

Strategies for reprocessing aggregated metadata

Muriel Foulonneau¹, Timothy W. Cole¹

¹ Grainger Library, University of Illinois at Urbana-Champaign
1301 W. Springfield Avenue
Urbana, IL 61801
+1 - 217 - 244 - 7809
{mfoulonn, t-cole3}@uiuc.edu

Abstract. The OAI protocol facilitates the aggregation of large numbers of heterogeneous metadata records. In order to make harvested records useable in the context of an OAI service provider, the records typically must be filtered, analyzed and transformed. The CIC metadata portal harvests 450,000 records from 18 repositories at 9 U.S. Midwestern universities. The process implemented for transforming metadata records for this project supports multiple workflows and end-user interfaces. The design of the metadata transformation process required trade-offs between aggregation homogeneity and utility for purpose and pragmatic constraints such as feasibility, human resources, and processing time.

1 Aggregating Metadata Describing Scholarly Resources

In recent years, large aggregations of metadata describing heterogeneous resources have been created using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). OAI service providers who build applications on top of such aggregations must amalgamate large amounts of metadata harvested in a range of formats. By reprocessing harvested metadata, service providers can adapt metadata for their specific use and present those metadata to end users in an integrated fashion.

The process of adapting metadata for another application than originally envisioned when the metadata records were created, i.e., repurposing metadata, requires analyzing the metadata harvested, identifying processes to apply to the metadata, and then building the reprocessing system to select, transform and organize the metadata. The present paper discusses issues related to metadata analysis and the implementation of a metadata reprocessing system. It suggests a range of strategies for metadata reprocessing and adaptation and identifies issues needing further study.

1.1 The CIC Portal: an Aggregation of 450,000 Metadata Records

The CIC metadata portal is a major metadata aggregation encompassing digital resources from 9 Midwestern universities in the U.S., mostly holdings of academic research libraries. It provides access to 450,000 descriptive metadata records from

152 defined collections. The CIC metadata portal has three main interfaces. A primary search and retrieval interface provides classic digital library access points for scholars: author, title, subject, type, and a number of filtering and grouping functionalities based on dates and collections,. A clickable geographic map allows users to browse by spatial coverage attributes of the resources indexed. Finally a second search and retrieval interface is provided that takes advantage of both collection-level and item-level descriptions in concert[3]. The interfaces were designed to improve information retrieval in aggregated collections, improve usability of heterogeneous information, and demonstrate the wealth of U.S. Midwestern digital library resources. To enable these three interfaces, metadata is reprocessed and then ingested by two distinct systems: the DLXS software developed by the University of Michigan on top of the OpenText XPat search engine; and a Microsoft SQL database. We describe below the implementation of our initial metadata reprocessing system, detailing workflow from harvesting to data publishing.

1.2 Metadata Reprocessing: Challenges and Objectives

"A metadata record is created in the objective of a specific use." [2] The challenge of repurposing metadata records is to reuse records created to fit one context in a different context with different constraints and objectives. Any attempt to reuse descriptive metadata runs the risk of misusing (misunderstanding) those records. The problem is exacerbated since even assuming that metadata records are generally well-adapted to the original context for which they were created, this original context typically remains partially or totally unknown to service providers. Moreover, OAI-PMH is designed to facilitate the harvesting of the same metadata records by multiple service providers, each likely to have their own unique purpose and context. However, since metadata records are expensive and resource consuming to generate, their reusability has great potential for benefit and is at the core of a number of current initiatives, notably the National Science Digital Library and the Digital Library Federation working group on best practices for OAI and shareable metadata.

These considerations dictate a thorough analysis of harvested metadata evaluated in terms of the service provider's context and typically applying different measures of metadata quality and utility [9] than were applied when the metadata records were originally created. Metadata records adequate in original local context may not be adequate in an aggregated context. In reprocessing harvested metadata, the service provider's challenge is to implement strategies to transform harvested metadata in a way that enhances usefulness, while avoiding misuse or misunderstanding.

1.3 Technical Implementation of Metadata Reprocessing

For the CIC portal, harvest of the OAI metadata provider repositories (18 in all) is customized according to a number of configurable parameters, including the strictness of XML validation, the specific sets to harvest, and the metadata format to harvest.

