



Measuring Up: Assessing Accuracy of Reported Use and Impact of Digital Repositories

Authors: Kenning Arlitsch, Patrick OBrien, Martha Kyrillidou, Jason A. Clark, Scott W.H. Young, Jeff Mixer, Zoe Chao, Brian Freels-Stendel & Cameron Stewart

This is a grant proposal narrative for the Institute of Museum and Library Services that was funded in September 2014.

Measuring Up: Assessing Accuracy of Reported Use and Impact of Digital Repositories

Statement of Need

Measurement has long been a part of the library profession. Librarians measure the size of collections, how often they are used, how much they cost, and how many visitors come through our doors, among other metrics. These areas of measurement have generally transferred to digital libraries as we routinely collect and report statistics about websites, digital collections, and institutional repositories using web analytics tools, such as Google Analytics or Webtrends. Web analytics should measure all digital access to a resources and should include metrics that help us gauge visitor behavior and needs: how much time they spent on websites, their general geophysical location, screen size, referring source, downloads, and where visitors went when they left the site. Libraries report the data they collect to their own institutions, professional organizations, and funding agencies like IMLS. Unfortunately the methods of collection and even the terminology lack standards or consistency. Current approaches make it difficult to assess the effectiveness of the investments being made in digital library initiatives.

In this proposal we intend to address two tracks associated with the assessment of use and impact of digital repositories. First, we will examine the challenges associated with web analytics; and second, we will recommend an assessment framework so that libraries may begin to measure the impact of open access institutional repositories (IR).

The Challenges of Web Analytics

Methods of data collection and reporting in the digital realm are complex, requiring expertise that tends to be outside the scope of traditional library skills. Powerful software tools exist that can help libraries gather data and monitor the health of digital repositories, but the skills and knowledge needed to configure the software correctly require IT expertise as well as an understanding of the entire landscape of the digital library. Assessment is most successful when driven by a strategic directive (Arlitsch, OBrien, & Rossmann, 2013) from the top of the organization. Based on previous IMLS-funded search engine optimization (SEO) research (Arlitsch & OBrien, Patrick, 2011) we assert there is ample evidence to suggest that most libraries struggle with accurate reporting of use of their digital repositories.

The SEO research conducted by Montana State University and OCLC Research corroborates findings at the Association of Research Libraries and the University of New Mexico. It suggests that current reporting can be grossly inaccurate, leading to a variance in numbers across the profession that makes it difficult to draw conclusions, build business cases, or engender trust. The inaccuracy runs in both directions, with under-reporting usage data as much a problem as over-reporting.

The reasons for inaccurate reporting range from incomplete configuration of web analytics software to a failure to filter robots and spiders from log analysis software. We believe that few libraries, if any, are comparing their web analytics statistics to parallel log analysis statistics (assuming they are using both) to contrast results and reveal anomalies. In addition, there is

no agreed upon definition of terms, or format, with which to share such data across institutions and with organizations such as ARL. Recent work by the Metridoc effort at the University of Pennsylvania suggests a potential framework for reporting such data that our study will investigate (Zucca, 2013).

Privacy issues are also a concern when using web analytics or log analysis software. While libraries and their professional associations vigorously opposed the USA PATRIOT Act and its implications for the privacy of library records (Matz, 2008), there has been relatively little outcry about the information that Google and other software vendors collect about library users through their reporting software. Few librarians are likely aware of the threats posed to privacy by web analytics software and this lack of awareness limits librarian engagement in demanding or developing methods for their users to opt in and out.

Open Access IR, Citation Rates and University Rankings

IR have been in development for well over a decade, and many have accumulated significant mass. While IR were initially defined as scholarly in content (Crow, 2002), some libraries now define them much more broadly and are including institutional records and other digitized materials. This section of the proposal focuses on the scholarly content within IR.

The business case for IR is built in part on the number of downloads of scholarly papers they sustain. Initial evidence we have gathered demonstrates that PDF and other non-HTML file types (which comprise most scholarly IR content) are often not counted by web analytics software when they are downloaded because search engines like Google Scholar can bypass the code that records the download transaction. Based on a representative five-day window of statistics recorded from one university we can extrapolate to a safe assumption that 8,000-10,000 PDF downloads per year are not currently being recorded or reported for that relatively small IR.¹ If even a fraction of that calculation holds true nationally it becomes obvious that the impact of IR is dramatically higher than currently perceived.

