July 2004

Integration of Non-OAI Resources for Federated Searching in DLIST

Anita Coleman
University of Arizona, acoleman@lpts.edu

Paul Bracke
University of Arizona

S. Karthik
University of Arizona

CONFERENCE REPORT

D-Lib Magazine
October 2004

Volume 10 Number 10

ISSN 1082-9873

Developing a Web Analytics Strategy for the National Science Digital Library

Casey Jones, University Corporation for Atmospheric Research
Sarah Giersch, iLumina Digital Library
Tamara Sumner, University of Colorado, Boulder
Michael Wright, University Corporation for Atmospheric Research
Anita Coleman, University of Arizona
Laura Bartolo, Kent State University

Point of contact: Anita Coleman, <asc@u.arizona.edu>

Introduction

In August 2004, a two-day workshop was held on "Developing a Web Analytics Strategy for the National Science Digital Library (NSDL)". The workshop was sponsored by the NSDL Educational Impact and Evaluation Standing Committee (EIESC) [1] and was jointly organized with the NSDL Technology Standing Committee (TSC) [2]. It brought together 26 representatives from government and industry, as well as some of the projects funded by the National Science Foundation (NSF) NSDL program [3], to discuss how web metrics could be implemented in a pilot study to identify current NSDL use and develop strategies to support the collection of usage data across NSDL in the future. This new pilot follows a study that the EIESC conducted in 2002 to identify and collect basic web metrics data for NSDL.

A bibliography on web metrics was prepared and distributed to the participants of the 2004 workshop [4]. During the workshop, participants first reviewed the processes and technology used to gather web metrics data by two different organizations: the Association of Research Libraries E-Metrics Project [5] and Sun Microsystems [6]. Through a series of breakout and plenary sessions, participants identified high-level goals for the new pilot study, formulated and prioritized a list of desired effects and requirements for collecting web metrics across NSDL, and developed recommendations for implementing web metrics data collection on a project and program level. The workshop concluded with the EIESC and TSC establishing a joint taskforce to lead the pilot study in NSDL over the next year. Web analytics will be used to address two high-level goals.

1. That high quality learning resources be accessible to a large spectrum of the US population
2. That there be value added to users and projects by participating in NSDL.

This workshop report provides a brief history of previous evaluation activities across NSDL and discusses the importance of web analytics to NSDL. After a review of the literature on web

metrics, the report identifies cross-cutting issues that affect implementing web metrics in the upcoming pilot study (e.g., build vs. buy, data ownership and storage, organizational structure that supports ongoing data collection, user privacy); describes the goals and requirements for the pilot study; and lists near term action items for the joint task force. Documents from the workshop, including a preliminary report entitled "Workshop on Web Metrics in NSDL", slides from ARL and Sun Microsystems presentations, participant statements and the web metrics bibliography can be found on the workshop website [7].

Web Analytics

Web analytics is the process through which a web site tracks visitor behavior to determine the reasons for that behavior. The data thus gathered is then used by the site to further its business goals. The web metrics used to track visitor behaviors include referring methods, search terms, technology use, page paths, entry/exit pages, and geo-segmentation. Compound metrics, such as ratios that combine 2 or more single metrics, can also be used for tracking visitor behavior.

Over the past few years, the field of web analytics has matured to the point where user behavior can be analyzed in an effort to improve website design and users' experiences of the websites, and to help companies become more effective in delivering their products and services. Web metrics can be dissected in many ways to produce a sophisticated level of information. However, analyses must be performed in such a way as to create *useful* information that will facilitate acting on that information. Strategies must be developed to successfully utilize web analytics, and currently these analytical tasks are performed by people.

Presently, two types of web metrics tools exist: those using web server log data (log aggregation) and those using page-tagging. Due to the distributed nature of NSDL, using web server log files to track NSDL user behavior has been determined to be infeasible with regard to maintenance, processing, and detail of results returned. On the other hand, the page-tagging method allows individual web sites to control how data are collected on their sites and demonstrate connectedness to NSDL as a whole.

Evaluating NSDL

The NSDL program was begun in Fiscal Year 2000 as an online library of resources for science, technology, engineering, and mathematics (STEM) education. The program's mission has been to support national improvements in STEM education with an emphasis on innovation.

