12-30-2012

Open Archives Initiative Protocol for Metadata Harvesting, Dublin Core and Accessibility in the OAIster Repository

Michael Peake
San Jose State University, michaelpeake123@gmail.com

Open Archives Initiative Protocol for Metadata Harvesting, Dublin Core and accessibility in the OAIster repository.

Michael Peake

San Jose State University

Abstract

This paper examines the use of Dublin Core as a minimum metadata standard for Open Archives Initiative Protocol for Metadata Harvesting in terms of its impact on end user experience in the OAIster repository. Specifically the study looked at the use of controlled vocabulary searches versus non-controlled vocabulary searches, as well as the impact of Dublin Core on the granularity and consistency records. Searches were performed in OAIster using Library of Congress Subject Headings and Name Authority Files, as well as non-controlled vocabulary searches for the same terms. The study concluded that controlled vocabulary searches are good for retrieving relevant results, but non-controlled vocabulary searches can retrieve more relevant results, at the cost of large numbers of non-relevant results also being returned. The openness of Dublin Core does lead to problems of granularity and consistency, but some records indicate the potential of Dublin Core for providing very useful records. The study concludes that institutions should give serious consideration to user experience and repository display before converting records to Dublin Core for harvesting.

Open Archives Initiative Protocol for Metadata Harvesting, Dublin Core and accessibility in the OAIster repository.

Introduction

The rapid increase in information resources on the World Wide Web has created a situation where an unprecedented amount of knowledge is potentially available to anyone with internet access. Indeed Tony Gill argues that "the Web is the largest and fastest-growing collection of documents the world has ever seen" (2008, p. 25). The publically indexable web alone had at least 25.21 billion pages in 2009 ("World Wide Web", 2012, Statistics section, para. 1), and this does not include the many resources in databases that require search forms to be filled out before they are accessible (Raghavan & Garcia-Molina, n.d., abstract section). With so many information resources available the problem becomes one of successfully finding relevant resources among the billions available.

Metadata, or data about data, is key to retrieving relevant information because it structures data about information resources in ways that can provide meaningful access points for searchers. For example the use of Machine-Readable Cataloging (MARC) records in online public access catalogs (OPAC's) in the library field provides searchers with understandable access points, such as subject or author, and a controlled vocabulary that aids retrieval by bringing together resources on the same topic or by the same author. Although OPAC's generally allow keyword searching, the combination of access points and controlled vocabulary can greatly aid retrieval of relevant resources as they bring together like resources that may be described differently and thus not show up in a keyword search. Works by an author who writes under more than one name, for example Stephen King/Richard Bachman, or subjects that may be

named differently within the literature, for example the Spanish Civil War/Spanish Revolution can be retrieved together in this way.

Paradoxically metadata is both a key and a hindrance to finding relevant information. Different knowledge communities work with different metadata schemas and standards that are best suited to their purposes and priorities. For example the "archival community has embraced standards such as" Encoded Archival Description (EAD) and Describing Archives: A Content Standard (DACS) (Spiro, 2009, The Role of Software in Addressing Hidden Collections section), while the library community currently tends towards MARC and Anglo-American Cataloguing Rules 2nd edition (AACR2). With researchers interested in a variety of resources, for example books, archival material, web documents, and images, the issue becomes one of metadata interoperability. As Woodley points out, bringing material together from a single community in a union catalog, like WorldCat for the library community, is possible because "the contributing community shares the same rules for description and access and the same protocol for encoding the information" (2008, p. 46). Bringing together data encoded in different metadata schemas and formatted according to different content standards poses issues of interoperability. In other words records from within a single knowledge community may have a high level of interoperability, but "it is when communities want to share their content in a broader arena, or reuse the information for other purposes, that problems of interoperability arise" (Woodley, 2008, p. 39). There are many approaches to improving metadata interoperability. This paper will examine one of them, namely the role that the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), in conjunction with Dublin Core, plays in improving the accessibility of records held by diverse institutions. The OAIster database hosted by WorldCat will be considered in terms of its ability to make records accessible.