Once harvested, records are first processed through a program that selects relevant records (for the purposes of the aggregation). Selected records are then sent through

an XSL pipeline (chain of XSLT files) which implements a series of transformations. Each pipeline is customized by repository and composed of a number of XSLT stylesheets (five to eight) that are named in a configuration file. Only two XSLT stylesheets per repository are repository-specific. Repository specific stylesheets mostly implement a subset of element-specific normalization and augmentation functions. To facilitate normalization, the stylesheets import a generic dictionary of XSLT templates and a number of data dictionaries encoded in XML: e.g., ISO639 language codes, Dublin Core Metadata Initiative and CIC type vocabularies, a date data dictionary, ISO3166 and ISO3166-2US geospatial codes, and a subset of Internet MIME Type (IMT) format string values.

The selected, normalized, and enriched metadata records are stored in a distinct location separate from where the as-harvested metadata records are maintained. Periodically a procedure is run to upload the enriched metadata records into the Microsoft SQL database. Another program applies an additional stylesheet and concatenates transformed metadata records as required for use in the DLXS-based service mentioned above. The concatenated files are then transferred to another server where a shell script routine rebuilds DLXS indexes. Specifics of the metadata record filtering, normalization, and enrichment tasks implemented for the CIC portal are described below in sections 3 & 4.

2 Metadata Analysis

Each new collection to be added to the aggregation is analyzed to define specific reprocessing needed. We also identify new mappings or transformation keys (e.g., to convert coded values to human-readable strings) for the data dictionaries used in reprocessing. We look both at a statistical characterization of the entire population of records and at the range of value strings encountered considered by metadata field.

2.1 Statistical Analysis of the Records Population

Analysis of the whole body of records harvested from a given provider is difficult to do manually due to the large numbers of records typically provided. A statistical analysis is therefore essential to obtain key indicators on the record population.

Earlier work here at the University of Illinois led to the development and refinement of a methodology for analyzing the use of DC metadata elements. [4, 7, 10] We have extended this methodology and applied it successfully to simple and qualified DC records of the CIC aggregation, and more recently to alternative formats offered by CIC metadata providers (e.g., ETDMS and MODS). Metadata records were analyzed either on a repository by repository basis, or where necessary on a collection by collection basis.

The metadata records are entered into a database and the summary information listed in Table 1 about the population of metadata records is generated. This analysis of the metadata population allows a service provider to identify reprocessing needs and anticipate impact new collection might have on the overall metadata aggregation.

Table 1. Statistics generated for metadata populations

General metadata population
The size of collection or set (i.e., number of records)
<hr/>
Structure of the records
The average number of metadata elements per record
The list and frequency of attributes used to specify encoding schemes or any other information
The use of the <i>About</i> section of the record to specify rights statement, provenance, or any other relevant information about the metadata
<hr/>
Elements used
The list and frequency of metadata elements of the specified metadata format that are actually used
Whether there is present in each record exactly one URL understood to link (directly or indirectly) to the individual resource described
Whether there is present in each record at least one of the fields displayed in results lists (i.e., title, subject, or description)
Whether all the fields used as browse or limit categories for portal interfaces are present in all records (eg. <i>Type</i>)
<hr/>
Value length
The number and percentage of empty metadata elements

2.2 Analyzing Metadata Values

All distinct text values are then extracted from the database for each metadata field present in the record set being analyzed. Also extracted are counts of how often each distinct text value occurs for the field considered. Distinct text values for each field are ordered by frequency of occurrence. This provides an indication of set consistency and the use of controlled vocabularies. Generally the characteristics of metadata elements are quickly identified using this approach, even in large metadata populations. The frequency of recurring values is also easy to identify since each distinct value is displayed with its number of occurrences in the collection.

This analysis of metadata values on a field-by-field basis also makes it easier to identify the location in the records of the concepts used in the CIC metadata portal, either as access points or for display purpose or as categories. For instance, the CIC metadata portal makes special use of resource type, format, language, URL, and rights information when present. Analysis facilitates identification of parameter values needed for XSL templates (e.g., what character to use for splitting concatenated terms) and any needed modifications or extensions of the data dictionaries used during reprocessing. When new unrecognized values appear (e.g., “technical reports” as a *Type*), they can be added. However, in several cases, the meaning may not apply

across all collections (e.g., The *Type* “other” means *software* in one collection, *text* in another). The data dictionaries are built to accommodate such variability.

For several fields, it is interesting to identify redundant values found in multiple records – e.g., in *Identifier*, in *Description*, and in *Title* – as the presence of such redundancy may indicate the presence of duplicate or overlapping metadata records. For other fields, redundancy is a good sign; it can suggest consistency and imply that the records in the set being analyzed will be easy to normalize. In one collection, duplicate records, having different URLs pointing to the same resource could be identified because their description and title fields were similar. It should be possible in the future to automatically identify records where the only difference is the page number of an online book. Page by page granularity does not make sense for the CIC metadata portal. Such records, while not duplicates in original local context, are essentially duplicative in the context of our CIC service provider implementation.

