# Is "Quality" Metadata "Shareable" Metadata? The Implications of Local Metadata Practices for Federated Collections

*Sarah L. Shreeves, Ellen M. Knutson, Besiki Stvilia, Carole L. Palmer, Michael B. Twidale, and Timothy W. Cole*

## Introduction

The federation of digital resources has become increasingly important in realizing the full potential of digital libraries. Federation is often achieved through the aggregation of descriptive metadata, therefore the decisions resource developers make for the creation, maintenance, and quality assurance of their metadata can have significant impacts on aggregators and service providers. Metadata may be of high quality within a local database or web site, but when it is taken out of this context, information may be lost or its integrity may be compromised. Maintaining consistency and fitness for purpose are also complicated when metadata are combined in a federated environment. A fuller understanding of the criteria for high quality, "shareable" metadata is crucial to the next step in the development of federated digital libraries.

This study of metadata quality was conducted by the IMLS Digital Collections and Content (DCC) project team (http://imlsdcc.grainger.uiuc.edu/) using quantitative and qualitative analysis of metadata authoring practices of several projects funded through the Institute of Museum and Library Services (IMLS) National Leadership Grant (NLG) program. We present a number of statistical characterizations of metadata samples drawn from a large corpus harvested through the Open Archives Initiative (OAI) Protocol for Metadata Harvesting (PMH) and interpret these findings in relation to general quality dimensions and metadata practices that occur at the local level. We discuss the impact of these kinds of quality on aggregation and suggest quality control and normalization processes that may improve search and discovery services at the aggregated level.

## Framework for Analyzing Metadata Quality

In general, quality problems arise when the existing quality is lower than the required quality in the context of a given activity (Gertsbakh 1977). Strong (1997) defines data quality problems as "difficulty encountered

*Sarah L. Shreeves (sshreeve@uiuc.edu) is a librarian; Ellen M. Knutson (eknutson@uiuc.edu) and Besiki Stvilia (stvilia@uiuc.edu) are students at the Graduate School of Library and Information Science; Carole L. Palmer (clpalmer@uiuc.edu), and Michael B. Twidale (twidale@uiuc.edu) are associate professors; Timothy W. Cole (t-cole3@uiuc.edu) is a math librarian, all at the University of Illinois at Urbana-Champaign.*

