The life cycle model described by DataONE was chosen as an initial template because it most closely aligned with the intended goals.

The following life cycle stages were included in the final project template: identifying, digitizing, cleaning, describing, storing and preserving, sharing, and analyzing. Questions for each stage were identified (Table 1). Prior to the project, the authors created a sample data-management plan using the Institute of Museum and Library Services template in the DMPTool, which is the data-management planning software available from the University of California Curation Center [17].

### Identifying

Desk statistics between July 2006 and August 2011 were archived in print form. Data tracking was transitioned to a digital format using Google Forms starting in September 2011. Due to the amount of data being captured and limitations of Google Tools, data were removed from the live repository on an annual basis at the close of the fiscal year and archived independently.

### Digitizing

To perform more robust analysis, the print archive required conversion into a digital format. To facilitate this, the authors defined coding procedures (e.g., time stamp should be recorded for the hour window in which the question was asked), and the conversion project was assigned to student staff and temporary support employees. Initially, these procedures were recorded informally on the first author's blog [18] but were then strictly documented as the project continued. Digitization began in May 2012, and the complete print archive was digitized by March 2014. Digitization required far more time and effort than initially anticipated and needed to be distributed over a total of nine different employees. Also, project management was inconsistent until August 2013, when a centralized ad hoc tool was developed by the first author.

### Cleaning

Over the course of the project, it became clear that record consistency correlated inversely with the number of people performing data entry. After reviewing professional literature on librarians performing data standardization [19, 20], the authors identified OpenRefine [21] as an open source tool that would allow efficient data set normalization. Single academic year subunits of the data set were uploaded to OpenRefine as digitization was completed and were standardized based on predetermined rules (e.g., time stamp format of HH:MM:SS). Techniques such as clustering and faceting were used to group similar fields and summarize information in a column or to filter related data aggregates. The standardization processes were documented to enable consistent repetition as new data became available.

### Describing

The project notes document was digitally shared among collaborators, allowing reliable coordination of timelines, instructions to the student employees performing the digitization, and project management

reviews. Records were maintained that contained each column header (time stamp, patron type, question type, notes) and definitions of standardized answers. Also included in a separate text file were known gaps, author contact information, and software packages used for digitization, cleaning, and initial analysis in order to synchronize appropriate metadata with the finalized data set. (Figure 1, online only).

A further text file, supplemental to the data set, provided additional information about the project, including the specific JSON code used to process the data in OpenRefine. Bibliographic and library ontologies were consulted, but most of those were targeted at library collections rather than library events. The Library, Information Science & Technology Abstracts [22] thesaurus was used to determine appropriate subject headings and key terms to associate with the data set to improve future discovery in an institutional repository.

### Storing and preserving

Storage and data access were fundamental issues. Options had to be evaluated for the short term, while the data set was being digitized and gathered, and for the long term, after study analysis had been completed and the data set was ready to be shared. To control access and localize initial data storage during digitization, the university installation of Box [23], an online file management service, was chosen for short-term storage. Using Box allowed convenient collaboration between the authors as well as manageable, limited, and easily revocable user access for employees performing data entry.

Many options and challenges for data sharing and long-term data storage were considered. These included using a subject repository versus a general data repository, cost, potential to embargo the data, and license requirements. The UIC repository, INDIGO [24], was selected as the long-term storage solution for this data set, with a supplemental version stored with the published manuscript on PMC. INDIGO, built on DSpace software, had the advantages of being locally maintained, readily available due to the University Library Open Access Mandate [25], and flexible in terms of file-level publishing. However, it also has limitations: files must be locally downloaded for access, and there is no mechanism for update management. The final data set was released in INDIGO in June 2014 [26].

In the absence of a grant requiring a specific length of time for data set preservation, the initial duration was identified as five years to meet professional journal standards [27]. Beyond the five-year point, the authors intend to preserve the data for as long as it continues to have utility.

### Sharing

In the interest of understanding data-sharing nuances and because desk statistics are not considered sensitive, the authors wished to share the data set under terms as broad as possible for reuse as an educational tool for other librarians looking to work through the data life cycle. This raised questions regarding ownership and rights, as the data were initially gathered as part of the regular work by publicly funded employees at a state institution.

