A Conceptual Approaches to the Metadata Harvesting

MS. INDIRA DODIYA
Assistant Librarian,
K.S.School of Business Management, Gujarat University
Gujarat (India)

Abstract:
This paper introduces conceptual approach for analyzing issues with metadata harvesting. The approach considers metadata work as an activity that shapes the way metadata is created and used in individual administrative contexts. Over time, governments can adopt and adapt metadata formats in different ways, which can lead to harvesting issues. This is illustrated a analysis of the issues encountered with harvests of Dublin Core from digital libraries. The analysis shows how the different histories of each library led to different metadata implementations, and thus to issues with the harvest. We suggest that Dublin Core metadata harvesting can be a two-stage process, requiring manual alignment and normalization in addition to automated harvesting, and conclude that metadata harvesting is supported by a clear understanding of data provider's history.

Keyword: *Dublin Core, Metadata Harvesting, Metadata, OAI-PMH*

1. Introduction

Metadata harvesting is support the exchange and aggregation of metadata between the digital libraries. Dublin Core metadata harvesting is supported by the Open Archives Initiative Protocol for Metadata Harvesting (OAI -PMH). The various barriers to metadata harvesting have been report. While these barriers can appear ad hoc in nature, it is useful to ask whether they might also be rooted in some wider underlying characteristics of digital libraries. If this is the case, then understanding these characteristics can support the development of more systematic harvesting approaches.

This paper proposes to analyze metadata harvesting. The approach describes digital libraries as organizations whose members are engaged in various kinds of metadata work. In this methodology, metadata work is conceptualized as a technical activity that, over time, shapes how metadata is created and used in individual organizational contexts. One result of this shaping is that, even when individual organizations adopt the same metadata schema, different uses of the same schema can emerge which can then affects metadata harvesting. In this paper, this approach is elaborated through a comparative analysis of harvesting issues of different libraries. This harvesting work was part of a wider plan that is seeking to generate Dewey Decimal class numbers automatically from metadata records. The harvesting issues were sufficiently interesting to warrant further investigation in them. While similar issues have been reported in the past, there has been little systematic analysis of such issues from an administrative perspective.

The paper is structured as follows and provides brief background to metadata harvesting, Dublin Core and OAI-PMH. Next Para describes of metadata harvesting from digital libraries. The metadata unavailable through OAI-PMH, and duplicate metadata records describing the same resource in different terms. It then supplies an administrative analysis that attempts to account for all of these issues in a coherent fashion.

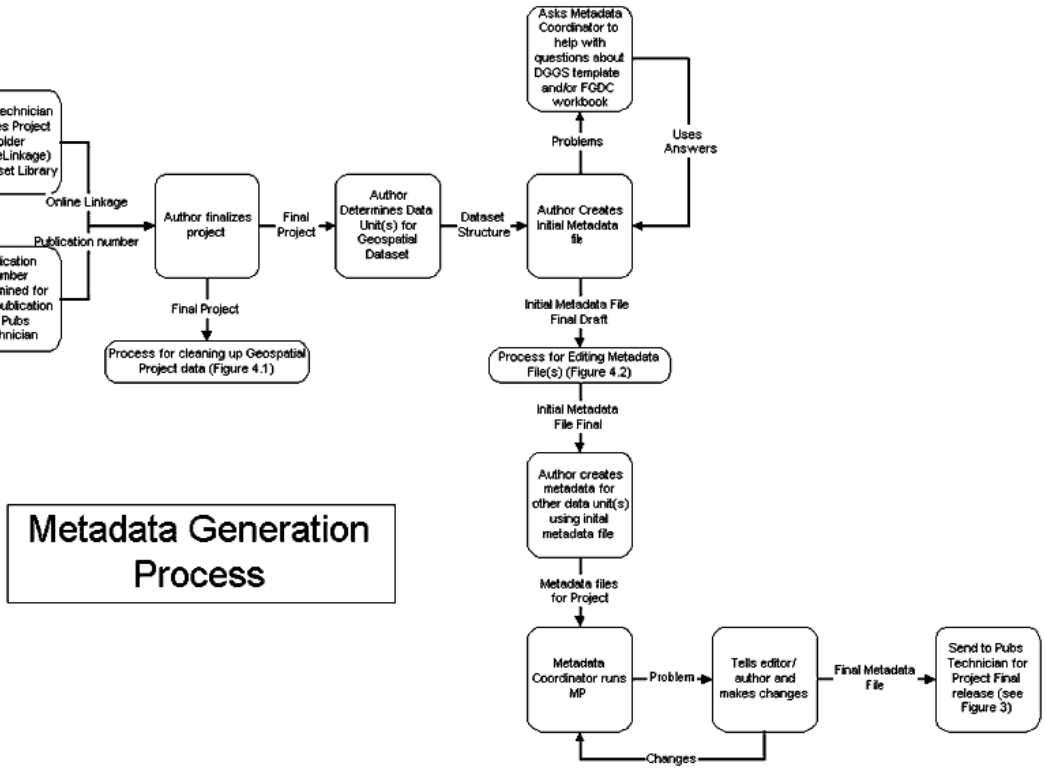
Metadata is by definition "data about other data" of any type and sort in any media. It is used to facilitate the understanding, characteristics and management usage of data. For effective data management, the Metadata should include data that is coherent with the context of use.

