Library Assessment and Data Analytics in the Big Data Era: Practice and Policies

Hsin-liang Chen
Long Island University
hsin.chen@liu.edu

Philip Doty
University of Texas
pdoty@ischool.utexas.edu

Carol Mollman
Washington University
mollman@wustl.edu

Xi Niu
IUPUI xiniu@iupui.edu

Jen-chien Yu
UIUC
jyu@illinois.edu

Tao Zhang
Purdue University
zhan1022@purdue.edu

ABSTRACT
Emerging technologies have offered libraries and librarians new ways and methods to collect and analyze data in the era of accountability to justify their value and contributions. For example, Gallagher, Bauer and Dollar (2005) analyzed the paper and online journal usage from all possible data sources and discovered that users at the Yale Medical Library preferred the electronic format of articles to the print version. After this discovery, they were able to take necessary steps to adjust their journal subscriptions. Many library professionals advocate such data-driven library management to strengthen and specify library budget proposals.

Keywords
Big data, data analytics, information privacy, information policy, library assessment

INTRODUCTION
Emerging technologies have offered libraries and librarians new ways and methods to collect and analyze data in the era of accountability to justify their value and contributions. Gallagher, Bauer and Dollar (2005) analyzed the paper and online journal usage from all possible data sources and discovered that users at the Yale Medical Library preferred the electronic format of articles to the print version. After this discovery, they were able to take necessary steps to adjust their journal subscriptions. Many library professionals advocate such data-driven library management to strengthen and specify library budget proposals, for example (Dando, 2014).

As libraries are offering more online resources and services, librarians are able to use emerging tools (i.e., analytics software) to collect more online data. Meanwhile, many libraries are using social media outlets (e.g., Facebook, Instagram) to promote their services and programs. Consequently, those social media outlets collect and own library user data. Several social scientists and librarians raise questions regarding the collection and availability of social media data. Conley and his colleagues (2015) are concerned about what they identify as three important threats to social scientists' collection and use of big data: privatization, amateurization, and Balkanization regarding research support and funding opportunities.

ASIST 2015,November 6-10, 2015, St. Louis, MO, USA.

Because libraries must assess their resources and services to support data-driven decisions, this panel will focus on the perspectives and future agenda of library data analysis/assessment in the big data era. The topics to be discussed are data assessment techniques and development, academic library management and practice, as well as legal and policy issues related to information security and privacy that educational analytics and big data give rise to. In examining the challenges of data collection and analysis, this panel will pose and address a number of questions, including: 1) What are the challenges of applying Big Data in the academic library world? 2) What are some of the emerging trends of analyzing big data in the libraries? 3) How can we thoroughly address the ethical issues surrounding the use of data sources and sets?

Managing Library Data to Support Evidence-Based Decision Making at Washington University Libraries/ Carol Mollman
Over the past decade, the Washington University Libraries, like most academic libraries, have shifted radically in our view of the value and utility of the data we collect. For many years, data collection amounted to filling in worksheets for submission to the Association of Research Libraries, or simply collecting data to prove that we were busy and productive. With the development of an assessment

program in 2006, our attention focused on using data to (1) better understand the students and faculty members we serve and (2) make better library decisions. While we did not think about it this way at the beginning, in hindsight we have gone through at least four stages in our evolutionary management of library data:

- **Phase One**, (around 2006) began with a sweep of all library units to create a master list of all data we collect. Some data collection was discontinued, and we identified a number of areas where data were erratic or non-existent.
- **Phase Two**: focused on data interpretation and in particular building our skills in data visualization. In 2010-2011 we launched a project to graphically interpret our key data sources. We called it "making the numbers speak," and the discussions enabled by this Statistical Report were powerful. The resulting report became the platform for awareness and discussion of our strategic direction, and was used as a briefing tool for our new leadership, as well as the National Council (our donor /advisory group). The graphs were developed in Excel, laying the groundwork to expand to a variety of visualization tools, most recently Tableau.
- **Phase Three**: Introduction of the Balanced Scorecard strategic management framework forced us to look at what data are most critical to our future direction as a library.
- **Phase Four**: Today, we are faced with a volume of data that is so great, we tend to view it in organizational compartments- collections curation, access services, space management, and emerging service lines such as Geospatial Information Systems or digitization projects. A task force is now forming to look at alternatives to our SharePoint intranet for storing and accessing these resources.
- **Next Phase**: finding the "common denominators" of our key data flows so that we can blend or harmonize the sources into more useful configurations. For instance, connecting student outcomes data with library usage data could provide important insights for library service development and programming.

A key question for discussion is: How can we parley the transactional big data analysis that libraries use into collaborations with other groups in the university (such as Institutional Data, IT, or faculty researchers?)

Library Data, Big Data or *Better Data*: Challenges from the Field/Jen-chien Yu

Academic libraries have a long history of collecting data and reporting their analyses. Traditionally library data collection focused on gathering information about library materials, expenditures, staffing, or service activities. The data were often compiled into library statistics and considered as a way to assess a library's resources and performance.