Improperly used (in the context of the service provider portal) and imprecise concepts should be normalized or renamed (i.e., put in a different field). When processing simple DC records, it may be necessary during processing to separate IMT format values from extent information. This is a qualified DC distinction. Spatial and temporal coverage data, which appear in the same element in simple DC records, are also considered for further processing. Information contained in records is assessed according to its function in the CIC portal. For instance, special effort is made to identify a single main URL per resource. We also verify the presence of elements needed for display in search results lists. Finally encoding consistency for any fields commonly retrieved across the whole metadata collection is assessed. This is an issue for *Author/Creator* values which may be encoded differently in different collections.

2.3 Analyzing Language Used in Metadata Fields

Metadata records harvested for the CIC aggregation include both "guide" metadata, expressed in natural language and intended for human consumption, and "control" metadata, intended for use in context of a database or other computer system application. [1] For the purposes of a search and discovery service, the fields used as access points should be clearly identified and controlled vocabularies used should be clearly labeled in all metadata records. This also implies that the terms by which the records are queried are the same as the ones used in the metadata records or at least retrievable through additional language processing functions commonly used in search engines (see for example [6]). Consider the following examples of Dublin Core *Description* field content: “First ed. Cf. BM.”, “D_North_American_1983_HARN”, “Added t.-p., engr.”, "Co. C". These strings are not really useful for direct retrieval since most users do not tend to query using abbreviations. Neither do they tend to query by codes. In order to search on such content, codes and abbreviations must be expanded to more human-readable forms. While this can be difficult for abbreviations or codes that are idiosyncratic to a particular local context, the service provider can expand recognized standard abbreviations and codes contained in harvested metadata.

OAI service providers also must display retrieved metadata records for end-users. Results listings of metadata records allow an end user to select resources of interest after gauging their likely usefulness for meeting his or her information need. All

metadata fields are not equally efficient for this purpose. It depends on the information they contain and the form of language used in the values (e.g., natural language versus codes). Each metadata field that will be displayed for the end user should be human interpretable. For example, “wln” is a code, created for machines, it is not human interpretable if the user does not know the ISO639 codes for languages. Some values may be both human and machine amenable. The value, “Text.Correspondence.Letter” in a Dublin Core *Type* field is easily understood by an end-user, although its stilted form suggests it is also intended to be machine parseable. However, even in such cases, it may be desirable to explicitly adapt controlled encodings to more human-readable forms. Thus, while the value, “197-“ is likely to be understood by most users, the value “created between 1970 and 1979” decreases the risk of end-user misunderstanding.

While the transformation of metadata field value strings to more human-readable form is generally desirable, a service provider displaying a value that was not originally created by the data provider risks betraying the original record, either intellectually or legally. This issue can be of major concern. In the museum community, a metadata record that describes an original object may be the result of an in-depth scientific analysis. Altering such a record before presenting it to end-users might risk providing less precise, duplicate, or less accurate information. This can have negative consequences for both service and content provider. The natural inclination of the service provider should be to maintain the original record values, unless he can safely and with a high degree of certainty decode an encoded value. Metadata normalization that focuses on adding machine readable values, expanding standard codes and abbreviations, and adding explicit labels for such information as type, language, URL, collection and rights, as we have done, is a fairly safe first step.

3 Metadata Reprocessing for the CIC Metadata Portal

After analyzing harvested metadata, we take the following steps to reprocess records for use in the CIC metadata aggregation.

3.1 Records Selection

The primary criteria in the collection development policy for the CIC metadata portal¹ are that all resources described by metadata in the aggregation should originate from a CIC institution and that the metadata records should be descriptive. Some repositories harvested include metadata describing both CIC and non-CIC resources. For those repositories the CIC portal collection development policy is implemented in the selection process using scripts to identify relevant metadata records. De-duplication of records, however, is not currently implemented at initial record filtering.