Some studies have demonstrated a positive, if tenuous link, between Open Access IR and author citation rates (Norris, Oppenheim, & Rowland, 2008). However, no comprehensive studies currently exist to prove or disprove this connection and that may be due to the fact that it can take years to produce trustworthy results for a single IR. The citation lifecycle of articles spans years and with embargos from commercial publishers the articles deposited in IR may suffer from an even longer period before potentially affecting citation rates (Brody, Harnad, & Carr, 2006).

Our previously published research (Arlitsch & O'Brien, 2012) and (Arlitsch & OBrien, 2013) demonstrates that IR can make scholarly works more accessible via search engines. Our hypothesis is that helping research become more accessible sooner, via an IR should lead to increases in author citation rates, which in turn would eventually improve worldwide university rankings. However, there are no datasets that would allow a researcher to perform this analysis with any statistical confidence. Moreover, the library community lacks any web metric standards a researcher would need to verify or compare results across institutions. *The Times Higher Education Supplement* uses citations to calculate 30% of a university's ranking score and goes so far as to call the metric "the most single influential of the 13 indicators" (TSL

¹ A preliminary analysis of the University of Utah IR conducted by Arlitsch and OBrien found that one method of access being undercounted (requests to download PDF files via Google Scholar users) was at least 110 and most

Education Ltd., 2013). The *Ranking Web of Universities* explicitly calculates the impact of IR by taking into account the number of “rich files (PDF, doc, docx, ppt) published in dedicated websites according to the academic search engine Google Scholar” (Cybermetrics Lab - CSIC, 2013). IR that can demonstrate their ability to affect citation rates could elevate their prominence in the academic research enterprise.

Measuring the intended results of SEO with metrics that demonstrate the effectiveness of IR represents a real value opportunity and requires more research to answer the following questions:

1. What metrics lead to better strategic and operational decisions for digital library resource management?
2. What situations and conditions lead to untenable variances in reporting, i.e. unexplained error variances?

Impact

This proposal supports Goal 3 and Goal 4 from the IMLS strategic plan for 2012-2016, *Creating a Nation of Learners* by providing the necessary frameworks, data models and best practices needed to establish baselines, measure progress and make informed policy decision concerning digital library resources.

The current state of web analytics reporting in libraries is ad hoc. Achieving IMLS goals requires development and implementation of a consistent web analytics framework and pragmatic data model. Our research will lead us to propose practices that will help libraries produce accurate reports on the use of their digital repositories, and they will help professional organizations assess performance of digital library efforts. Funding agencies, in turn, will use the data to make stronger cases to their source of funds, such as the U.S. Congress (Institute of Museum and Library Services, 2012).

We plan to recommend an assessment framework that will help libraries collect data and understand root causes of unexplained errors in their web metrics. Our proposed research will position us well to offer best practice recommendations and engage in data collection and benchmarking with data that are trustworthy and reliable. The recommendations will provide a foundation for reporting metrics relevant to outcomes based analysis and performance evaluation of digital collections and IR.

Working closely with professional associations, such as the Association of Research Libraries (ARL), we will introduce standard web metrics for performance evaluation and outcomes based assessment of access and use to digital collections and IR. The research will provide the framework for the reintroduction of web metrics for digitized special collections and IR as an assessment measurement in ARL statistics reporting. ARL in collaboration with the University of Tennessee have examined the capability of Google Analytics as a way of measuring usage of digitized special collections. The approach is promising and through the grant activities we can expand this application in a more consistent and systematic way across institutions.²

² LibValue: Digitized Special Collections webcast, Aired August 15, 2013. Available on the ARL YouTube channel <http://www.youtube.com/watch?v=X_7uXYKgSSc&feature=youtu.be>

A second impact will be the development of a research framework to capture the web metrics needed for outcomes based analysis of IR efforts. If the recommendations for SEO and web metrics are adopted, the library community will have data that enables more advanced mathematical and bibliometrics models to evaluate the statistical relationships between IRs, citations and University rankings. This data will also help ensure decision makers can make informed policy decisions around the creation and maintenance of IRs in the future (Institute of Museum and Library Services, 2012).