Statement of the Problem

There are a vast and continually growing number of information resources on the World Wide Web. Many of these resources remain hidden from commonly used search methods, such as Google or Yahoo search engines. By exposing their records to OAI compliant harvesters, institutions hope to make their records easier to access by allowing them to be retrieved from larger repositories, such as OAIster or Europeana. These repositories bring records together and allow users to search through the records of various institutions in one place. However, different institutions use different metadata schemas and data standards, so in order to ensure the interoperability of the records, OAI requires all records, at a minimum, to be exposed in simple Dublin Core. There is a concern that the use of simple Dublin Core may lead, in itself, to retrieval issues due to lack of granularity and crosswalking problems. The question arises as to whether the use of simple Dublin Core hampers the retrievability of records in OAI compliant repositories?

Literature Review

"The primary role of the OAI-PMH is to facilitate resource discovery when resources are stored in a number of distributed, independent repositories by exporting metadata about items in those repositories" (Fegen, 2007, What is the OAI Protocol for Metadata Harvesting for? section, para. 1). In doing this the OAI-PMH attempts to address the problem of metadata interoperability. According to the National Information Standards Organization (NISO), "interoperability is the ability of multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality" (2004, p. 4). Chan and Zeng argue that metadata interoperability is needed to

make it "possible to facilitate the exchange and sharing of data prepared according to different metadata schemas and to enable cross-collection searching" (2006a, Introduction section, para. 1). Haslhofer and Klas consider metadata interoperability "a perquisite for uniform access to media objects in *multiple* autonomous and heterogeneous information systems" (2010, p. 1). With Woodley's declaration that "global access to the universe of traditional print materials and digital resources has become more than ever the goal of many institutions" (2008, p. 38), the importance of metadata interoperability can easily be seen.

Chan and Zeng identify three different levels of approach to achieving Metadata interoperability, the schema level, the record level and the repository level. They note that individual projects may combine more than one approach (2006a, Metadata Interoperability Projects at Different Levels section, para. 5). The OAI-PMH approach is focused on both the repository level and the schema level. Repositories are identified by Woodley as a means of bringing metadata records together in a single database "with links from individual records back to their home environments" (2008, p. 47). Chan and Zeng differentiate between two types of repository, those that harvest records based on converted metadata, for example using the OAI-PMH, and those that harvest records without needing record conversion (2006b, Achieving Interoperability at the Repository Level section). This research paper only considers repositories that use the OAI-PMH.

Woodley describes repositories as a "recent model for union catalogs" (2008, p. 47) that physically bring together records "in a single database, with links from individual records back to their home environments" (2008, p. 47). According to the Open Archives Initiative website, two types of entity are involved in creating OAI-PMH compliant repositories, data providers and service providers (n.d., Interoperability through Metadata Exchange section, para. 1). "*Data*

Providers are repositories that expose structured metadata via OAI-PMH. *Service Providers* then make OAI-PMH service requests to harvest that metadata" (n.d., Interoperability through Metadata Exchange section, para. 1). The service providers add what Woodley describes as "an extra 'layer'" to the records from the data providers that "manages the mapping and searching of heterogeneous metadata records within a single aggregated resource" (2008, p. 49). Fegen notes that contrasted to federated searching across multiple databases, search within an OAI-PMH compliant repository has the "aim of providing the end user with an increase in responsiveness, reliability and possibly functionality" (2007, How OAI-PMH works section, para. 2). One of the main ways that OAI-PMH attempts to achieve this is by demanding that data providers adhere to a minimal standard at the schema level. Haslhofer and Klas, point out that "Most standardized metadata schemes are designed for a specific domain and a certain purpose" (2010, p. 20) so the use of one standard schema by all knowledge communities is highly unlikely. Furthermore, Chan and Zeng point out that "there are often two or more options for metadata standards" for the same subject-domain or resource type" (2006a, Introduction section, para. 1). In an attempt to maintain a low barrier to entry, OAI-PMH requires the use of unqualified Dublin Core at the schema level. Unqualified Dublin Core consists of only fifteen elements, none of which are required, although they are all repeatable (Hutt & Riley, n.d., Slide 3).