¹ CIC collection development policy <<http://cicharvest.grainger.uiuc.edu/collection.asp>>

3.2 Metadata Cleaning

This process consists of deleting erroneous characters at the beginning and end of strings, most commonly extra occurrences of characters used to delineate concatenated values (typically, certain data providers concatenate multiple values in a single metadata element, e.g., <description>16 History; 17 Geography;</description>, which can result in an unnecessary semi-colon at the end of the value), removing empty metadata (in the first analysis of the raw data collected that was done at the beginning of the project, 17.5% of the records contained at least one empty metadata element such as <description />), removing meaningless values (<date>--</date>), and splitting data when a metadata element contains multiple values (such as in the *History* and *Geography* example mentioned above).

3.3 Metadata Normalization

This step consists in disambiguating concepts and metadata semantics across the collections. It maps the elements used by data providers to a number of normalized fields of a specific metadata format used internally in the CIC portal (and derived from qualified DC). The *Format* field is renamed as *Extent* in the following case: “<format>163 pages</format>”. Values also are reprocessed at this stage for machine interpretability. This is the case for the resource type which is mapped to a controlled vocabulary used in a drop-down list on the search interface and for a filtering option in the search results. The format and the language are also normalized. One URL suitable for linking to each resource also is identified at this stage if possible.

3.4 Metadata Augmentation

Metadata augmentation consists in adding information to the records from external sources. While normalizing *Type* information, the values are translated through the DCMI-Type standard in addition to the local terminology used for the CIC portal. For several collections, a default *Type* is applied. A collection name is also added to the record to help provide additional context for the end user for when the record might be displayed in a results list. A provenance element is created in order to trace the record source.

3.5 Customizing Records for Use in Portal Interfaces

Even after filtering, cleaning, normalization, and enrichment, records may need to be modified further for use in a specific interface. The records to be ingested by the DLXS application are transformed from our internal, qualified DC-based metadata format to the Bibclass record format used by DLXS. Similar elements also are concatenated so that when the record is displayed *Subject* fields, for instance, are not

displayed on multiple lines. For our SQL-based interface, searchable metadata elements are concatenated into an additional field to facilitate use of built-in Microsoft SQL Server full-text search functions and features.

3.6 Performance Issues

Ingestion of new collections has been streamlined. As an example, once initial data analysis was performed for a new collection ingested in February 2005, the actual adaptation of reprocessing filters took less than one hour since there was no specific new processing to write (not always the case). However, any new metadata format to be ingested by the system can require significant new adaptation of the XSLT chain.

The execution of metadata reprocessing should not take too long, otherwise data would not be updated on the portal in a timely manner and the service credibility would be consequently jeopardized. Selection, clean up, normalization and augmentation can currently be performed on the full aggregation within 30 hours using parallel batch programs processing at a speed of about 2 records per second for each process. Processing collection by collection entails additional ongoing maintenance as noted by D. Hillmann and N. Dushay of Cornell University. [5] Metadata provider practices can change over time, meaning that the original, precise analysis of a certain data provider practice does not remain valid forever. This may happen with a new version of a turnkey system which might considerably change the nature of records exported when upgrading to the new version. In a specific repository harvested to build the CIC metadata portal, collection to record associations were originally recognized automatically through processing the URL of each resource as recorded in each item-level metadata record. A number of rules had been defined to identify a collection code in resource URLs. One day, the CIC portal appeared to have lost several of its collections, this altered considerably the value of some of the services offered. The data provider had changed the form of its URLs. Collections to item associations were no longer being recognized for this repository.

4. Specific Metadata Properties Featured in CIC Portal Interfaces

The overall metadata repository is composed of 445210 records (as of February 2005). The metadata reprocessing described above facilitates automatic recognition of specific concepts (properties) present in harvested metadata records, notably the resource URL, associated collection, format, type, and language. Normalization insures encoding of properties when present in a standard manner to facilitate use in portal interfaces. These properties are used in portal interfaces either as search access points, for filtering results, or for display of search results. Table 2 shows pre-processing count of metadata records including at least one occurrence of a field potentially containing a property of interest. (For brevity DC field names are shown -- records in Qualified Dublin Core were included using appropriate field names) Presence of a field does not guarantee property of interest is present and recognizable. An *Identifier* field may contain something other than a resource URL. The value of a

Type field may not correspond to the type vocabulary used in portal interfaces. For *Type* and *Format* we include in parenthesis a count of how many as-harvested records contained a string value from controlled vocabularies used for these two concepts in CIC portal interfaces.