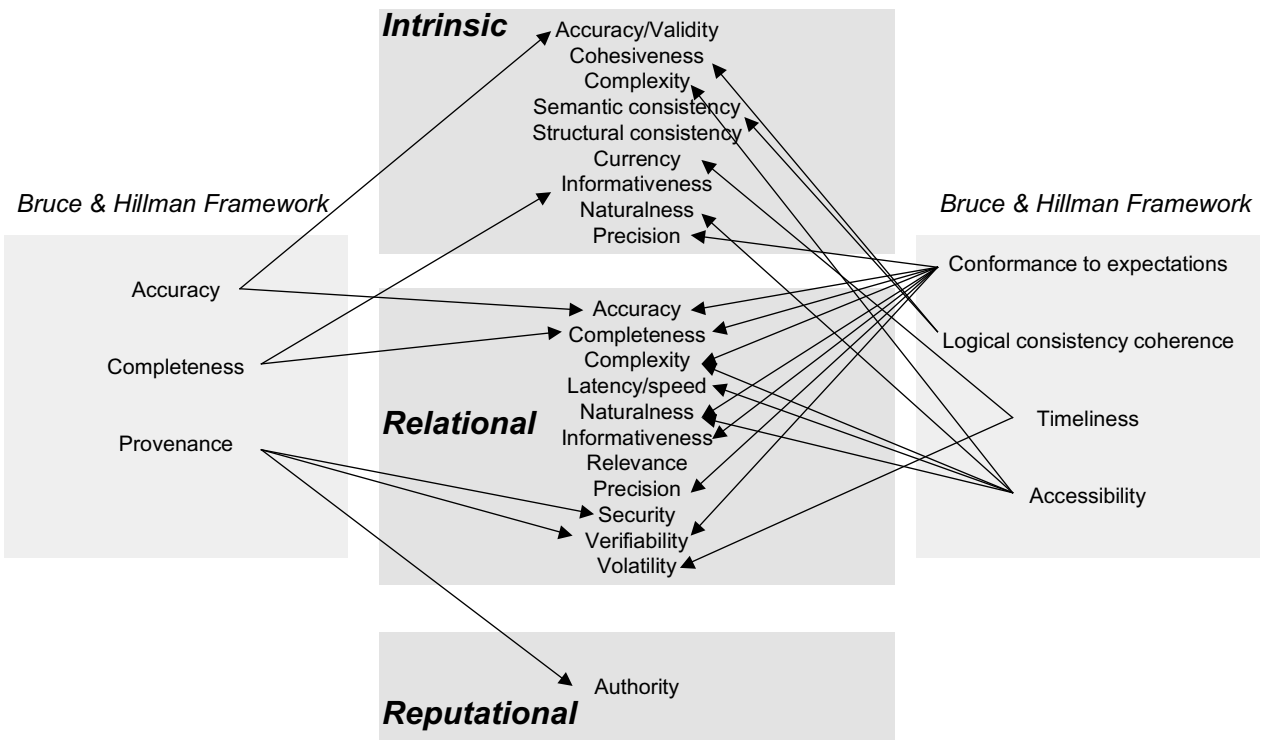
on one or more quality dimensions that renders data completely or largely unfit for use" (104). Consequently, to assess the size of the problem and its consequences on the outcome of the activity, one needs to have defined quality dimensions, measurements of the object's current quality, as well as information about the activity's specific quality requirements.

Until recently there has been little focus on developing measurements specifically for metadata quality. Bruce and Hillman (2004) offer a useful examination of characteristics of metadata quality particularly in light of its importance to aggregated collections. They outline seven general characteristics of metadata quality: completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility. In addition, they offer some possible criteria and compliance indicators for each, noting that shared metadata may require additional quality efforts.

In this study we rely on an information quality framework proposed by Gasser and Stvilia (2001) and Stvilia et al. (2004), which they have derived from the analysis of 32 representative quality assessment frameworks from the information quality literature. The framework is intended to be general enough to apply to different kinds of information as well as sufficiently specific to allow easy operationalization. Over one hundred characteristics of quality were extracted from the literature, examined for redundancy as well as for composite attributes which could be represented in combination, and then reduced to twenty-one quality dimensions (see Table 5 in Appendix One for all descriptions). The resulting set was organized into three information quality (IQ) categories: intrinsic IQ, relational/contextual IQ, and reputational IQ. The first two are relevant to the analysis presented here, and are described below.

**1. Intrinsic Information Quality (IQ):** Dimensions of information quality that can be assessed by measuring attributes of information items themselves in relation to a reference standard, such as spelling mistakes and conformance to a date encoding standard. In general, intrinsic IQ attributes are persistent, depend little on context, and can be measured more or less objectively. The dimensions within the intrinsic information quality category include: accuracy/validity, cohesiveness, complexity, semantic consistency, struc-

**Figure 1. Mapping between the Bruce and Hillman framework and the Gasser and Stvilia framework (Note: We have divided Bruce and Hillman measures into two sections for clarity.)**



tural consistency, currency, informativeness, naturalness, and precision.

**2. Relational/Contextual IQ:** Dimensions of information quality that depend on relationships between the information and some aspect of its usage context. This includes representational quality dimensions—those that measure how well an information object reflects some external condition (e.g., actual accuracy of addresses in an address database). Since metadata records are surrogates for information objects, many relational dimensions apply in measuring metadata quality. The dimensions within the relational/contextual information quality category are: accuracy, completeness, complexity, latency/speed, naturalness, informativeness, relevance (aboutness), precision, security, verifiability, and volatility.

