Enhancing Content Visibility in Institutional Repositories: Overview of Factors that Affect Digital Resources Discoverability

Ahmet Meti Tmava
University of North Texas, College of Information
Ahmet.Tmava@unt.edu

Daniel Gelaw Alemneh
University of North Texas, Digital Libraries Services
Daniel.Alemneh@unt.edu

Abstract

In the last decade, a growing body of the scholarly information and instructional materials produced by universities existed primarily in digital format. New digital technologies increased the productivity of scholars. The purpose of an Institutional repository (IR) is to manage their scholarly work in ways that facilitate interdisciplinary collaboration, and accelerate the pace of discovery and innovation. Academic institutions have increasingly recognized that IRs are a vital part of the scholarly dissemination infrastructure. The goal of an IR is essentially to collect, preserve, and make persistently accessible a variety of scholarly materials. This paper explores digital curation activities that enhance the visibility of IR in an ever-changing digital landscape.

Keywords: institutional repositories, metadata standards, metadata consistency, search engines

Introduction

Institutional repositories (IR) are defined by Piorun et al. (2007) as electronic systems that capture, preserve, and provide access to the intellectual output of a community in a digital format. The goal of an IR is to collect, preserve, and make persistently accessible a variety of scholarly materials (Palmer et al., 2008). Open access digital repositories have implemented the Open Archive Initiative-Protocol for Metadata Harvesting (OAI-PMH) as a mechanism to achieve interoperability in the exchange of meta-information with other systems (Muhammad et al, 2007). With all the different software platforms employed by IRs their systems need to be able to support interoperability in order to provide access via multiple online search engines.

IRs are responsible for not only creating metadata but also for ensuring that metadata is available in standard schemas and formats that comply with the OAI-PMH protocols. With all the digital content available online today, it is a constant struggle to attract users. One of the challenges facing IRs today is resistance by faculty to voluntarily depositing their works (Albanese, 2009). In order to make their cases, IRs need to assure faculty that their work can be widely disseminated. In this paper we explore factors that enhance the visibility of IR items in an ever-changing online environment.

Institutional Repositories and the Changing Landscape of Scholarly Communication

New digital technologies have increased the productivity of scholars and enabled them to manage their own digital content in ways that facilitate interdisciplinary collaboration and accelerate the pace of discovery and innovation. IRs provide long-term sustainable storage, preservation, and open access to resources. Foster & Gibbson (2005) found that faculty want to be able to make their own work available to others and have easy access to other people's work. IRs also serve as tangible indicators of an institution's productivity, and thereby increasing an institution's visibility, prestige and value.
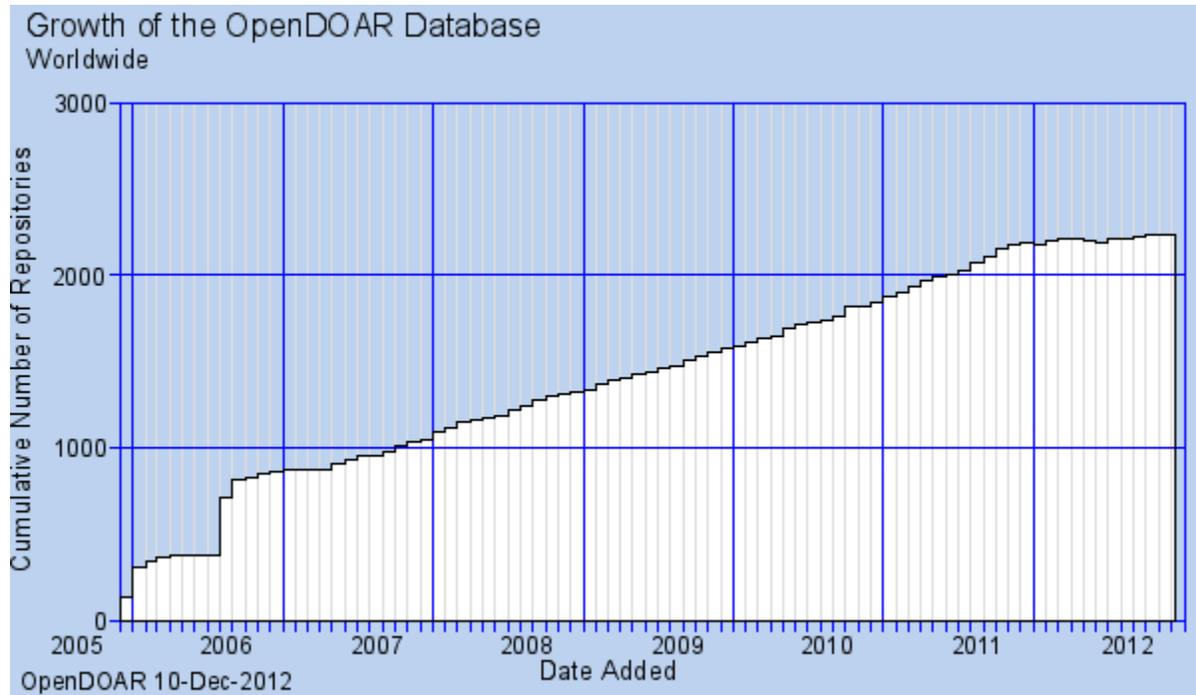