Table 2. Presence of information in the original records

Property	Field potentially containing the property	# of records having field before processing	% of records in the repository
Type	Type field	344816 (297756)	77%
Format	Format field	319157 (42501)	72%
Language	Language field	268994	60%
Collection	Relation field	167990	38%
Resource URL	Identifier field	430848	97%

4.1 Acceptable Threshold of Normalization for Implementing a Search Interface

Not all properties used for search access or result filtering in portal interfaces apply to full range of resource descriptions included in the metadata aggregation. Language does not apply to images and URLs are not available for analog-only resources. The share of the records in the aggregation for which a concept can be identified and consistently included as a property in the augmented record will often be less than 100%. This may or may not be significant according to the nature of the property, how the property is exploited in the interface, and why it is not present in all records. For instance, 23% of the as-harvested records do not contain a *Type* field. Since *Type* is used as a way to filter results this means that 23% of the records in the aggregation are excluded when an end user chooses to limit search by *Type*. This is a serious problem. On the other hand, if a user limits his or her search by *Language*, we can reasonably assume he or she is not interested in still images. Excluding records that have no language property because they describe pictures is acceptable.

4.2 Applying Default Values

While record-by-record reprocessing facilitates the recognition of properties used as search access points, for filtering results, and/or for display of results, this technique alone may not insure presence of an essential property in a sufficient percentage of records in the aggregation. You cannot normalize the *Type* property value of a record if no *Type* field is present. Often explicit information is left out of item-level metadata records because in the local context for which the records were created this information was understood implicitly. In order to insure the presence of a property in all or almost all records, default property values are sometimes applied to (i.e., added to) all records in a given collection if the appropriate property value can be inferred with confidence from collection-level information.

The identification of the collection a record belongs to is typically not based on metadata values. A *Relation* field only appears in 38% of the as-harvested records. Even assuming this field is generally being used to express collection association, collection information would be present in at most 38% of the records in the aggregation. This percentage is clearly insufficient. Collection associations are determined primarily by harvest repository and/or OAI set membership.

Assignment of default values by collections is also used to identify restricted access resources. This method may be extended in the future, through the use of collection-level descriptions and property inheritability in order to complete the information contained in the item-level metadata record (information completeness) and to increase confidence in information added (information accuracy). Table 3 shows final post-processing counts of records containing *recognized* property values for the concepts listed in Table 2.

Table 3. Presence of information in records after processing

Property	# of records having property after processing	% of records in the repository
Type	441788	99%
Format	295803	66%
Language	268989	60%
Collection	445209	100%
Resource URL	320005	72%

4.3 Trade-Off between Accuracy and Share of Aggregation

Not all automated interpretations of concepts and meanings are precise and exact. In some cases, the nature of the resource (digital / analog) is not absolutely certain. After processing, a slight difference (3,322 records) exists between the number of resources without a resource URL in the augmented records and the number of resources to which a *Type* “analog resource” is applied. Even though those records do not contain a resource URL, they are still displayed when filtering “digital resources only.” The probability of inaccuracy can be increased by the application of default values. For example a service provider might apply a default *Type* of “analog resource” to all members of a specific collection (where *Type* was not given in metadata records provided) based on initial inspection of a sample of the collection. If in fact the collection contains a few online resources mixed in with mostly analog resources these resources will be mislabeled as analog only.

The content of the language property of the original record is accurately recognized in almost 100% of cases -- only 5 records contain unrecognized language values (much less than 1%). Unrecognized values either come from values impossible to identify such as “other” or from failure in the data analysis to correctly define the condition for splitting elements (a slash in the following case: "English/Japanese"). The character used as a separator to split metadata elements is identified when analyzing the records in a new collection. It is possible that when the collection was

analyzed, the case did not occur. Records being dynamic (updated from uncontrolled sources), new phenomena will appear from time to time.

5 The Future: Targeting Transformations for User Benefits

Metadata reprocessing can be a resource-intensive (i.e., expensive) process for a service provider. There also are potential hazards to the fidelity and integrity of the harvested metadata. The processing performed for the CIC metadata portal may improve the level of completeness of information, but it represents a risk of altering several features of the metadata, notably by making it less accurate. Further work is needed to measure and quantify the magnitude of this risk, especially when the metadata processing includes not only direct interpretation and normalization of string values, but also adds default properties to records based on association with a collection or OAI set or repository. The same metadata reprocessing workflow will be applied in the future to more complex data, such as dates. Complementary methodologies might also be used better adapt human-readable and machine readable data to the different functions of the portal.

Further work is also needed to quantify the impact of metadata quality on service to end-users. This should include, a comparison of user queries found in transaction logs to the content of specific metadata fields and an assessment of the impact of uncertain metadata created during service provider metadata reprocessing on recall and precision of information retrieval. User expectations and tolerance to inaccuracy of information displayed or retrieved (or not retrieved) may be different depending on whether the service is provided by libraries or in context of Web search engines [8].