This framework has considerable overlap with the set of dimensions suggested by Bruce and Hillman as illustrated in Figure 1.

As mentioned above, metadata quality must be judged in relation to the activities for which they are used. To get to the point of a useable metadata aggregation, several layers of activities transpire. In this paper we are principally concerned with three of these layers: the information design and creation activities of the resource provider, the "value-added" activities of the aggregator, and the use of the metadata aggregation itself by end-users. While we have qualitative and quantitative data to support discussion of the first two activities, few user studies of aggregated metadata collections exist. For the purposes of this paper, however, we are supposing that the primary purpose of the aggregated metadata system is similar to the purpose of other online library catalogs and databases, that is, to find, identify, select, and obtain items (IFLA 1998). Also important is the activity of collocating like resources (whether by subject, author, or other criteria) that Svenonius (2001) describes as the "primary act of information organization" (18) and which traditional library catalogs and databases do as a matter of course. This is typically difficult to do in a metadata aggregation because of the heterogeneity of the metadata harvested.

## Methods

The DCC project is currently harvesting metadata from 28 NLG funded digital collections using the OAI Protocol for Metadata Harvesting, with approximately 200,000 metadata records from 26 digital collections harvested to date. For this paper a subset of harvested metadata from four digital collections was analyzed, as detailed in Table 1. All records were harvested in unqualified or simple Dublin Core (http://www.dublincore.org/) via the OAI protocol. We should note that the OAI protocol requires that compliant OAI data providers provide metadata in at least simple Dublin Core; therefore, to meet the minimal requirements of the OAI protocol, many data providers map whatever metadata format is in use in their native database to simple Dublin Core. In the case of our four collections, none were using simple Dublin Core in their native system, so mapping is one of the activities required for this aggregation. Collection 1 uses a variation of qualified Dublin Core, and Collection 2 uses a locally developed metadata format. Collection 3 uses a slight variation of simple Dublin Core (an addition of a <note> element), but does not export this element in its OAI data provider, so in effect Collection 3 is exporting the Dublin Core record in use in their native database. Collection 4 uses a locally developed metadata format similar to qualified Dublin Core. Metadata from two of the digital collections are also available in additional metadata formats via the OAI protocol, however the project team only analyzed the common simple Dublin Core records.

We first performed a descriptive statistical analysis of the use and frequency of Dublin Core elements for each collection as described in Ward (2003) and Shreeves et al. (2003). In addition, we manually assessed a random sample of 35 records from each collection (except for Collection 4 where we took the entire set) for quality problems using the framework introduced above.

Two notes should be made about these samples. First, the 35 records from Collection 1 are from only 12 of the institutions contributing to the project. Second, the sample metadata from Collection 3 contained 14 (40%) records with only two elements, <title> and <identifier>, in use. They were kept in the data sample because they accurately represent the number of nearly-empty records in the entire collection (43%). These nearly-empty records demonstrate a type of quality issue caused by the systems in use. They are essentially placeholders for pieces of a compound object, such as a multi-paged document, and were exported through a content management system with a built-in OAI data

**Table 1. Characteristics of the four analyzed collections**

| | Collection 1 | Collection 2 | Collection 3 | Collection 4 |
|---|---|---|---|---|
| **Total number of records** | 27,444 | 14,425 | 1,599 | 35 |
| **Type of institution** | Large collaborative digitization project | Large academic library | Small academic library and public library collaboration | Small academic library |
| **Metadata from multiple institutions?** | Yes | No | Yes | No |
| **Type of resources described** | Photographs, artifacts, text. | Photographs | Legal documents, letters, government documents, maps | Texts |
| **Metadata mapped to simple Dublin Core from other metadata format?** | Yes; variation of Qualified Dublin Core in use. | Yes; local metadata format in use. | No; variation of simple Dublin Core in use, but only Dublin Core elements exported. | Yes; local metadata format similar to qualified Dublin Core. |
| **Notes about 35 record sample** | Represents metadata from 12 institutions | None | Contains 14 nearly empty records exported by the content management system. | Represents entire collection. |

provider which does not allow these records to be suppressed from exposure via the OAI data provider.