This Metadata is used for recovered access to the huge amounts of data stored and managed by different companies. Metadata provides context for data. In data processing, for example, Metadata is definitional; it gives documentation of other data in the application. The term "Metadata" should be used carefully since all data is about something and hence is "Metadata". This page itself includes metadata in the structure of Meta tag. Right-click anywhere on the page, select "View source" or "View page source" it is depending on your browser, and search for "meta." The content of the Meta tags you find provides information about various elements on perticular page; this data is not visible to the reader, but is searchable and used in many ways.

A library's online catalog displays much of the metadata about a publication.

2. Metadata Harvesting:

Metadata harvesting is a procedure for aggregate metadata from individual digital libraries, for instance to build a central metadata repository. The harvesting discussed in this paper used the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) to harvest Dublin Core metadata. The development of OAI-PMH was based on earlier work on E-print interoperability in repositories that sought to make E-prints more widely available. In contrast to formats such as Z39.50, which convey richer and more granular detail but which also require more effort to assume early E-print metadata harvesting was therefore intended to be a low barrier technique to resource discovery, which could then support third party development of repository tools and services. As the wider possibilities of metadata harvesting became apparent, OAI-PMH was subsequently developed, in which libraries (data providers) could expose their metadata on servers where it could be harvested by third parties (service providers) and then reuse in various ways in appropriate field.



Metadata Generation Process

In terms of functionality, the technical characteristics of metadata such as standardized elements, attributes, definitions, and relationships provided suitable benchmarks for evaluating harvesting. From this perspective, the early challenges for metadata harvesting were anticipated to be technical in nature. While OAI-PMH supports the harvesting of unevaluated Dublin Core, the survival of different flavors of qualified Dublin Core in different repositories was seen as potential barrier to harvesting describes a comparative study of OAI-PMH implementations in the Mellon Metadata Harvesting Initiative; among the issues encountered were lack of resources hindering the work of data providers;

varying and conflicting applications of Dublin Core by different data providers, which required metadata normalize after harvesting duplicate metadata records and a general lack of awareness of the full metadata lifecycle associated with harvested metadata reports how the use of OAI-PMH to harvest metadata from NSDL way to a central NSDL repository was demanding of Pathway's organizational resources (time, staff, expertise, etc.), and in cases where those resources were inadequate, the quality of harvesting was affected unfavorably. The harvesting essential ongoing human intervention by the harvesting team, in order to ensure success describe metadata harvesting to build a Web portal for library resource, and a number of issues arising from the heterogeneous nature of the collections developed by different communities. From that some of the 39 harvest partners were themselves aggregators and the team had to manage metadata from institutions. Inevitably, there was variation in the quality of the retrieved metadata involved in the harvest (archives, academic libraries, and digital libraries). There were difference in descriptions of the same resources by different institutions, and differences in the use of item- level and collection- level metadata, all of which required normalization work. In terms of the comment by on the experimental nature of OAI-PMH implementation, it is useful to ask whether the reported issues (as well as a number of similar examples collected anecdotally by the authors) in fact constitute a series of simulated outcomes of such experiments (c.f. If so, it is then also useful to see to what extent these replicated outcomes might be explained in terms of underlying characteristics of metadata and/or digital libraries.

The rest of this paper approaches this question by considering digital libraries as organization. The specific organizational approach adopted in this paper treats digital libraries components of sociotechnical systems of technologies, institutions, practices and other phenomena. From this perspective, digital libraries, metadata, and metadata work, all mutually shape and re-shape each other in response to internal and external organizational factors (similar observation have been with regard to the mutual co evolution of data curation practices and cyberinfrastructure. Over time, these mutual interactions can produce divergences in metadata format and usage between organizations, which can impede metadata harvesting

3. Remote Metadata

The most important issue is that the digital libraries collections sometimes encountered metadata not immediately retrievable through OAI-PMH queries. This 'unavailable' metadata was often discovered by coincidence, and required further work in order to locate, understand, and then (if possible) harvest it.