Guidelines and best practices² for metadata creation and transformation by data providers contribute to improve the efficiency and accuracy of normalized information. However, service providers usually cannot impose common rules to all data providers and data providers cannot apply a single rule that is valid for all service providers. Ultimately a better understanding is needed of what metadata cleanup, normalization, and enrichment can reasonably and safely be done by service providers (i.e., harvesting agents) versus what processing really should be done by metadata providers prior to making records available for harvesting. While metadata records will always be authored with an immediate and specific local implementation in mind, in an environment that increasingly encourages sharing, reuse, and repurposing of digital metadata and content, authoring “shareable” metadata will benefit development of more robust and full-featured Web services.

² E.g., DLF/NSDL best practices for OAI and shareable metadata <<http://oai-best.commsdl.org/cgi-bin/wiki.pl>> and CIC-OAI project recommendations for Dublin Core metadata providers <<http://cicarvest.grainger.uiuc.edu/dcguidelines.asp>>

Acknowledgements

This work was supported by a grant from the Committee of Institutional Cooperation's Center for Library Initiatives. We acknowledge the libraries of the following participating CIC member institutions for providing metadata and collection descriptions: University of Chicago, University of Illinois at Chicago, University of Illinois at Urbana-Champaign, Indiana University, University of Iowa, University of Michigan, Michigan State University, Northwestern University, Pennsylvania State University, and the University of Wisconsin-Madison.

References

1. Bretherton, F.P., Singley P.T.: Metadata: a User's View. In: Proceedings of the Seventh International Working Conference on Scientific and Statistical Database Management. IEEE Computer Society, Washington, DC (1994) 166-174.
2. Coyle, K.: Data with a purpose. Talk at the California Library Association Meeting, November 2004. <http://www.kcoyle.net/meta_purpose.html>
3. Foulonneau, M., Cole, T.W., Habing, T.G., Shreeves, S.L.: Using collection descriptions to enhance an aggregation of harvested item-level metadata. In: Proceedings of the Fifth ACM / IEEE-CS Joint Conference on Digital Libraries, Denver, CO (2005, in press).
4. Halbert M., Kaczmarek J., Hagedorn K.: Findings from the Mellon Metadata Harvesting Initiative. In Koch, T., Sølberg, I.T. (eds.), Research and Advanced Technology for Digital Libraries 7th European Conference, ECDL 2003, Trondheim, Norway. Proceedings, Lecture Notes in Computer Science vol. 2769, Springer-Verlag GmbH, (2004) 58-69.
5. Hillmann, D.I., Dushay, N., Phipps, J.: Improving Metadata Quality: Augmentation and Recombination. In: DC-2004: Metadata Across Languages and Cultures, Shanghai, China. <http://purl.oclc.org/metadatasearch/dccconf2004/papers/Paper_21.pdf>
6. Ross, S., Donnelly, M., Dobreva, M., Abbott, D., McHugh, A., Rusbridge, A.: Natural Language processing. In: Core technologies for the cultural and scientific heritage sector. Digicult technology watch report 3 (2005) 67-103. <<http://www.digicult.info/downloads/TWR3-lowres.pdf>>
7. Shreeves, S.L., Kaczmarek, J., Cole, T.W.: Harvesting cultural heritage metadata using the OAI protocol. Library Hi-Tech. 21 (2003) 159-169.
8. Shreeves, S.L., Kirkham, C.M.: Experiences of educators using a portal of aggregated metadata. Journal of Digital Information 5 (2004) Article No. 290, 2004-09-09. <<http://jodi.ecs.soton.ac.uk/Articles/v05/i03/Shreeves/>>
9. Shreeves, S.L., Knutson, E.M., Stvilia, B., Palmer, C.L., Twidale, M.B., Cole, T.W.: Is quality metadata shareable metadata? The implications of local metadata practice on federated collections. In Thompson, H.A. (ed.): Proceedings of the Twelfth National Conference of the Association of College and Research Libraries, Minneapolis, MN. Association of College and Research Libraries, Chicago, IL (2005, in press).
10. Stvilia, B., Gasser, L., Twidale, M., Shreeves, S.L., Cole, T.W.: Metadata quality for federated collections. In Proceedings of ICIQ04 - 9th International Conference on Information Quality. Cambridge, MA (2004) 111-125.