Supplementing the statistical analysis of the metadata records, qualitative data in the form of interview and focus group transcripts, and open-ended survey responses, were also examined. These qualitative data are being collected from the entire group of NLG awardees in the first and third years of the DCC project to monitor progress and change in metadata practices and perceptions. At the same time, we are conducting a series of case studies of selected projects based at academic, public, and state libraries, museums, historical societies, and other cultural heritage institutions, to capture the full range of operations and requirements of various services and users. This multi-method approach allows us to perform analysis across a large sample of projects to address general research questions while addressing specific research questions by a fuller analysis of a smaller, representative sample. For a fuller discussion of our qualitative methods and results see Palmer and Knutson (2004).

For this paper we focused on data from 13 interviews from 10 institutions represented in our sample set (including 7 institutions participating in Collection 1, already an aggregated set). We conducted the interviews with project managers and metadata specialists to discuss their experiences with collection building and metadata application. The interviews covered: 1) the history and background of the project, 2) elements considered important for collection level description, 3) potential use of a collection registry, 4) staffing and technical issues encountered applying item level metadata, and 5) current and expected usage of their digital collection. The interviews lasted between 40 and 80 minutes.

## Results and Discussion

For this paper, we have focused on a small set of quality dimensions: completeness of the metadata records, structural and semantic consistency, and ambiguity (a composite of relational precision, intrinsic naturalness and informativeness**)**. Throughout the discussion we consider local metadata practices and identify possible strategies aggregators could use to ameliorate metadata quality issues and strategies.

### *Completeness*

Completeness is a relational information quality dimension and is defined as the degree to which an information object matches the ideal representation for a given activity. Ideally, completeness should be judged on a record's sufficiency for use in the aggre-

gated database, that is, does it meet the requirements of finding, identifying, selecting, obtaining, and collocating? However, there has been little research into the utilization of specific Dublin Core elements for specific purposes. Greenberg (2001) demonstrates that most elements support discovery to some degree, but determining which elements are more important than others is largely dependent on the context and use of the system in question. This, of course, can change when metadata are taken out of their native environment and moved to an aggregated system, as in the case of the metadata analyzed in this study.

We judged completeness based on a published best practices guideline in use by one of the collections. This guideline states that a metadata record should contain at least eight elements of simple Dublin Core (<title>, <creator>, <subject>, <description>, <date>, <format>, <identifier>, and <rights>) as opposed to all fifteen elements. Both the characterization of incompleteness according to the reference standard (see Table 2) and the collection profiles (see Tables 6–13 in Appendix Two) indicate that none of the collections comply with this completeness standard. Collection 3 is an example of unintentional incompleteness because of the nearly-empty records discussed above. If those nearly-empty records were dropped, it would in fact meet the completeness requirement.

The use of Collection 4 in an aggregated environment is challenging because of the sparseness of its records; most contain only six distinct elements: <title>, <creator>, <type>, <language>, <identifier>, and <source>; 17 percent also include a <contributor> element. The result for an aggregator is that these records are essentially lost since the lack of descriptive metadata (such as subject headings or an abstract of the text) is likely to keep records from being retrieved even if they are immediately relevant. It is important to note that this is not a problem for use in their native environment, because navigation and search also rely on marked-up text and a richer metadata scheme. However, this context is lost once the descriptive metadata are exported via OAI in simple Dublin Core.