For example, at the time of the pilot harvest, IPL (Internet Public Library) metadata was stored in any database, in three separate datastreams: a DC stream, which held the 15 unqualified DC metadata fields; an IPL1.0 datastream, which held further metadata; and the RELS -EXT datastream, which held reports regarding item-collection relationship. While some of the metadata required for MASH (e.g. dc:title, dc:description, and dc:subject) was stored in the DC datastream, further potentially useful elements (such as ipl:subject) were stored in the IPL datastream, and were not originally planned to be part of the harvest. In another example, some metadata from an earlier version of the IPL was stored in a FileMaker Pro database, and as this had not been crosswalked to Dublin Core, it was not accessible for harvesting. Sympathetic the occurrence of this extra metadata required familiarity with IPL history before harvesting decisions could be made.

Finally, similar issues were faced in the metadata harvest with NSDL. The NSDL is a joined multi-disciplinary STEM library, with a central metadata repository at nsdl.org, which was provided as test-bed repository by the Mining into Data program funders. NSDL served as an early operation of OAI-PMH; and early issues get up from the fact that many NSDL projects lacked the organizational resources (skills, expertise, time, etc.) necessary for successful harvesting. In the current work, the initial batch of NSDL records obtained through OAI-PMH was found to contain different metadata to that displayed on the corresponding catalog record page in the NSDL Website. Email communication with the NSDL's metadata staff explained this position, which gets up from the fact that the metadata

aggregated for the NSDL central repository sometimes included records for the same resource that had been provided by multiple NSDL partner. In this case the Website showed a summarized and normalized version of these multiple records, although the original OAI-PMH query retrieved a concatenated version of the same multiple record. This clarification allowed for the alteration of the harvesting process, through a reconfigured set of OAI-PMH queries.

4. Conclusion

This paper has introduced the analysis of digital library metadata harvesting. The digital libraries and metadata are mutually constitutive occurrences that co-evolve over time. One consequence of this co-evolution is that even where a metadata standard (such as Dublin Core) is accepted, individual factors can produce different applications of that standard in different digital libraries. This paper also considers the metadata generation process and it is useful for the librarians for metadata harvesting.

References

1. Arms, W., Dushay, N., Fulker, D., Lagoze, C. (2003). A case study in metadata harvesting: The NSDL. Library Hi Tech 21(2), 228-237.
2. Bishop, A. P., Van House, N. A., Buttenfield, B. P. (Eds.). (2003). Digital Library Use. Social Practice in Design and Evaluation. Cambridge, MA: The MIT Press.
3. Chowdhury, G., McMenemy, D., Poulter, A. (2006). Large-Scale Impact of Digital Library Services: Findings from a Major Evaluation of SCRAN. In: J. Gonzalo et al. (Eds): ECDL 2006, LNCS 4172, pp. 256-266.
4. Electronic Libraries Program (n.d.). e-Lib: The Electronic Libraries Programme 1995-2001. http://www.ukoln.ac.uk/elib/
5. Gonçalves, M. A., Fox, E. A., Watson, L. T., Kipp, N. A. (2004). Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries. ACM Transactions on Information Systems, 22(2), 270-312.
6. Halbert, M., Kaczmarek, J., Hagedorn, K. (2003). Findings from the Mellon Metadata Harvesting Initiative. 7th European Conference, ECDL 2003 Trondheim, Norway, August 17-22, 2003. Pp. 58-69. DOI: 10.1007/978-3-540-45175-4_7
7. Hartswood, M., Procter, R., Taylor, P., Blot, L., Anderson, S., Rouncefield, M., Slack, M. (2012). Problems of Data Mobility and Reuse in the Provision of Computer-based Training for Screening Mammography. CHI ’12, Austin, TX, USA, 909-918.
8. Hiom, D. (2006a). Retrospective on the RDN. Ariadne Issue 47, http://www.ariadne.ac.uk/issue47/hiom/
9. Khoo M., Hall, C. (2013). Managing metadata: Networks of practice, technological frames, and technical work in a digital library. Information and Organization 23, 81–106.
10. Lagoze, C., Van de Sompel, H. (2001). The Open Archives Initiative: Building a Low-Barrier Interoperability Framework. First ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'01), 54–62.
11. Lynch, C. (2001). Metadata Harvesting and the Open Archives Initiative. ARL, A Bimonthly Report, no. 217. http://www.arl.org/resources/pubs/br/br217/br217mhp.shtml
12. NISO (National Information Standards Organization) (2004). Understanding Metadata. Bethesda, MD: NISO. Retrieved from: http://www.niso.org/publications/press/UnderstandingMetadata.pdf