Tennant argues that the requirement to use unqualified Dublin Core creates problems, for example granularity may be lost and can be hard to recover (2004, Granularity section, para. 2). Tennant also argues that institutions sometimes create sets that have no meaning in a broader repository environment, for example naming sets after university departments (2004, Sets section, para. 1). Although Tennant provides suggestions to help remedy some of these issues, it is unclear from looking at the openarchives.org website whether any of these suggestions have

been acted on (Open Archives, 2008). Using unqualified DC as a minimum standard has the advantage of being a relatively simple metadata schema which provides a low barrier to entry, but it may end up creating poor search retrieval results, somewhat defeating the purpose of creating a metadata repository. The problem is that most data providers will have their original data in a format other than DC, meaning the data will have to be crosswalked.

Crosswalks, defined by NISO as "a mapping of the elements, semantics and syntax from one metadata scheme to those of another" (2004, p. 11), are identified in the literature as a crucial means of obtaining metadata interoperability. Chan and Zeng describe crosswalks as "by far the most commonly used method to enable interoperability" (2006a, Crosswalks section, para. 1) and Haslhofer and Klas that "schema mapping can deal with all kinds of heterogeneities on the schema level" (2010, p. 34). Woodley notes that "mapping metadata elements … is only one level of crosswalking" and that data content standards also need to be mapped (2008, p. 42). However, Haslhofer and Klas assert that in reality the issue of data content standards, which they refer to as "instance transformation", is often left out of crosswalks (2010, p. 27). Crosswalks lack of focus on the actual data in metadata records is a potentially serious issue, the more so because of the ubiquity of crosswalk use for metadata conversion. Woodley notes that differences in data value structure can make search results less successful (2008, p. 42). Chan argues that crosswalking works best when converting from a more complex to a less complex schema, for example MARC to Dublin Core (2005, Crosswalks/Mapping section, para. 3). Even this is not immune to problems as "data values may be lost when converting from a rich structure to a simpler structure" (Chan & Zeng, 2006b, Conversion of Metadata Records section, para. 5). Tennant describes mapping from a rich format to a simpler one like Dublin Core as "dumbing down" (2004, Simple DC is Too Simple section, para. 1). Issues can also arise with

equivalencies between schemas, such as many-to-one and one-to-many (Zeng, as cited in Chan, 2005, Crosswalks/Mapping section, para. 3). Haslhofer and Klas argue that what they define as "metadata mapping", which includes scheme mapping and instance transformation, has the potential to deal with these issues (2010, pp. 28-31, 34).

Unfortunately, simple Dublin Core does not address the issue of instance mapping and "leaves content rules to the particular implementation" (NISO, 2004, p. 3). Taylor argues that unqualified Dublin Core elements are themselves too vague to express anything meaningful other than author, title and date (Taylor, 2010, The Dublin Core, metadata - made dumb section para. 2). Tennant has found that even the date element is vague, due to the lack of authority control on how date terms are entered – he found twenty different ways of entering date information among only five data providers (2004, Encoding Variances section), showing a lack of control even at a local level. Tennant identifies part of the problem as being the lack of granularity available in unqualified Dublin Core, for example the need to place the" various constituent components of a personal name into one unstructured field" (2004, Granularity section, para. 2). On the other hand, Taylor argues that even Qualified Dublin Core suffers from a lack of granularity. He gives the example of the bibliographicCitation element that is intended to contain the information for a journal citation (journal title, volume, issue number and page range), but in an entirely uncontrolled format (2010, Even qualified Dublin Core can't describe a journal article section, para. 4). Placing these items in one element makes efficient search and retrieval harder, as does the lack of any specific citation format. Overall the literature does suggest that the use of unqualified Dublin Core may create problems with retrievability in OAI-PMH service providers.

The repository being looked at in this paper is OAIster. OAIster was started by the University of Michigan with funding provided by grants from the Andrew W. Mellon Foundation in 2002 (Online Computer Library Center (OCLC), 2012a, History of OAIster section, para. 1). In 2009 the OCLC partnered with the University of Michigan to provide continued access to OAIster which now hosts over 25 million open access records from 1,100 contributing organizations (OCLC, 2012a, History of OAIster section, para. 2).