**Table 2. Percentage of Incomplete Records as Measured Against Use of Eight Dublin Core Elements**

| | Collection 1 | Collection 2 | Collection 3 | Collection 4 |
|---|---|---|---|---|
| **Percentage of incomplete records** | 69% | 71% | 43% | 100% |

Completeness also relates to the goals specific to individual projects and the metadata's "fitness for purpose." For example, as a collaborative project Collection 1 focuses on having an open approach and does not press contributing institutions for "perfect" metadata. They are aware of and willing to accept the trade-off in metadata quality to meet their goal of wide participation. One aspect of completeness that this collection faced in general is the use of the <title> element. They find that their museum partners do not always use the <title> element, preferring instead to use the <description> element for their objects. After all, what is the title of a rock? Participants interviewed for this study were generally aware of the possible tensions between interoperability and their local practice, but for a variety of reasons immediate local needs tended to take priority over the needs of interoperability.

***Consistency***

Within the intrinsic information quality category, there are two consistency dimensions. The first is semantic consistency or the extent to which the collections use the same values (vocabulary control) and elements for conveying the same concepts and meanings throughout. For example, is the type of material described included in the same element throughout the collection? The second is structural consistency or the extent to which similar attributes or elements of a metadata record are represented with the same structure and format. This dimension covers issues such as consistently encoding dates as YYYY-MM-DD throughout the collection.

In the interviews metadata providers' notion of consistency extended beyond structural and semantic consistency to include precision and informativeness dimensions, which are problems that affect their own end users, whether or not the collection is aggregated. For instance, the level of detail included in a description can vary dramatically from brief one-liners to historical diatribes. Furthermore, the granularity of subject headings varies by the cataloger who assigns them. These problems are of particular concern to participants from Collection 1 where the different professional practices and principles of libraries and museums have to be understood and negotiated.

Consistency is highly significant for aggregators. While the ideal is to have

semantic and structural consistency across all harvested collections, in general it is easier for aggregators to normalize metadata across collections if internal consistency for each collection exists. This is true even when the use of an element or the way a value is encoded is regarded as generally incorrect. If the type of material is consistently described in the <source> element, an aggregator can better cope with normalizing this information to, for example, make the information appear in the <type> element, than if this information appears variously in the <type>, <source>, and <description> elements.

Although we report here on only one area—the placement and encoding of date information in the four collections—we observed similar issues for several other types of information about the resources including their type, format, hosting institution, and geographic coverage.

### *Date*

Variation in where and how dates are recorded is mentioned frequently in the literature as a problematic area for aggregators and practitioners alike (Dushay and Hillman 2003, Shreeves et al. 2003, Barton et al. 2003). Based on the general Dublin Core Metadata Initiative (DCMI) guidelines (DCMI 2004) we expect to find the date "associated with an event in the life cycle of the resource", typically the creation of the resource, in the <date> element while we expect to find dates relating to "the extent or scope of the content of the resource" in the <coverage> element (http://dublincore.org/documents/dcmi-terms/). In addition, date information associated with the resource creator is commonly found in the <creator> and/or <contributor> elements. The <description> element sometimes contains date information as appropriate to the text included there. In addition, date information is often included in the <subject> element within subject headings (particularly Library of Congress Subject Headings).

Table 3 shows our analysis of where date information relating to the creation of the resource and coverage of the resource is found in each of the four collections. In general Collection 2 and 4 are consistent about where date information is placed and how the date is encoded. However, Collection 4 includes the date of publication at the end of a string in the <source> element. This string is typically the name of the publisher of the resource, though sometimes it also includes complete bibliographic information (author, title, and publisher). An aggregator could typically create a program to normalize these dates, for instance to copy the date information in the <source> element in Collection 4 to the <date> element, and to copy and truncate the date information in Collection 2 to the common encoding scheme of YYYY.

Collections 1 and 3 are more complicated for aggregators. Date information is not consistently recorded in one location, and is, in fact, often recorded in multiple locations in a single record. Generally encoding schemes for the dates are not consistent. Obviously it can be expensive for an aggregator to try to cope with these internal inconsistencies. Particularly problematic are the cases of date information recorded in multiple locations within a single record. As Barton et al. (2003) note, an aggregator must determine what event(s) in the resource life-cycle is documented by the date(s). In the case of Collection 3 the date information in the <date> element and <title> element are essentially the same. However, in Collection 1, the 20 records with multiple dates recorded generally include the date that the original print resource was published or created (usually an older date) and the date the digital resource was created (generally from 1998 to present). This ambiguity will be discussed further below, but, suffice it to say, in order to normalize this information, an aggregator must determine where it is recorded and which date or dates to normalize.