A third impact will be raising awareness regarding user privacy and building consensus for establishing privacy policy language and best practices for web analytics in libraries. Research enabled by this IMLS grant will help advance the understanding of web analytics as it relates to privacy for the end-user. Many libraries utilize web analytics software to understand user behavior, yet the extent to which end-user privacy is considered and protected is unknown. Developing and communicating privacy best practices will help ensure that web analytics are collected anonymously and in aggregate.

Project Design

We propose a research and outreach partnership that will address two **specific areas of research** related to assessment of digital special collections and IR.

1. Improve the accuracy and privacy of web analytics reporting on digital library use
2. Recommend an assessment framework and web metrics that will help evaluate digital library performance to eventually enable impact studies of IR on author citation rates and university rankings.

Goals

1. **Goal: Gather data about the state of web analytics reporting in academic libraries**
 - a. Objective: Conduct survey of academic libraries to gather data about existing practices and priorities.
 - b. Objective: Identify common scenarios that contribute to inaccurate reporting about access and use of digitized special collections and IR
2. **Goal: Establish best practices for web metrics for digital collections and IR**
 - a. Objective: Research options and propose solutions to improve existing standards that reduce variance of web metrics across institutions
 - b. Objective: Make recommendations for configuring web analytics software
3. **Goal: Raise awareness of privacy issues associated with web analytics software and propose solutions**
 - a. Objective: Develop best practices libraries can use to avoid the undue passing of private user information to software vendors
 - b. Objective: Develop policy language describing risks and benefits of web analytics software to users
 - c. Objective: Identify areas where solutions are lacking or create potential issues that may compromise libraries as a trusted source of information and learning.
4. **Goal: Establish a common understanding of what will be measured and provide an assessment framework that identifies digital collection and IR metrics in the spirit of IMLS Strategic Goals 3 & 4.**

- a. Objective: Identify and define metrics that evaluate the public's access to the information and content found in library digital collections and IR
 - b. Objective: Develop recommendations on a framework that IMLS, and other funding providers, can use to help ensure decision makers have the data needed for making informed policy decisions about digital collections and IR.
- 5. Goal: Develop a structured data model that improves machine access and measurement of digital collections and IR content**
- a. Objective: Recommend an IR data model that defines metrics required for an IR citation impact study.
 - b. Objective: Recommend a pragmatic data model that defines metrics to assist in evaluating open access to digital collections and IR.
 - c. Objective: Recommend semantic web standards needed to implement metrics for IR and Digital Special Collection that improve web analytics reporting.

Research Methods

We will apply the Object Management Group's (OMG) Business Motivation Model (BMM) to identify common goals and objectives of funding providers and libraries, as well as measures of performance. Based on the conceptual BMM model developed for library digital collections and IR, we will recommend a pragmatic semantically structured data model that describes digital collection and IR performance evaluation using web metrics. We will use the three phased research approach described below to evaluate and refine our strategic frameworks, data models and web metric accuracy.

Research Phase 1 - Baseline: we will collect historical data to evaluate data reporting variances and privacy issues to establish a web metrics baseline.

Research Phase 2 - Basic Evaluation: we will assist ARL libraries in preparing metadata and systems to implement basic evaluation metrics in the spirit of IMLS Goals 3 and 4

Research Phase 3 - Advanced Evaluation : we will develop recommendations for ARL members willing to implement web metrics that enable advanced evaluation and longitudinal impact studies, i.e. the impact of IR on author citation rates and university rankings.

We will iterate each research phase running early experiments at MSU, confirming results in small scale Pilots at UNM, Alpha testing with 1 to 2 initial ARL members, and finally, Beta testing with a larger ARL member group (i.e, 5 to 10). In the process we will compile, refine, or create the necessary "vocabulary" to ensure a common understanding of terms to compare evaluation metrics across library digital collections and IR. For example, we will define relevant visitor segments and identify high value "events" such as the download of a scholarly work by a government agency, higher education institution or K-12 school.