Research Questions

The research in this paper intends to address the following questions:

- Does the use of simple Dublin Core inhibit searches in OAIster using Library of Congress Subject Headings (LCSH) and Library of Congress Name Authority Files (LCNAF)?
- Does the open nature of the data fields in simple Dublin Core create search and retrieval problems in OAIster due to inconsistent placement of data?
- Is there much data lost in OAIster records due to differences in granularity between Dublin Core and the original metadata schema used by the originating institution?

Methodology

Searches were performed using two subjects from the LCSH using the advanced search options screen of the FirstSearch database accessed through San Jose State University, King Library. Each subject was searched using the limitations subject, subject phrase and keyword. The number of hits was noted, followed by a count of the number of those hits that were relevant to the subject. Relevancy was judged on the basis of interest to a researcher studying the topic.

The same method was used for two personal names from the LCNAF, with the searches being performed with the limitations author, author phrase, named person, personal name, personal name phrase and keyword. After each controlled vocabulary search, searches were performed using non-controlled vocabulary looking for the same subjects/personal names. Several records chosen at random from those retrieved were looked at to determine the element location of relevant data, as well as granularity compared to the original records and any other observations of interest.

Results

Tables 1 through 4 show the results of searches performed in OAIster in terms of the number of records retrieved and the number of those records that were relevant to the search. A Relevant record was defined as a record that may be of interest to someone researching the topic in question.

Table 1 Spain--History--Civil War, 1936-1939

	Subject search number of records retrieved	Subject search number of relevant results	Subject Phrase search number of records retrieved	Subject Phrase search number of relevant results	Keyword search number of records retrieved	Keyword search number of relevant results
Spain--History--Civil War, 1936-1939 (LCSH)	77	77	29	29	77	77
Spanish Civil War	325	287	159	155	1493	988
Spanish Revolution	36	6	0	0	515	67
Spanish Revolution 1936-1939	3	3	0	0	8	8

Table 2 Great Britain--History--Civil War, 1642-1649

	Subject search number of records retrieved	Subject search number of relevant results	Subject Phrase search number of records retrieved	Subject Phrase search number of relevant results	Keyword search number of records retrieved	Keyword search number of relevant results
Great Britain--History--Civil War, 1642-1649 (LCSH)	569	569	355	355	569	569
English Civil War	60	46	28	28	1067	630

Table 3 Orwell, George, 1903-1950

	Author Search Retrieved /Relevant	Author Phrase Search Retrieved/Relevant	Named Person Search Retrieved/ Relevant	Personal Name Search Retrieved/Relevant	Personal Name Phrase Search Retrieved/ Relevant	Keyword Search Retrieved/Relevant
Orwell, George, 1903-1950 (LCNAF)	40/40	29/29	0/0	0/0	0/0	80/80
George Orwell	71/71	3/3	0/0	0/0	0/0	368/350
Eric Arthur Blair	19/15	0/0	0/0	0/0	0/0	33/26

Table 4 Cromwell, Oliver, 1599-1658

	Author Search Retrieved/Relevant	Author Phrase Search Retrieved/Relevant	Named Person Search Retrieved/ Relevant	Personal Name Search Retrieved/Relevant	Personal Name Phrase Search Retrieved/ Relevant	Keyword Search Retrieved/Relevant
Cromwell, Oliver, 1599-1658 (LCNAF)	57/57	57/57	0/0	0/0	0/0	185/185
Oliver Cromwell	64/	3/2	0/0	0/0	0/0	317/291

Discussion

The first thing that stands out about the search results is the 100% retrieval to relevancy rate obtained by all four controlled vocabulary searches. This retrieval rate was across every type of search, for example subject, author or keyword. There are two interesting things to note from these results. Firstly controlled vocabulary searches provide the best retrieval to relevance ratio. Secondly there is a difference in the number of records retrieved in some of the searches, depending on the search limiter used. For example Orwell, George, 1903-1950 retrieved 40 results with an author search, 80 with a keyword search and only 29 with an author phrase search. The author phrase search only retrieves results with the exact data "Orwell, George, 1903-1950" and misses records such as the Orwell papers from the AIM 25 Archives, because the author element contains the data "Blair | Eric Arthur | 1903-1950 | novelist and journalist known as George Orwell". Incidentally this record is retrieved by an author search for the term "George Orwell" or "Eric Arthur Blair". This may give the impression that an uncontrolled vocabulary search gives better results when looking for an author, but other issues can arise.