In our interviews date came up time and again as a problematic field for practitioners. Maintaining structural consistency did not seem as difficult for metadata creators as semantic consistency. Decisions about whether to cite the publication date of the original resource or the date it was digitized were not straightforward. For non-published items the date was sometimes unknown for the original, and circa or date ranges would be used. This is largely where the structural consistency problems came in. If you have resources with an exact day, month, and year as well as resources with only a date range, they will necessarily have a different structure (MM-DD-YYYY vs. YYYY-YYYY).

## Ambiguity

The last quality point we examined was adherence to the Dublin Core One-to-One principle. The principle states that: "Dublin Core metadata describes one manifestation or version of a resource, rather than as-

| Table 3. Location of date information relating to the creation and coverage of the resource | | | | |
|---|---|---|---|---|
| **Date information included in:** | **Collection 1** | **Collection 2** | **Collection 3** | **Collection 4** |
| <date> element (used once) | 9 (26%) | 35 (100%) | 20 (57%) | 0 |
| <date> element (used at least twice) | 20 (57%) | 0 | 0 | 0 |
| <coverage> element (used once) | 0 | 0 | 17 (49%) | 0 |
| Date in other element | 0 | 0 | 21 (60%)<title> | 35 (100%)At end of <source> string |
| Not recorded | 6 (17%) | 0 | 14 (40%)(nearly empty records) | 0 |
| Notes | Inconsistent encoding schemes within individual records with multiple date elements as well as between records. | Consistent encoding scheme (YYYY-MM-DD). | Consistent encoding scheme in <coverage> element (YYYY-YYYY); Inconsistent encoding scheme in all others. | Consistent encoding scheme. (YYYY). |

suming that manifestations stand in for one another" (Hillman 2003). The classic example is that Leonardo da Vinci's painting of Mona Lisa is not the same resource as a digital photograph of the painting which is not the same resource as a physical photograph of the painting. Theoretically, there should be a metadata record for each of these manifestations and links made among them (if desired) through use of the relation and source elements.

In practice, however, metadata authors find it difficult to maintain this one-to-one mapping. Metadata records—particularly those describing digitized resources—often are a composite of descriptions of both the physical and digital item, as was the case with some of the multiple date instances discussed above. This sort of many to one mapping—when the metadata record represents two or more resources simultaneously—results in an ambiguity problem, which can be evaluated using three quality dimensions from the Gasser and Stvilia framework: relational precision, intrinsic naturalness, and informativeness. For aggregators reliant on automated processing, this ambiguity can be particularly problematic. If two dates are present—one for the digital resource and one for the source or physical resource—how does an aggregator determine which to use? Often it is not clear, even from a visual examination of a record, to which resource the date information refers.

In the analysis of our sample records, no collection maintained a consistent one-to-one mapping between the metadata record and a single resource. Table 6 shows that all records in Collection 2 contained some sort of ambiguity; in this case it was the inclusion of a <format> element describing the digital image (image/jpg) and the physical image (35 mm slide). In Collection 3 all elements (including <date> and <rights>) except one describe the physical object, yet the format for each record was 'jpeg'—obviously referring to the digital object. Note that if all the nearly empty records in Collection 3 (43%) were dropped out, 100 percent of the records would have this ambiguity.