In determining potential rights-holders of the data set, the following possibilities had to be considered: the authors, the University Library (as an independent college), the UIC institution, the entire University of Illinois system, and potentially, the state of Illinois itself. While a few institutional guidelines surrounding copyright assignment are available from the University of Illinois Board of Trustees [28], ultimate licensing policy remained unclear. Due to both the generative and transformational work done by the authors to compile the data electronically, standardize it for analysis, and then format it for potential reuse, it was decided that, in the absence of clear policy and with the permission of the dean of the University Library, the data set could be released through the institutional repository.

Further, in order to facilitate as much potential reuse of the data set as possible and with no perceived risk or income loss to the authors or the university by commercial use, the authors opted for a Creative Commons Attribution 4.0 International License.

### Analyzing

Data set constraints and limitations had to be identified before analysis could occur.

1. Prior to developing and implementing an electronic data-capture solution, question type and questioner could not be resolved to a one-to-one relationship. Historically, the information services metrics had tracked patrons individually and independently of each separate question. For example, if a student asked a directional and an in-depth reference question, two separate questions would be marked, but patron details would only be captured once.
2. Without requiring explicit patron type identification, the default entry was ''UIC Student,'' potentially introducing significant skew toward that category.
3. Similarly, when a patron was recorded without a correlating question type, the default entry was ''Ready Reference,'' introducing skew to that category as well.
4. There were numerous gaps in the data.
5. Human error is believed to have introduced omissions and inaccuracies in actual patron counts.
6. Indirect staffing challenges and changes, including reference desk closures, were not tracked.
7. Acknowledging these issues, the authors considered tools for exploring the data that included SPSS [29], STATA [30], and R [31]. Considerations included cost, size and data set complexity, and learning curve. Analysis was finally undertaken with the most easily available tool, Microsoft Excel 2007, in conjunction with the charting functions and general accessibility of Google Spreadsheets.

## OUTCOMES

The data set from this case study was used to directly argue for hiring additional student employees. Patron

frequency and question type were analyzed, identifying a clear trend of many directional questions with few in-depth reference questions. From this analysis, the information services department petitioned for increased student employee hours beyond the single student employee position that had existed, with the rationale that the repurposed time would enable the faculty to better meet demand for consultations and classes. This review led to an increase in the student employee budget and the addition of more student employee positions to the department.

## DISCUSSION

This data set enabled the authors to achieve three goals: redistributing faculty workload, obtaining practical experience with data through each life cycle stage, and informing future data collection practices.

On a more fundamental level, this exercise changed the authors' general approach to working with data. In addition to the data tools specifically mentioned, it provided the opportunity to explore Event Ontology [32], GitHub [33], DataDryad [34], figshare [35], and OpenDepot [36]. The data set was evaluated against its limitations to determine if there was value in preservation for further analysis. Having completed this data life cycle review, the ongoing data capture of desk metrics was converted from paper to a standardized Google form, which remedied many limitations. Finally, the authors gained familiarity with data collection and standardization challenges as well as facility with commonly used tools and techniques.

This case study focused on creation of a data-management plan, the metadata standardization challenges, and exploration of data storage and sharing options. Future studies may incorporate other information services statistics, including reference desk metrics, and consider the complexities for librarians as they navigate multiple independent data sets. Discussion and clear policies will be critical to minimize all the potential issues that may occur as data sets are shared more broadly in the future.

As legislation and funding requirements emerge surrounding data management, sharing, and reuse, academic libraries are being called on to provide guidance to their institutions and training to both researchers and students. Librarians with a working knowledge of data skills are a resource not only for patrons with questions about their own data, but for their libraries as a whole. By understanding the data generated by the library itself, librarians have the opportunity to apply evidence-based librarianship and demonstrate library efficacy to their institutional administrations. This case study demonstrates that librarians can start with nonsensitive data (e.g., circulation, electronic resources, or facilities usage statistics; publication data; bibliometric analyses) and consider the same questions engaged in this case study, detailed in Table 1, as a means to gain valuable experience.