Figure 1: Worldwide growth of Institutional Repositories*.
Source: Open DOAR (Directory of Open Access Repositories), December 2012.

In spite of the increase popularity of IRs worldwide (see Figure-1), a survey conducted at universities in United States and Canada in 2009 found that only 56.7% of faculty were aware that their institution had a digital repository where they can publish their work. And out of which, less than 10% of faculty have ever submitted material to their institution's digital repository (Primary Research Group, 2009, p.17).

In this regard, one of the main challenges for Institutional Repository managers is to make faculty aware that IRs offer them new dissemination opportunities, to reach a much broader audience than what is available through other formal means of publishing. Through open access IRs make their digital object available to a worldwide audience. This can increase citations made to their scholarly work, thus it would maximize their impact factor. The "impact factor" refers to the number of subsequent citations a work receives (Donovan and Watson, 2011).

IRs reform scholarly communication by stimulating innovation in a disaggregated publishing structure (Crow, 2002). In light of the continually evolving information environment and user needs, digital curators (among other stakeholders that are actively involved in supporting IR development) must promote IR use, maintain consistency and participate in the creation of and adherence to national/international standards and institutional-specific policies. Furthermore, it is vital that they understand open standards for searching and effective retrieval of digital content. Open standards enhance interoperability between different IR software platforms, such as DSpace (see Figure-2), and online search engines, such as Google, to index metadata elements consistently and offer effective search capabilities.
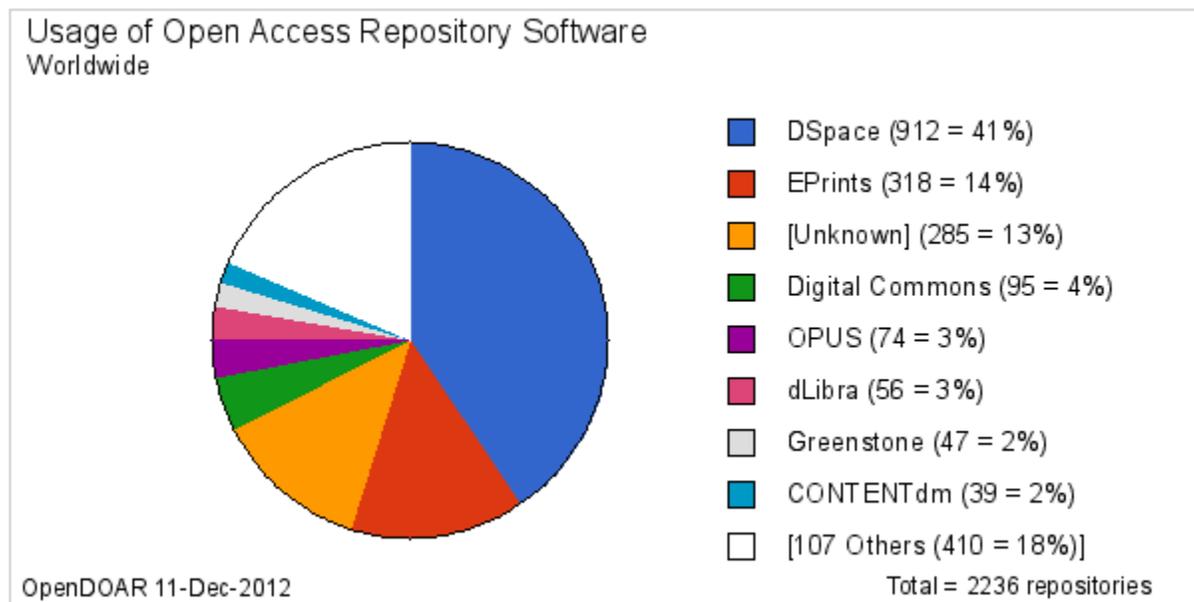
Usage of Open Access Repository Software
Worldwide