Data Collection

1. Conduct surveys and gather feedback from ARL and OCLC member institutions to better understand their plans, practices, issues, and priorities for web metrics.
2. Data samples from MSU, UNM and participating ARL and OCLC members will include: Analytics reports data; Webmaster Tools data; Web server logs data; Link resolver logs
3. Survey of publicly accessible digital collection and IR websites to check for the presence and basic configuration of web tools and services that create privacy issues.
4. Develop a feedback loop from internal prototyping and testing at MSU and UNM
5. Examine publicly accessible strategic plans of funding agencies and libraries.

Data Analysis

We plan to:

1. Use the survey data to assess which web metrics are being collected, how they are being used in the organization, and identify gaps.
2. Use data analysis tools (e.g., Splunk, R, OpenRefine) on the data above to identify the frequency and size of reporting inaccuracies and privacy issues.
3. Conduct event studies at MSU and UNM to isolate and verify solution impact on reporting accuracy and privacy.
4. Use Semantic Web tools and methods to develop a pragmatic structured data model that improves public access and visibility of digital library resources via Search Engines and enables accurate metrics reporting needed to support IMLS Goals 3 and 4.

Reporting of findings

We will report our findings to IMLS and the library community through publications, conference presentations, webinars, required IMLS reports and other methods described in the Communications Plan section

Data Management and Accessibility

Please see the Digital Content Supplementary form for details on our research data management plans.

Roles of the Partners

Montana State University (MSU) - MSU will provide the following:

1. Work with ARL, and other library organizations, to facilitate the development of a business model incorporating web analytics relevant to libraries and begin to address strategies with special attention to digital collections and IR.
2. Conduct research to identify and measure reporting accuracy and privacy issues related to current web analytics tools. The team will propose a model for collecting and reporting library-relevant web metrics with more consistency, accuracy, and with appropriate privacy.
3. Develop policy language for libraries to use in describing the risks and benefits of web analytics software to users.
4. Work with ARL to develop training that helps research libraries understand how to develop a web metrics plan to evaluate their digital collections and IR performance.

OCLC Research - OCLC Research will work with MSU on the following:

1. Provide data modeling support that accurately reflects existing digital collection and IR data relevant to library web metrics. They will also provide support in managing Research Phase 1 - Baseline data and help prepare metadata for Phase 2 - Basic Evaluation and Phase 3 - Advanced Evaluation studies. They will provide pragmatic guidelines, and best practices, to help libraries create, or migrate, digital collection and IR metadata based upon research findings and recommendations.
2. OCLC Research Domain Experts will reach out to their community to solicit feedback on their web analytics issues and priorities. A survey may be conducted to support these activities.

Association of Research Libraries (ARL) - The ARL team will work with MSU on following:

1. Act as the primary communication bridge into the research library community. This includes surveying their membership to help guide our research, getting input and feedback on proposed strategic frameworks and solutions, co-publishing and presenting research findings, and organizing training activities.
2. Help gain access to strategic initiatives, thought leaders, and library executives for input and feedback on developing relevant web metrics for evaluating digital collections and IR.
3. Provide support in accuracy analysis with analytics expertise and access to ARL data, and data products (e.g., LibValue), for research studies.

University of New Mexico (UNM) - The UNM team will provide the raw dataset and assist with small Pilot studies to verify web metrics plan and structured data model are practical and effective. UNM has a fairly large IR containing over 8,000 scholarly works that uses common open source software (DSpace) and freely available web analytics tools (Google Analytics). They have not been optimized for SEO and are an excellent representative starting case for this research.

Project Resources: Personnel, Time, Budget

The team we've assembled is well equipped to conduct the proposed research and produce solutions that will be useful for the library community. Our key project staff includes significant digital library development experience, as well as IT, data modeling, SEO, statistical analysis, and community outreach expertise. Please see the attached document titled "Key Project Staff" and their resumes.

We have also assembled an Advisory Council of considerable expertise to help guide the Key Project Staff as they conduct research and develop best practices. Two-page resumes for each of the Advisory Council members are included in the Supplementary Documents. The Advisory Council will meet in person at least once annually, most likely at ALA conferences, and will have quarterly conference calls with the team.