Sometimes author search can be misleading, for example searching either “George Orwell” or “Eric Arthur Blair” under author search will retrieve the record for “Orwell papers: Eileen Blair papers”. This happens because it is stated in the author element that Eileen Blair was the “first wife of Eric Arthur Blair (George Orwell)”. Another case appears with Mary Howgill’s letter to Oliver Cromwell where he is listed in the author element despite being the recipient of the correspondence not the writer. These seem to be an inappropriate use of the author element as it produces ambiguous search results. Presumably a person performing an author search is only looking for works authored by the person whose name they are searching. Additionally using uncontrolled vocabulary can retrieve unwanted records, for example an author phrase search for “Oliver Cromwell” retrieves three records, one of which is by a different Oliver Cromwell.

When searching for a personal name it was determined that the order in which the name is entered, for example “George Orwell” or “Orwell, George”, has no effect on the results retrieved. The placement of terms on the other hand is important. For example Oliver Cromwell appears in some records only in Author and Abstract elements, meaning a search for Oliver Cromwell as a subject will not retrieve these. It’s reasonable to suppose that such a search would hope to find these items. Named person, personal name and personal name phrase searches in all cases produced no records, so the merits of these search options are unclear. When documentation for searching in OAIster was looked at it was determined that there is no expert search option for these search types available, suggesting that they provide no useful function(OCLC, 2012b, Index labels and examples of an expert search in OAIster section) . If this is the case their presence in the search options is unnecessary and distracting.

Keyword searches tended to retrieve more records than other searches, such as author search, when looking for personal names. This was to be expected as works about, as well as by, a given person are likely to be available. On the other hand, when searching LCSH heading by subject search and keyword search the exact same records are retrieved. Searching for subjects using uncontrolled vocabulary produced disparate results. For example a subject search for "Spanish Civil War" produced 325 records, of which 287 were relevant. A keyword search for the same term retrieved 1493 records, of which 988 were relevant. These figures are far higher than the maximum 77 records retrieved using the appropriate LCSH term, "Spain--History--Civil War, 1936-1939". Does this mean searches using uncontrolled vocabulary are the best option in OAIster?

Several factors contributed to these search results. For example the search term "Spanish revolution 1936-1939" produced better results as keyword rather than subject because 5 out of 8 records contained English in the abstract section, but Spanish in the identifier section. The abstract is not searched in a subject search and so any records with the relevant information in the abstract only will not be retrieved by a subject search. The number of records retrieved is also misleading, because many records show up multiple times in the same result set. For example two records for "Spanish civil war refugees on a train" are retrieved with different OCLC accession numbers. Both are from university of San Antonio, Texas, one via Contentdm, the other from the university direct. In a worse example, five records for the same resource, "Remembering Franco: Spanish collective memory from the civil war till today" by Rebecca Beeson can be found in OAIster. As they contain the same data a search that retrieves one of these records will also retrieve the other four, which inflates retrieval rates and wastes researcher time.

In addition to the issues outlined above concerning retrieval, the following observations were made on the quality of the data available in OAIster. Some very poor title choices appeared among the records retrieved. Some suggested that they had been copied from an institutional title, despite the fact that in a repository like OAIster they would be meaningless. Examples included “Interview 181”, “page08” and “Book Review”. In the case of “Book Review” the abstract contained a list of the actual reviews, but these are only accessible through a keyword search. The worse title encountered consisted only of “:”. Some titles and other elements contained artifacts like presumably from HTML versions of the records.

Records were encountered with a wide range of granularity. Figure 1 below shows an example of an exceedingly sparse record. It should be noted that, other than a search for the very generic sounding title, there is no way given to aid in finding this record in the originating institutions database.