## ACKNOWLEDGMENTS

The authors thank University Library Dean Mary Case for approving a broad sharing policy for this data set. The authors thank the students and staff who assisted in digitizing the data and Cleo Pappas and Ray Mathew for providing input on early drafts.

## REFERENCES

1. National Institutes of Health. Final NIH statement on sharing research data [Internet]. Washington, DC: The Institutes; 26 Feb 2003 [cited 25 Jun 2014]. <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html>.
2. Goben A, Salo D. Federal research: data requirements set to change. Coll Res Lib News. 2013 Sep;74(8):421–5.
3. Martin ER, Creamer AT, Kafel DM, eds. New England collaborative data management curriculum [Internet]. Worchester, MA, c2011–2014 [cited 25 Jun 2014]. <http://library.umassmed.edu/necdmc/index>.
4. DataONE [Internet]. Albuquerque, NM: University of New Mexico; c2014 [cited 25 Jun 2014]. <http://www.dataone.org>.
5. University of California Curation Center of the California Digital Library. California Digital Library [Internet]. Oakland, CA: The Center; c2014 [cited 25 Jun 2014]. <http://www.cdlib.org>.
6. De Groote SL, Hitchcock K, McGowan R. Trends in reference usage statistics in an academic health sciences library. J Med Lib Assoc. 2007 Jan;95(1):23–30.
7. Barrett F. An analysis of reference services usage at a regional academic health sciences library. J Med Lib Assoc. 2010 Oct;98(4):308–11. DOI: http://dx.doi.org/10.3163/1536-5050.98.4.009.
8. Bayley L, Ferrell S, Mckinnell J. Practicing what we preach: a case study on the application of evidence-based practice to inform decision making for public services staffing in an academic health sciences library. New Rev Acad Lib. 2009;15(2):235–52. DOI: http://dx.doi.org/10.1080/13614530903245311.
9. Carter S, Ambrosi T. How to build a desk statistics tracker in less than an hour using forms in Google Docs. Comput Lib. 2011 Oct;31(8):12–6.
10. Cooper ID, Crum JA. New activities and changing roles of health sciences librarians: a systematic review, 1990–2012. J Med Lib Assoc. 2013 Oct;101(4):268–77. DOI: http://dx.doi.org/10.3163/1536-5050.101.4.008.
11. Bardyn TP, Resnick T, Camina SK. Translational researchers' perceptions of data management practices and data curation needs: findings from a focus group in an academic health sciences library. J Web Lib. 2012;6(4):274–87. DOI: http://dx.doi.org/10.1080/19322909.2012.730375.
12. Carlson JR. Opportunities and barriers for librarians in exploring data: observations from the data curation profile workshops. J eScience Lib [Internet]. 2013 [cited 25 Jun 2014];2(2):article 2; [18 p.]. <http://escholarship.umassmed.edu/cgi/viewcontent.cgi?article=1042&context=jeslib>.
13. Marshall B, O'Bryan K, Qin N, Vernon R. Organizing, contextualizing, and storing legacy research data: a case study of data management for librarians. Issues Sci Technol Lib [Internet]. 2013 Fall [cited 25 Jun 2014]; [10p.]. <http://www.istl.org/13-fall/article1.html>.
14. Digital Curation Center. DDC curation life cycle model [Internet]. Edinburgh, UK: The Center; c2014 [cited 25 Jun 2014]. <http://www.dcc.ac.uk/resources/curation-lifecycle-model>.