In recent decades, higher education has grown significantly in the area of assessment as a way to demonstrate value and accountability to various stakeholders. Academic libraries have been playing a prominent and leading part in this movement as well. The libraries have developed sophisticated assessment tools and methods and expanded our data collection to include library survey data, qualitative data (interviews, chat transcripts, etc.), social engagement data (from social media sites), usability testing, and collection analysis, just to name a few (Association of College and Research Libraries, 2010). Furthermore, the rise of Big Data makes some data collection tasks easier and faster; it also has enabled libraries to move beyond simply counting and compiling statistical measures and to engage in complex data analysis such as learning analytics (Cox & Jantti, 2012) and research performance analysis (Elsevier).

The University of Illinois at Urbana-Champaign (UIUC) has a population of 43,300 students and 3,400 faculty members (data from FY2014). The UIUC University Library has a collection of more than 13 million volumes (second largest research university library in the United States), 12 branch libraries and employs more than 500 librarians, staff members, and student workers. The Library has not only collected a wealth of data, we also have acquired or developed a wide range of tools for managing, computing and reporting the data. With all these advantages, however, analyzing library data can still be challenging. Why? What are the issues that the technological developments still cannot solve? This presentation will first give an overview of library data management and how it has evolved with the help of new hardware and software tools and the growing focus on evidence-based librarianship. We will also discuss the challenges (new and old) that academic libraries continue to face in the Big Data era.

Integrating Behavioral User Studies with Log Analysis/Tao Zhang and Xi Niu

"Big data" has been a multifaceted and evolving term. The library catalog transaction logs are believed as one of the important sources of big data in libraries, because: 1) the logs are big in size; 2) they are larger than the typical size that traditional

technologies can deal with; 3) the velocity (speed of in and out) of the data is high; and 4) the potential for extracting knowledge is promising.

Although researchers can mine detailed information about users' search behavior from logs, one of the obstacles of log analysis is the lack of contextual information such as users' motivations, information needs, and step-by-step actions. Analyzing logs alone also has the danger of reaching oversimplified conclusions about search behavior without appropriate understanding of the tasks' contexts and users' preferences, recalling the amateurization of "big data" analysis noted by Conley et al. (2015).

In this presentation, we propose a new search behavior assessment methodology by integrating transaction log analysis and behavioral user studies (Niu, Zhang, & Chen, 2014). We believe integrating behavioral user studies with log analysis to uncover the contextual information of search tasks is a valuable approach to addressing these obstacles.

We will introduce our analytics techniques on the library transaction logs, including data collection, sampling, preprocessing, analyzing, visualization, and storage. Then we will review search behavior results from a number of log analysis studies we have conducted for library catalogs and discovery tools. Common findings and behavior patterns on search field usage, facet selections, and query formulations include these: (1) users predominantly use keyword search; (2) use of facets is low, and nested facet selections are rare; (3) most search sessions involve fewer than four queries; (4) the average number of words per query is generally less than three; and (5) more than half of search sessions reformulate the search by adjusting the original keywords. The content coverage of catalogs and discovery tools can affect users' search behavior. For example, users of discovery tools tend to have a higher percentage of keyword searches and a lower percentage of title, author, subject, and call number searches.

We present a case study of behavioral observations driven by log analysis results. We discuss the design of testing tasks for observing how participants selected search fields, used facets to limit search results, adjusted search queries, and selected relevant results. We will correlate the behavioral observations with the results of log analysis to show the utility of integrating data-driven user study with log analysis for assessing users' search activities. Finally, we will review the lessons learned from our experience of integrating behavioral observations and log analysis and discuss how this approach could help avoid some common pitfalls of mining "big data." At the very end, we will have questions for the audience to discuss:

- Should we define "big data" in terms of size (volume, velocity, and variety) or the insights we can get from the data?
- Is bigger size of log data necessarily better than smaller size of log data?
- Does being able to access logs necessarily mean being ethical to analyze logs?

Policy Framework for Academic Library Analytics/Philip Doty

The use of big data analytics in academic libraries involves the questions "how?" and "when?" "If?" has long been answered (e.g., Hinchliffe & Asher, 2014) by reasons such as the supposed need to use empirical data to justify investments in libraries; to demonstrate libraries' contributions to educational outcomes demanded by university administrators as well as state and federal legislators and agencies; and to show that academic libraries align themselves with the growing use of business analytic tools and strategies in the university (e.g., ACRL, 2010; Cox & Jantti, 2012; Dando, 2014; Gallagher et al., 2005; and Murphy, 2014). Whether these and the myriad other reasons for adopting big data analytics in the academy are justified or correct may be moot. In any case, we must: (1) ask how to mobilize the ethical values that matter to academic libraries, (2) maintain compliance with legal requirements under which academic libraries operate, (3) and then use those values and requirements to help us choose which analytics to use, how to use them, and how to maintain those values and requirements in the face of supposedly irresistible technical change.