Timeline

Months 1-6:

We plan to launch our project on December 1, 2014, which will allow us to conduct a search for the project manager in the months between notification of award and the project start. Once launched the MSU team will begin researching accuracy and privacy issues surrounding page tagging and log analysis software. This will include establishing Research Phase 1 - Baseline at MSU and UNM so that progress can be measured over time. The team will also begin drafting a BMM and web metrics measurement plan relevant to libraries. MSU and OCLC will use UNM data to begin developing a practical data model for gathering the web metrics necessary to measure the goals and objectives identified in the library relevant BMM analysis.

OCLC Research, ARL, and CLIR will publicize the grant award through their communication outlets. ARL will initiate a needs assessment asking research libraries about the desired evidence for successful management and decision making in digital collections and IR.

Months 7-12:

ARL conducts Google Analytics training with authorized Google Analytics consultants on a

cost recovery basis (not part of grant budget) and will coordinate with the team in identifying best practices. ARL will initiate a call for Research Phase 1 - Baseline Alpha participation for testing solutions and allow for the collections of comparable data across DC and IR.

MSU, UNM and OCLC will begin verifying accuracy and privacy solutions in small scale Pilots at UNM. We will meet to refine a draft of the library relevant BMM, web metrics measurement plan, and structured data model. We will seek conference and blog publication outlets to begin raising awareness of our research and findings.

Months 13-18

Continue Research Phase 1 - Baseline Alpha with 1 to 2 initial ARL members.

The team will work closely with ARL Alpha testing libraries to help them understand the relationship between our strategic framework, the structured data model and the web metrics plan. We will begin planning and preparing Alpha testing member data for (Research Phase 2 - Basic Evaluation). We will also continue communicating research through presentations, blog posts, and publications.

Months 19-24

Expand Research Phase 1 - Baseline into a Beta by initiating a second call for participation with an aim of adding 5-10 libraries. We will continue implementing Research Phase 2 - Basic Evaluation with initial ARL Alpha members. The research teams and advisory council will meet to refine the BMM, web metrics measurement plan and structured data model for baseline evaluation of digital collections and IR. Propose web metrics plan and structured data model enhancements needed for Research Phase 3 - Advanced Evaluation. We will continue to communicate research findings. MSU team will begin development of privacy policy language and privacy-related best practices.

Months 25-30

Continue Research Phase 1 - Baseline Beta and start Research Phase 2 - Basic Evaluation Beta with early adopters. During this expansion we anticipate identifying scaling and organizational barriers that will slow adoption of the proposed web metrics plan and eventually impede IMLS Goals 3 and 4. We also plan to begin implementing Research Phase 3 - Advanced Evaluation Pilot at MSU and UNM. Increase communication of research as our findings and recommendations begin to stabilize.

Months 31-36

In the final months the team will focus on outreach and communications. We will continue to offer training through webinar and conference workshops. We will also perform data analysis using descriptive statistics to identify variances across Phase 2 - Basic Evaluation data. While a full analysis will not be possible before the end of the grant we hope to encourage other researchers to continue working with the data by making it accessible via the MSU open access IR ScholarWorks. MSU will conclude development of privacy policy language and privacy-related best practices and will communicate findings.

Budget

The proposal requests \$499,775 in direct funding from IMLS. Although no cost share is required for this research category proposal the partner institutions are offering \$174,333 in in-kind match. A budget spreadsheet is attached to the proposal along with a budget justification.

Communications Plan

To achieve maximum impact this project will need to involve research libraries, both for input and for testing and sharing findings. The unique partnership in this proposal offers powerful possibilities for disseminating our research. **OCLC Research, ARL, and CLIR/DLF** all have significant networks for ensuring that we are addressing genuine needs and that what we learn is propagated to their communities. Each of those organizations has committed to playing a role in communication and dissemination.

OCLC Research has a partnership with 160 research libraries through the OCLC Research Library Partnership (ORLP). Those Partners are predominantly university libraries, but also include independent research libraries, national libraries, museum libraries, and archives. OCLC Research regularly communicates with over 1300 staff from those institutions, and hosts webinars on topics ranging from the very technical to high level policy. Additionally OCLC Research maintains a listserv consisting of 200 ORLP staff who self-identified as being interested in topics related to research information management. OCLC publishes and promotes reports, issues news releases, and keeps in touch with the community via blog posts and Twitter – all of these outreach activities extend beyond the grant partnership to the broader cultural heritage community. Throughout the project we will send progress reports to our listservs and blog about the project at hangingtogether.org.