As NSDL has matured, the types of funded project activities have expanded from creating discrete collections to curating content for specific audiences via specialized web portals, and from developing services for a limited audience to conducting workshops around a suite of digital resources and services. Throughout the development of NSDL, key stakeholders (NSF, NSDL Core Integration, EIESC, project PIs, and the NSDL National Visiting Committee) have placed a high priority on developing an evaluation strategy to ensure that future design, development, and evaluation activities contribute effectively towards the larger NSDL educational mission.

To achieve its educational goals, NSDL must assess its use and effectiveness. As part of an evaluation process begun in 2001 to guide the early design work of NSDL, the EIESC conducted a pilot study [8] that considered four questions, one of which was: "How are people using the Library?" Six pilot sites participated in the initial pilot study by providing a standard set of information [9] from their web server logs covering the period April - June 2002. The web metrics portion of the 2002 pilot study proved to be time-consuming for the participating sites, especially while they were getting set up for the study and also while the sites were

reporting data. Additionally, data from the six pilot sites proved to be inconsistent due to differences in reporting standards [10] and tools used [11].

At the NSDL Annual Meeting in December 2002, the EIESC determined that the costs of the web metrics portion of the pilot were greater than could be justified in light of the information gained. EIESC also advised that web metrics collection, in the form undertaken during the pilot study, should not be scaled up across NSDL. Because NSDL is comprised of distributed digital libraries containing a range of digital materials (e.g., learning objects, data sets, metadata, annotations, online communications tools, and web services), the Committee recognized that although web metrics could provide useful data, they felt that a different approach from that used in the 2002 pilot was necessary. While other NSDL evaluation efforts proceeded, the EIESC waited for NSDL to mature, for the field of web analytics to become more robust, and for the tools used to measure web metrics to be extended to aid web sites in making the data more meaningful. By the time of the 2003 NSDL Annual Meeting, there was support from the NSDL community to undertake another pilot study using web metrics. Recognizing that web metrics approaches involve technical implementation, the NSDL Technology Standing Committee (TSC) joined the EIESC to help develop a web metrics strategy for NSDL [12].

Goals and Requirements for the Next NSDL Web Analytics Pilot Study

Workshop participants agreed that the next NSDL web analytics pilot study should include five to seven web sites, comprised of NSDL.org, three or four mature digital libraries and one or two small sites. Data collection should occur over three to six months after deployment of the pilot study technology. Using a phased approach, straightforward metrics should be studied first; more interesting and complicated questions will be addressed later as part of implementing a larger web analytics strategy.

As mentioned earlier in this report, Web Analytics Workshop participants developed two high-level goals to provide direction for the upcoming pilot study, with the objective that web metrics would be used to analyze achievement of those goals. In addition, the goals developed are broad enough to contribute to a larger web metrics strategy for NSDL. Note that the first goal has been separated into two parts so the appropriate metrics could be associated with each part:

1. [Part A] That high quality learning resources be accessible / [Part B] accessible to a large spectrum of the US population
2. That there be value added to users and projects by participating in NSDL.

Through the discussions related to achieving these goals, many potential metrics, issues, solutions, and analytical options were proposed; a complete list can be found in the preliminary workshop report and notes. Below is a list of requirements that workshop participants regarded as high priority for implementing the web analytics pilot study. The technical, organizational and methodological issues implied in these requirements are discussed in the next section.

High-Priority Requirements for Implementing an NSDL Web Metrics Pilot

- Collecting web metrics data must not interfere with user experience (e.g., site performance, 508-compliance).
- There should be a low barrier to entry; data collection technology should be easy for projects to implement and use.
- Projects should have access to their own data as well as to a global view of NSDL-wide data.
- NSDL should own the data.

- Security of data must be maintained constantly.
- Methodology for data collection should be transparent to ensure confidence in analysis.
- Data collection technology must be infrastructure independent.
- It should be possible to track cross-site traffic.
- User demographics must be tracked anonymously.

Issues Affecting Implementation

The high-priority requirements listed above touch on overlapping issues that relate to users, technology, methodology and an organizational structure that supports data collection and analysis. This section considers workshop participants' thoughts and decisions about these issues in relation to the pilot study. We also consider the implications of the issues for a larger web metrics strategy.