This presentation identifies and addresses some of the major concerns with the use of big data and educational analytics that concern academic librarians, especially the need for libraries, their home institutions, and other organizational actors to develop comprehensive privacy and security practices (Bollier, 2010; Brantley et al., 2014; Conley, 2015; Hinchliffe & Asher, 2014; and Uprichard, 2014). Major reasons why include:

- The complexity of privacy, e.g., the necessity for informed consent to use library data, especially going beyond easy assumptions about waivers of privacy by users (e.g., boyd & Crawford, 2012, and *Lochner v. New* York, 1905) and librarians' cooperation with vendors' surveillance of use of copyrighted information,
- Inevitable data leaks, e.g., transmission of passwords and real-world identities over the public Internet and Wi-Fi without encryption

- Librarians' direct involvement with individually identifiable educational outcomes, therefore, need for compliance with Federal Educational Records Privacy Act (FERPA) regulations
- Contribution to the creation of a data mosaic or digital persona of individual students and others
- Increased pressure to release research data held by libraries thought protected by Institutional Review Board assurances of confidentiality given by researchers, e.g., conflicts about subpoenas requesting Boston College's (Irish) Troubles interviews (Palys & Lowman, 2012)
- Questioning "data," e.g., as per Gitelman (2013) and Science and Technology Studies.

We must address the imperative to big data and data analytics in academic libraries while engaging questions about legal and ethical requirements a priori and continuously rather than post hoc and sporadically, perhaps beginning with principles and procedures for libraries in Hinchliffe & Asher (2014):

1. Primacy of audits of privacy and data collection
2. Balance of privacy with analytic specificity
3. Elimination of transaction-level data and of collections of users' demographic information
4. Import of informed consent and opt in/opt out
5. Ensuring that vendors maintain high standards of protection and avoiding those who do not
6. Need for institutional and library codes of practice about data and learning analytics.

Thus attendees of the session will better understand the application of big data analytics to academic library data, some of the most important challenges inherent in such applications, and how to develop means to address those challenges.

REFERENCES

Association of College and Research Libraries. (2010). *Value of academic libraries: A comprehensive research review and report*. Research by Megan Oakleaf. Chicago: Association of College and Research Libraries.

Bollier, D. (2010). *The promise and peril of big data*. The Aspen Institute, Communications and Society Program. http://www.emc.com/collateral/analyst-reports/10334-ar-promise-peril-of-big-data.pdf

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*, 662-679.

Brantley, P., Breeding, M., Hellman, E., & Price, G. (2014, December). Swords, dragons, and spells: Libraries and user privacy. Coalition for Networked Information Membership Meeting. Retrieved March 16 from http://www.cni.org/topics/information-access-retrieval/swords-dragons-and-spells-libraries-and-user-privacy/

Conley, D., Aber, J. L., Brady, H., Cutter, S., Eckel, C., Entwisle, H.,…Scholz, J. (2015, February 2). Big data, big obstacles. *Chronicle of Higher Education*, https://chronicle.com/article/Big-Data-Big-Obstacles/151421

Cox, B., & Jantti, M. (2012). Discovering the impact of library use and student performance. *EDUCAUSE Review Online*. Retrieved February 27, 2015, from http://www.educause.edu/ero/article/discovering-impact-library-use-and-student-performance

Dando, P. (2014). *Say it with data: A concise guide to making your case and getting results*. Chicago: American Library Association.

Elsevier. Elsevier Research Intelligence. Retrieved March 3, 2015, from http://www.elsevier.com/research-intelligence

Gallagher, J., Bauer, K., & Dollar, D. M. (2005). Evidence-based librarianship: Utilizing data from all available sources to make judicious print cancellation decisions. *Library Collections, Acquisitions, and Technical Services*, *29*(2), 169-179.

Gitelman, L. (Ed.). (2013). *"Raw data" is an oxymoron*. Cambridge, MA: MIT Press.

Hafner, A. W. (1998). *Descriptive statistical techniques for librarians*. Chicago: American Library Association.

Hinchliffe, L. J., & Asher, A. (2014, December). Analytics and privacy: A proposed framework for negotiating service and value boundaries. Coalition for Networked Information Membership Meeting. http://www.cni.org/news/video-analytics-and-privacy/

Lochner v. New York. (1905). 198 *U.S.* 45.

Murphy, S. (Ed.). (2014). *The quality infrastructure: Measuring, analyzing, and improving library services*. ALA Editions.

Niu, X., Zhang, T., & Chen, H. (2014). Study of user search activities with two discovery tools at an academic library. *International Journal of Human-Computer Interaction*, 30(5), 422-433.

Palys, T., & Lowman, J. (2012). Defending research confidentiality "to the extent the law allows": Lessons from the Boston College subpoenas. *Journal of Academic Ethics*, *10*(4), 271-297.

Uprichard, E. (2014, October 13). Big-data doubts. *Chronicle of Higher Education*, https://chronicle.com/article/Big-Doubts-About-Big-Data-/149267/