DSpace (912 = 41%)
EPrints (318 = 14%)
[Unknown] (285 = 13%)
Digital Commons (95 = 4%)
OPUS (74 = 3%)
dLibra (56 = 3%)
Greenstone (47 = 2%)
CONTENTdm (39 = 2%)
[107 Others (410 = 18%)]

OpenDOAR 11-Dec-2012   Total = 2236 repositories

Figure 2: Worldwide Usage of Open Access Repository Software*.
*Source: Open DOAR (Directory of Open Access Repositories), December 2012.

Current Best Practices and Ways to Increase Content Visibility

Search engines are the most popular way users search for information; they account for 88 percent of users' search time when they are looking for information (Bifet and Castillo, 2005). Most users like to search by keyword, and they expect to retrieve useful results. Google is the most popular search engines because it is very easy to use, it is free, it is extremely fast and produces useful results. For the most part Google generates search results by matching search terms entered with Web page content, usually referred to as full text searching. This form of searching usually has shortcomings, mainly in precision. Precision refers to the proportion of the relevant documents retrieved in a search to the total number of documents retrieved in a search.

Contrary to the long-held belief that Google ignores the data encoded in a Web page's meta tags, Zhang and Dimitroff (2005) show that "metadata is a good mechanism to improve webpage visibility." Metadata describes the nature of the digital items stored in a repository, including content, structure, and access rights. In other words, metadata should provide multiple access points (e.g. author, title, subject). To fulfill its purpose it is important for the digital curators to understand the ability of search engines to index metadata elements and retrieve digital content using their embedded metadata elements.

Web crawlers operated by search engine companies harvest metadata and other information about online objects and send that information back to the search engine. Specialized algorithms (that take many factors into account) analyze the harvested data and decide whether or not to add the metadata to the search engine's index. Farajpahlou and Tabatabai (2011) note that XML, as a syntax ground for implementing the metadata elements of DC and MARC 21, maximizes the interoperability between search engines and metadata initiatives. According to Taheri and Hariri (2012), Google and Yahoo treat XML-based metadata with language-based tags (like DCXML) and without language-based tags (such as MARCXML) in the same way. The same authors noted that all metadata elements of the MARCXML and DCXML standards are compatible with the Google and Yahoo indexing software.

Institutional Repository systems need to be able to support interoperability in order to provide access via multiple search engines. According to Hirwade (2011), "interoperability is the ability of systems, services and organizations to work together and exchange information and use exchanged information without special effort of either system"(p.59). Metadata interoperability facilitates the exchange of information between repositories and enables World Wide Web searching. To make items discoverable in the diverse online environment, the database needs to provide standards-compliant database servers and expose the fullness of that metadata to a metasearch search client (Dorman, 2008). Metasearching, or the process of metadata being searched, enables connections to multiple

resources. However, as described by Dorman (2008), in cases when the IR does not supply a standards-compliant database server there are two alternative access methods that can be used; a proprietary gateway or HTML parsing.

While the metadata standards and standards-compliant servers are helpful, consistency in local practices is still needed because inconsistent metadata adversely affect search results. If the metadata is incorrect, the resources in the IR are not adequately represented and will remain invisible to the users. Yasser (2011) extensive literature review identifies five categories of metadata problems: incorrect values, incorrect elements, missing information, information loss, and inconsistent value representation. Similarly, Alemneh (2008) and Shreeves et al. (2005) explain that metadata problems occur when elements are not applied properly and when values are not consistently or accurately recorded. According to Dorman (2008), no open communication standard can compensate for metadata that lacks content or encoding standards. In the end, the visibility of digital items in an IR depends on the quality and richness of the descriptive metadata that content providers and digital curators provide.

Conclusion

The digital environment has introduced new resource types and new user expectations into the information landscape. In order to deliver a richer user experience and maximize visibility of digital resources, it is critical to have metadata that complies with standards, both in its completeness and its adherence to metadata creation standards. Consistency is the key factor in successfully managing different scholarly works in IRs and in making information more widely and easily available to users. Regardless of the repository software, metadata consistency plays a key role in describing and managing digital objects of different formats to effectively integrate the contents of IR into the existing services and collections. Although there are a number of contributing factors that affect digital resources visibility in IRs, it is the rich metadata that is consistently encoded that makes the digital items more discoverable.