Build vs. Buy

Many workshop discussions revolved around how to collect web metrics data. Web server logs are designed to reflect individual web servers. As such, these log files provide disparate information for each NSDL web site, and user interaction among distributed NSDL sites cannot be determined. Additionally, applying rules across all participating sites (such as excluding a robot or crawler) requires time and effort but could still result in differing datasets from sites that are unable to implement the rule or that use different interpretive methods for analysis.

Fortunately, technology has now evolved to the point that page-tagging has become a viable, successful tracking method. To track visitor behavior, this approach involves including one line of JavaScript code on each page that refers to a JavaScript file on the web site. This method allows all specified information to be tracked with minimal effort. In addition, data from distinct web sites can easily be aggregated to demonstrate cross-site traffic and page flows between sites. There is one drawback to the page-tagging method, and that is the inability to capture data from users who do not have JavaScript enabled.

In the first phase of the new pilot study, comparative assessments of web server logs and page-tagging data will be conducted to determine confidence in the page-tagging method. Eventually, all pilot sites will use the same tool to determine the validity of that method. The utilization of the same tool by all pilot sites enables uniform reporting and understanding of the gathered web metrics. This understanding and interpretation of results will enable sites to use the data to form conclusions about the effectiveness of the tool. An analysis working group will focus on cost-benefit analysis regarding use of the data by both individual projects and the NSDL program as a whole.

Due to required implementation time and development costs, workshop participants decided against building a tool from scratch. Instead, buying an existing tool may be more feasible, because NSDL will be able to begin using the tool within a few months of the workshop, and the total cost would be lower. Having determined that buying the appropriate tool is more feasible than building a new one, the key issue with regard to NSDL web analytics became which tool to use. Page-tagging tools are priced based on page views for a site, so the more page views, the more it costs to use a page-tagging service provider. Thus, tool decisions are reduced to considering cost and desired features. In this regard, an analysis [13] of several web analytics tools performed by the online journal *Network Computing* provided NSDL with a basis for making its decision about which tool to buy.

Collecting Web Metrics with Mixed Methods

Workshop participants agreed that: a) there are additional means of collecting web metrics in addition to server logs or page tagging (e.g., online surveys); and b) analyzing web metrics data by itself does not reveal completely how sites are used. Participants highlighted the importance of maintaining a knowledge bank about NSDL users (e.g., use cases, user profiles, models of successful and failed visits, and site classifications) to aid in analyses and interpretations of web metrics. Participants also agreed that it is important that web metrics should supplement, and be supplemented by, other NSDL data (e.g., the number of items accessible through NSDL.org and NSDL partner sites; the number of search engines that have indexed the site; and the movement of content via OAI within NSDL and harvested from NSDL partners). Collecting qualitative and quantitative data will help with more complete analyses of user behavior and needs at NSDL sites.

Data Ownership and Storage

As a result of adopting the expanded approach, discussed above, NSDL will need to develop the practices and storage mechanisms for collecting the qualitative and quantitative data that the page-tagging tool cannot. However, since service providers of page-tagging tools host and maintain all data collected, the chosen provider will need to ensure that NSDL owns the information collected and that NSDL can download the information at any time. In addition, due to the distributed projects within NSDL, an agreement will need to be outlined in which both the NSDL program and individual projects are comfortable with the use, analysis, and ownership of the data collected from the individual project sites. These detailed requirements reflect projects' uncertainty about how data will be used to demonstrate the level of success of individual web sites and the NSDL program as a whole. As the field of web analytics advances, their use to assess business goals is constantly evolving as well. NSDL will be challenged to develop sound data collection and storage practices that prevent uninformed conclusions.

Support for Ongoing Data Collection

Analyzing web metrics data is just one part of understanding usage of NSDL and developing meaningful interpretations and actions. Supporting data, as well as staff to support web metrics —both in analysis and implementation—are also necessary for successful web analytics results. In addition, many projects may need to be convinced to participate. This will be especially true if people outside a project will have access to the project's web metrics data. Yet outside access will be necessary on at least a small scale for successful analyses and interpretations. In the future, it is possible that project participation in NSDLweb analytics efforts may require additional data or assistance from NSDL projects. However, such requirements cannot be determined at this time.