Availability: **Check the catalogs in your library.**

- Libraries worldwide that own item: 1

Title: **The Spanish civil war**

Language: French; French

SUBJECT(S)

Genre/Form: Text

Note(s): text/xml

Document Type: i

Entry: 20111005

Database: OAIster

Figure 1

Other records clearly showed the dumbing-down effect of crosswalking from MARC to Dublin Core. For example the record for "The Plundering Time: Maryland and the English Civil War, 1645-1646 (review)" can be found in keyword search for English Civil War because the title, featuring those words is used as an identifier along with the LCSH heading Maryland -- History -- Colonial period, ca. 1600-1775 (the item will also be retrieved because keyword also searches the title). Looking at the original record at John Hopkins University it can be seen that the original record is in MARC format. Maryland -- History -- Colonial period, ca. 1600-1775 appears as the first of three entries in the 651 field, the third entry being Great Britain -- History -- Civil War, 1642-1649 – Influence. The record has suffered sufficient loss of granularity to make it impossible to retrieve in OAIster using a search under this subject heading, even though its presence in the original record suggests its relevance.

The "Harry S. Holcomb papers" provide an example of crosswalking from EAD to Dublin Core. The record has a large amount of information in the abstract section, which is only retrievable through a keyword search. This in itself is not necessarily a problem, but there is very little information in the subject searchable area, which contains only the following:

Genre/Form: Correspondence; Certificates; Pamphlets; Scrapbooks

Identifier: undefined; UMAbroad—Politics, Government, and Law; UMAnarrow--Military

It is hard to tell from this information that these are the papers of a man who fought in the Philippines and with the American Expeditionary Force in World War One, information that can easily be garnered from the abstract.

The problem of inconsistent data standards was also evident in other elements of the records looked at. For example the values “text”, “text(article)”, “other”, “text, thesis” and “thesis” were found in the Genre/Form element showing varying degrees on granularity in the records. Within the description element there was little uniformity either. For example the following description values were found in the space of five records:

1 broadside

[2], 14p.

8p.

[8] p.

[6], 140 [i.e. 138] p.

Conclusion

In his 2004 article “Bitter harvest: Problems & suggested solutions for OAI-PMH data & service providers”, Tennant identified several areas of concern with OAI-PMH records, most of which he saw as stemming from the use of unqualified Dublin Core as a minimum schema for data providers to adhere to. Running searches through OAIster today reveals that many of these issues are as real today as they were eight years ago. The lack of granularity in many records is apparent. As the examples provided above show, this lack can be obvious or obscure to the searcher. No one will think the example in figure one is very granular, but on first look the record for “The Plundering Time: Maryland and the English Civil War, 1645-1646 (review)” looks reasonable, with an LCSH subject heading and so on. It is only when you go back to the originating institutions record that you realize what you are missing, a procedure a researcher is

unlikely to follow if they think the information available is present in the record they are looking at. Unqualified Dublin Core does seem "too simple" (Tennant, 2004, Simple DC is Too Simple section) and ambiguous, the lack of elements and clear direction leading to crosswalking issues, like the information in an EAD abstract only being searchable through a keyword search. The metadata artifacts and encoding variances that Tennant referred (2004, Metadata Artifacts section & Encoding Variances section) to are still very much in evidence, as is the problem of local naming procedures being used for records that will be looked at by a wider audience. "Interview 181" as a title may make sense within the originating institution, but it is worthless in the broader context of a repository like OAIster.

Clearly unqualified Dublin Core does create some issues, particularly due to its lack of granularity and data content standards. On the other hand much responsibility lies with the originating institutions. The act of making their records available for harvesting implies that they wish to make their records widely accessible and available. In order to achieve this goal they must pay attention to the way searches in OAIster work and consider the way they crosswalk their metadata appropriately. If the institution is relying on exposing Dublin Core records for harvesting, careful attention needs to be paid to mapping of elements to Dublin Core, making sure data that is suitable only for local use is changed to make it meaningful and that data value standards are adhered to as much as possible. Additionally it should be remembered that Dublin Core elements are repeatable and that supplying all data valuable for searching, such as multiple subject headings is retained. The existence of detailed and well laid-out records in OAIster show what is possible if an institution is willing to put the necessary effort and forethought into the creation of their records.