References

Albanese, A. (2009). Thinking Beyond the Box. *Library Journal, 134*(4), 26-28. Retrieved from http://www.libraryjournal.com/article/CA6639327.html

Alemneh, D. G. (2008). Maintaining Quality Metadata: Toward Effective Digital Resource Lifecycle Management. In S. Hawamdeh et.al. (Eds.), Knowledge Management: Competencies and Professionalism. Proceedings of the 2008 International Conference, October 23–24, 2008, Columbus, Ohio (pp. 313-324). World Scientific (Series on Innovation and Knowledge Management – Vol.7). Retrieved from http://digital.library.unt.edu/ark:/67531/metadc29317/

Bifet, A. & Castillo, C. (2005), An analysis of factors used in search engine ranking. Retrieved from http://airweb.cse.lehigh.edu/2005/bifet.pdf

Crow, R., (2002). "The Case for Institutional Repositories: A SPARC Position Paper" Research on Institutional Repositories: Articles and Presentations. Paper 27. Retrieved from http://digitalcommons.bepress.com/repository-research/27

Directory of open access repositories (2012, September). Retrieved September 12, 2012, from http://www.opendoar.org/index.html

Donovan, J. M. and Watson, Carol A. (2011). Citation Advantage of Open Access Legal Scholarship. Research on Institutional Repositories: Articles and Presentations. Paper 4. Retrieved from http://digitalcommons.bepress.com/repository-research/4

Dorman, D. (2008). The potential of metasearching as an "open" service. *Library Hi Tech, 26*(1), 58-67.Retrived from www.emeraldinsight.com/0737-8831.htm

Farajpahlou, A. H., & Tabatabai, F. (2011). How are XML-based Marc 21 and Dublin Core records indexed and ranked by general search engines in dynamic online environments? *Aslib Proceedings, 63*(6), 586-592. doi:10.1108/00012531111187234

Foster, N., & Gibbons, S. (2005). Understanding Faculty to Improve Content Recruitment for Institutional Repositories. Retrieved from http://www.dlib.org/dlib/january05/foster/01foster.html

Hirwade, M. (2011). A study of metadata standards. *Library Hi Tech News, 28*(7), 18-25. doi:10.1108/07419051111184052

Muhammad, A., & Burk, A. (2007). New possibilities for metadata creation in an institutional repository context. *OCLC Systems & Services, 23*(4), 403-410. doi:10.1108/10650750710831547.

Palmer, C. L., Teffeau, L. C., & Newton, M. P. (2008). Strategies for Institutional Repository Development: A Case Study of Three Evolving Initiatives. *Library Trends, 57*(2), 142-167. Retrieved from http://muse.jhu.edu/journals/library_trends/v057/57.2.palmer.html

Piorun, Mary E., Comes, James F., & Billings, M. (2007). The Role of the Institutional Repository in Preserving the Scholarly Record of the University of Massachusetts". Research on Institutional Repositories: Articles and Presentations. Paper 63. Retrieved from http://digitalcommons.bepress.com/repository-research/63

Primary Research Group. (2009). The survey of higher educational faculty: Use of digital repositories and views on open access. New York: Primary Research Group. Retrieved from http://www.primaryresearch.com/view_product.php?report_id=167

Shreeves, S., Knutson, E., Stvilia, B., Palmer, C., Twidale, M., & Cole, T. (2005). Is "quality" metadata "shareable" metadata? The implications of local metadata practices for federated collections. In H. A. Thompson (Ed.), Proceedings of the Twelfth National Conference of the Association of College and Research Libraries, April 7–10, 2005, Minneapolis, MN (pp. 223–237). Chicago, IL: Association of College and Research Libraries.

Taheri, S.M. and Hariri, N. (2012). A comparative survey on the indexing and ranking of the content objects including the MARCXML and Dublin Core's metadata elements by general search engines, *The Electronic Library, 30*(4), 480-491   doi:10.1108/02640471211252193

Yasser, C. M. (2011): An Analysis of Problems in Metadata Records, *Journal of Library Metadata, 11*:2, 51-62. doi.org/10.1080/19386389.2011.570654

Zhang, J. & Dimitroff, A. (2005). The impact of webpage content characteristics on webpage visibility in search engines' results (Part I), *Information Processing & Management, 41*(3), 665-90. Retrieved from: www.elsevier.com/locate/infoproman