End User Issues

Two issues were raised at the workshop regarding effects of web analytics on NSDL users: continuity in service and privacy. Continuity in service needs to be addressed between NSDL and the service provider in terms that demonstrate that any failure of the provider's system to work properly will not affect what the visitor sees in his web browser, the time required to display content in the web browser, and interactions with the web site. Privacy issues need to be addressed at both the NSDL-wide level and within each participating web site. Participating sites will be encouraged to explain in their privacy policies that web metrics data are being gathered but that they will not be associated with any person's name. Additionally, NSDL can pursue efforts to determine if opting-out of the data collection is technically feasible.

One of the workshop participant breakout groups focused on business requirements relating to search. The requirements were stated in terms of NSDL goals in analyzing search and web metrics to discover user actions and goals. Issues that arose included the need to outline what should be tracked at both the individual project and NSDL-wide levels to help guide projects with search analyses and goals.

Organizational Issues

Pilot sites (and any new site implementing a web metrics solution) require identification of the actual tasks that need to be completed in order to participate. Pilot sites also require a low barrier to implementation, and they will need additional aid in both implementation and analysis. Thus, pilot sites need a coordinated contact with whom they can interact. This contact needs to be knowledgeable about all the issues involved in participation and should be able to provide answers to questions as well as provide direction for the project.

Conclusion

Who is coming to NSDL and to its individual sites? What do these users want from the sites? What works and does not work for these users? NSDL will implement a web analytics pilot study using the page-tagging method to determine if it can answer questions such as these. However, answering these questions isn't enough. NSDL must also determine if meaning can be derived from answering these questions to help make web sites easier to use and more effective, to better meet business goals, and to understand NSDL's role as a digital library. While time consuming to implement, collecting the metrics data is easy when compared to interpreting the data.

For the forthcoming pilot to be most effective, NSDL must elucidate its criteria for pilot site selection, structure pilot activities around meeting stated NSDL goals, and develop a process for integrating web metrics data and data analyses back into the NSDL program and web site development processes. A web metrics panel will discuss these isuses and more at the upcoming NSDL Annual Meeting in November, and all those interested are encouraged to participate [14].

Notes and References

1. NSDL Community Portal - Educational Impact, <http://eduimpact.comm.nsdl.org/>.

2. NSDL Community Portal - Technology, <http://technology.comm.nsdl.org/>.

3. National Science, Technology, Engineering, and Mathematics Education Digital Library (NSDL) Program, <http://www.ehr.nsf.gov/ehr/DUE/programs/nsdl/>.

4. Web Metrics Bibliography, <http://dlist.sir.arizona.edu/archive/00000393/>.

5. ARL E-Metrics, <http://www.arl.org/stats/newmeas/emetrics/index.html>.

6. Sun® Microsystems, <http://www.sun.com>.

7. Web Metrics Task Force, <http://webmetrics.comm.nsdl.org>.

8. For a full documentation and reports on the pilot study, see <http://eduimpact.comm.nsdl.org/events/?pager=90>.

9. The Web metrics used by the six pilot study sites included: incoming top-level domains, browser types, referring URLs, number of sessions, pages served, queries conducted and downloads.

10. Reporting problems occurred with needed data not available, lack of time, and technical issues/capabilities.

11. Six sites reported using seven tools to collect the data for the Users and Usage metrics: 1) wusage 7.0; 2) Summary 2.1.1; 3) Sawmill; 4) Netracker; 5) webTrends; 6) in-house; 7) awstats & sourceforge stats.

12. For more information on the joint Standing Committee effort, see <http://eduimpact.comm.nsdl.org/events/?pager=225>.

13. Boardman, Bruce. "Web Analytics Services: Inside Information" in *Network & Systems Management Review*. August 5, 2004. Available at <http://www.nwc.com/showArticle.jhtml?articleID=20003001>.

14. NSDL Annual Meeting 2004. Available at: <http://nsdl.comm.nsdl.org/>.

---

doi:10.1045/october2004-coleman