OAIster could focus on a way of removing duplicate records from search results. As it stands keyword searching is necessary to find all relevant records, but duplicate results and the high number of non-relevant records that are often retrieved do not make this the best search process from a user perspective. The results achieved through searches using LCSH and LCNAF showed the value of a controlled vocabulary, all results retrieved were relevant. Only the failure of some data providers to use controlled vocabulary prevents this from a best search solution. In the final analysis, using unqualified Dublin Core is useful in that it allows for the broadest range of participation in a repository like OAIster. However, those who choose to be data providers need to understand how the search mechanisms work and strive to provide the best data possible and comply with data content standards even though they are not mandatory. Those that do will find that their records are the most accessible and that should be motivation enough.

References

Andresen, L. (2009). Europeana [PowerPoint slides]. Retrieved from http://*files.itslearning.com/data/826/open/CO15/667.**ppt***

Chan, L. M. (2005). Metadata interoperability: A study of methodology. Retrieved from http://www.white-clouds.com/iclc/cliej/cl19chan.htm

Chan, L. M., & Zeng, M. L. (2006a). Metadata interoperability and standardization – a study of methodology part I. *D-Lib Magazine, 12*(6). Retrieved from http://www.dlib.org/dlib/june06/chan/06chan.html

Chan, L. M., & Zeng, M. L. (2006b). Metadata interoperability and standardization – a study of methodology part II. *D-Lib Magazine, 12*(6). Retrieved from http://www.dlib.org/dlib/june06/zeng/06zeng.html

Fegen, N. (2007). What is the OAI Protocol for Metadata Harvesting. In *JISC cetis*. Retrieved from http://wiki.cetis.ac.uk/What_is_the_OAI_Protocol_for_Metadata_Harvesting

Gill, T. (2008). Metadata and the web. In M. Baca (Ed.), Metadata (pp. 20-37). Los Angeles, CA: Getty Research institute.

Haslhofer, B. & Klas, W. (2010). A survey of techniques for achieving metadata interoperability. *Computing Surveys (CSUR) 42*(2). Retrieved from http://eprints.cs.univie.ac.at/79/1/haslhofer08_acmSur_final.pdf

Hutt, A., & Riley, J. (n.d.). Semantics and syntax of Dublin Core usage in Open Archives Initiative data providers of cultural heritage materials [PDF document]. Retrieved from http://www.lib.unc.edu/users/jlriley/presentations/jcdl2005/jcdl2005.pdf

National Information Standards Organization. (2004). Understanding metadata. Retrieved from http://www.niso.org/publications/press/UnderstandingMetadata.pdf

Online Computer Library Center. (2012a). The OAIster database at a glance. Retrieved from http://www.oclc.org/oaister/about/default.htm

Online Computer Library Center. (2012b). OAIster. Retrieved from http://www.oclc.org/support/documentation/firstsearch/databases/dbdetails/details/OAIster.htm

Open Archives Initiative. (2008). Implementation Guidelines. In *The Open Archives Initiative Protocol for Metadata Harvesting*. Retrieved from http://www.openarchives.org/OAI/openarchivesprotocol.html#ImpGuid

Open Archives Initiative. (n.d.). The Open Archives Initiative Protocol for Metadata Harvesting. Retrieved from http://www.openarchives.org/pmh/

Raghavan, S., & Garcia-Molina, H. (n.d.). Crawling the hidden web. [Abstract] Retrieved from http://www10.org/cdrom/posters/p1049/index.htm

Spiro, L. (2009). Introduction. In *Archival management software: A report for the Council on Library and Information Resources*. Retrieved from http://www.clir.org/pubs/reports/spiro/report.html

Taylor, M. (2010). Bibliographic data, part 2: Dublin Core's dirty little secret. In *The Reinvigorated Programmer: Everything except sauropod vertebrae.* Retrieved from http://reprog.wordpress.com/2010/09/03/bibliographic-data-part-2-dublin-cores-dirty-little-secret/

Tennant, R. (2004). Bitter harvest: Problems & suggested solutions for OAI-PMH data & service providers. Retrieved from http://roytennant.com/bitter_harvest.html

Woodley, M. S. (2008). Crosswalks, metadata harvesting, federated searching, metasearching: Using metadata to connect users and information. In M. Baca (Ed.), *Metadata* (pp. 38-62). Los Angeles, CA: Getty Research institute.

World Wide Web. (2012). In *Wikipedia, the free encyclopedia*. Retrieved from http://en.wikipedia.org/wiki/World_Wide_Web