ARL represents 126 research libraries that employ more than 12,000 professional librarians. ARL regularly communicates with its member libraries through the *arl-directors* list and other programmatically focused lists like *arl-assess*. The StatsQUAL gateway is accessible by all ARL libraries, including about 2,000 of their employees. The LibQUAL+ database is accessible by more than 1,300 libraries and includes more than 3,000 professionals. ARL regularly communicates its programmatic agenda through *arl-announce* and hosts a bi-annual ARL Library Assessment Conference that attracts more than 600 professionals from both ARL and non-ARL libraries. Throughout the project ARL will utilize these communication outlets to inform its community and to draw participants into the research.

The **Council on Library and Information Resources (CLIR)** and the **Digital Library Federation (DLF)** will further serve as conduits to the community by publishing short pieces in CLIR's *Thinking* blog written by team members describing research developments. Longer articles will be published in *CLIR Issues* and presentations will be promoted at annual DLF Forums.

When the grant is awarded OCLC Research, ARL and CLIR will actively publicize it via news releases, various lists, and social media. Later in the project they will communicate project findings through those same outlets, followed by webinars to relay more specific information and demonstration and to allow for discussion. Input will feed back into our work as we wrap up the project. At the conclusion of the project, we will write a full report and publicize it widely throughout the professional library, archives, and museum community.

Throughout the project the team will seek conferences, workshops, and publication opportunities through which to share the outcomes of the investigation and the benefits of implementing the recommendations.

References

- Arlitsch, K., & O'Brien, P. S. (2012). Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar. *Library Hi Tech*, 30(1), 60–81. doi:10.1108/07378831211213210
- Arlitsch, K., OBrien, P., & Rossmann, B. (2013). Managing Search Engine Optimization: An Introduction for Library Administrators. *Journal of Library Administration*, 53(2-3), 177–188. doi:10.1080/01930826.2013.853499
- Arlitsch, K., & OBrien, P. S. (2013). *Improving the visibility and use of digital repositories through SEO*. Chicago: ALA TechSource, an imprint of the American Library Association. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=578551>
- Arlitsch, K., & OBrien, Patrick. (2011). *Getting Found: Search Engine Optimization for Digital Repositories* (National Leadership Grants proposal No. LG-07-11-0345). Institute of Museum and Library Services.
- Brody, T., Harnad, S., & Carr, L. (2006). Earlier Web Usage Statistics as Predictors of Later Citation Impact - ePrints Soton. *Journal of the American Association for Information Science and Technology*, 57(8), 1060–1072. doi:10.1002/asi.20373
- Crow, R. (2002). *The case for institutional repositories: a SPARC position paper*. (No. 223). Retrieved from http://works.bepress.com/ir_research/7
- Cybermetrics Lab - CSIC. (2013, July). Updated methodology: Ranking Web of Universities. *Webometrics Ranking of World Universities*. Retrieved from <http://www.webometrics.info/en/node/19>
- Institute of Museum and Library Services. (2012). *Creating a Nation of Learners; IMLS Five-Year Strategic Plan 2012–2016* (Brochure). Institute of Museum and Library Services. Retrieved from http://www.imls.gov/assets/1/AssetManager/StrategicPlan2012-16_Presentation.pdf
- Matz, C. (2008). Libraries and the USA PATRIOT Act: Values in Conflict. *Journal of Library Administration*, 47(3-4), 69–87. doi:10.1080/01930820802186399
- Norris, M., Oppenheim, C., & Rowland, F. (2008). The citation advantage of open-access articles. *Journal of the American Society for Information Science and Technology*, 59(12), 1963–1972. doi:10.1002/asi.20898
- TSL Education Ltd. (2013, 2014). World University Rankings 2013-2014 Methodology. *Times Higher Education World University Rankings*. Retrieved from <http://www.timeshighereducation.co.uk/world-university-rankings/2013-14/world-ranking/methodology>
- Zucca, J. (2013). Business Intelligence Infrastructure for Academic Libraries. *Evidence Based Library and Information Practice*, 8(2), 172–182.