15. DataONE. Best practices [Internet]. Albuquerque, NM: University of New Mexico; c2014 [cited 25 Jun 2014]. <http://www.dataone.org/best-practices>.
16. University of California Curation Center of the California Digital Library. Our approach [Internet]. Oakland, CA: The Center; c2014 [updated 13 May 2013; cited 25 Jun 2014]. <http://www.cdlib.org/about/approach.html>.
17. University of California Curation Center of the California Digital Library. DMPTool [Internet]. Oakland, CA: The Center; c2010–2014 [cited 25 Jun 2014]. <https://dmp.cdlib.org>.
18. Goben A. Open access tenure: slog slog slog. Hedgehog Librarian [Internet]. Chicago, IL: Abigail Goben; c2014 [cited 25 Jun 2014]. <http://hedgehoglibrarian.com/2012/05/09/open-access-tenure-slog-slog-slog/>.
19. Heller M. A librarian's guide to OpenRefine. ACRL TechConnect [Internet]. 2013 [cited 25 Jun 2014]. <http://acrl.ala.org/techconnect/?p=3276>.
20. Bedoya J. Analyzing usage logs with OpenRefine. ACRL TechConnect [Internet]; 2014 [cited 25 Jun 2014]. <http://acrl.ala.org/techconnect/?author_name=jbedoyaexchange-fullerton-edu>.
21. OpenRefine [Internet]. c2010–2014 [cited 25 Jun 2014]. <http://openrefine.org>.
22. Library, Information Science & Technology Abstracts [Internet]. Ipswich, MA: EBSCO; c2014 [cited 25 Jun 2014]. <http://www.ebscohost.com/academic/library-information-science-and-technology-abstracts>.
23. Box [Internet]. Chicago, IL: University of Illinois; c2014 [updated 10 Jun 2013; cited 25 Jun 2014]. <http://accc.uic.edu/service/box>.
24. Library, University of Illinois at Chicago. UIC INDIGO repository [Internet]. Chicago, IL: The Library; c2005–2009 [cited 7 Apr 2014]. <http://indigo.uic.edu>.
25. Library, University of Illinois at Chicago. UIC library faculty open access policy [Internet]. Chicago, IL: The Library; c2011 [updated 9 May 2012; cited 25 Jun 2014]. <http://researchguides.uic.edu/libraryoapolicy>.
26. Goben A, Raszewski R. 2006–2011 Library of the Health Sciences Chicago Reference Desk interactions data set [Internet]. Chicago, IL: Library, University of Illinois at Chicago; c2014 [cited 15 Jul 2014]. <http://hdl.handle.net/10027/18931>.
27. Medical Library Association. Journal of the Medical Library Association: instructions to authors: data retention [Internet]. Chicago, IL: The Association; c1999–2014 [cited 25 Jun 2014]. <https://www.mlanet.org/publications/jmla/jmlainfo.html#retention>.
28. Board of Trustees of the University of Illinois. General rules concerning university organization and procedure [Internet]. Urbana-Champaign, IL: The University; c2014 Section 4: copyrights; [updated 24 Jan 2013; cited 25 Jun 2014]. [about 20 screens]. <http://www.bot.uillinois.edu/general-rules>.
29. SPSS software [Internet]. Armonk, NY: IBM; c2014 [cited 25 Jun 2014]. <http://www-01.ibm.com/software/analytics/spss/>.
30. STATA 13 [Internet]. College Station, TX: StataCorp; c1996–2014 [cited 25 Jun 2014]. <http://www.stata.com>.
31. The R project for statistical computing [Internet]. Vienna, Austria: R Foundation; c2014 [cited 25 Jun 2014]. <http://www.r-project.org>.
32. Raimond Y, Abdallah S. The event ontology [Internet]. London, UK, c2007 [updated 25 Oct 2007; cited 25 Jun 2014]. <http://motools.sourceforge.net/event/event.html>.
33. GitHub [Internet]. San Francisco, CA, c2014 [cited 25 Jun 2014]. <https://github.com>.
34. North Carolina State University. Dryad [Internet]. Durham, NC: The University; c2014 [updated 2 Apr 2014; cited 25 Jun 2014]. <http://datadryad.org>.
35. figshare [Internet]. London, UK, c2014 [cited 25 Jun 2014]. <http://figshare.com>.
36. University of Edinburgh. OpenDepot.org [Internet]. Edinburgh, UK: The University; c2014 [cited 25 Jun 2014]. <http://opendepot.org>.

## AUTHORS' AFFILIATIONS

**Abigail Goben, MLS,** agoben@uic.edu, Assistant Information Services Librarian and Assistant Professor; **Rebecca Raszewski, MS, AHIP,** raszewr1@uic.edu, Assistant Information Services Librarian and Associate Professor; Library of the Health Sciences–Chicago, University of Illinois at Chicago, 1750 West Polk Street, Chicago, IL 60612

*Received April 2014; accepted August 2014*