Collection 1 contained some of the most ambiguous records. As noted earlier, 57 percent of the records contained at least two <date> elements, one of which referred to the date the digital resource was created and the other to the creation of the physical source. In addition, 51 percent of the records contained references to at least two formats with one in the <format> element and the second in another element. Format is defined as the "the physical or digital manifestation

| Table 4. Percentage of metadata in collections that do not meet the one to one principle | | | | |
|---|---|---|---|---|
| **Percentage of records that describe at least two manifestations of a resource** | **Collection 1** | **Collection 2** | **Collection 3** | **Collection 4** |
| | 86% | 100% | 57% | 69% |

of the resource" (DCMI 2004) and usually refers to the media type or software requirements for digital resources or the specific manifestation (e.g., a 35 mm slide is a manifestation of a photograph) or the dimensions of a physical resource. Most of these records, like those in Collection 2, refer to a format for the digital resource (like 'GIF') and a format for the physical source (like '21 x 26 inches'). However, 14 (40%) of the records contain a reference to at least two, if not three, digital formats. For example, one record contained the following format information:

**<description>** 100 x 70 cm
**<description>** image/tiff
**<format>** image/jpeg
**<format>** Any machine capable of running graphical Web browsers, 640x480 minimum monitor resolution

Presumably, the first <description> element refers to the size of the physical source (a lithograph). But are we to assume that the second <description> element refers to the tiff image that is commonly made as the archival copy, and from which the jpeg image (in the <format> element) is derived? Many of the records in Collection 1 include this mix of information. Several records include in the <description> elements an account of the digitization process as well as a description of what is pictured in the photograph. Other metadata from Collection 1 document format information about the digital resource in the <format> element and format information about the physical resource in the <source> element. This is technically correct, but it represents a dilemma for the aggregator. Which information—that about the physical resource or the digital resource—is most important to normalize? On which descriptive information do we base collocation decisions?

These practices represent a very real and understandable tension between the need for standardized, accurate description of digital objects and description that meets the needs of end users. Interview participants expressed frustration with this situation, noting that strictly adhering to the one-to-one principle and describing only the digital object was not helpful to the users of their collection. For example, if the collection contains a document created in 1908, that date is important to the searcher, not the fact that it was digitized in 2002. The date of origin helps users identify content from a particular time period, and it helps differentiate among similar documents with different original dates of publication or creation. With the one-to-one description, photographs that span a century, but are digitized at the same time, would have the same date. Likewise, interview participants also considered descriptions of the original objects more important than the digitized object for end-user discovery. The decision as to where to put the descriptions of the original objects varied, however from the <description> element, to the <format> element, to the <source> element and others.

## Conclusion

Even though the majority of interviewees expressed concern with the consistency of their metadata, the analysis of Collections 2, 3, and 4 suggests that metadata created by a single or pair of institutions is less susceptible to varied interpretation in part because it is created in the same local circumstance with the same use in mind. However, in Collection 1, where collections from multiple institutions are aggregated, the variance in the metadata sharply increases, which can complicate most of the activities aggregators wish to support. Such variance might force aggregators to orient their services towards the minimum level of quality in the collection. Aggregated search services may not be able to implement the well functioning, standard services offered by online catalogs, such as browse interfaces or searches targeted at specific fields such as title, creator, or subject.

Although this study was limited to a handful of quality dimensions on a small sample, it points to at least two specific strategies that resource creators interested in sharing their metadata can take to aid aggregators in using their metadata most effectively. Both structural and semantic consistency can allow aggregators to easily process and normalize metadata. Eliminating as much ambiguity as possible helps ag-

gregators interpret what specifically is described and to process metadata accordingly.

There are other ways that resource developers can aid aggregators that cannot be explored within the constraints of this paper. These include the provision of semantically rich metadata (such as MARC, MODS, or qualified Dublin Core) which may allow aggregators to make better use of the metadata values. Further research is needed to determine whether provision of metadata more complex than simple Dublin Core is effective in alleviating some of the quality problems outlined above. Moreover, making metadata documentation—such as metadata formats, controlled vocabularies, and mappings used—publicly available can also help aggregators better interpret harvested metadata. Finally, further exploration of the metadata quality framework outlined in this paper is much needed. Connecting metadata quality not only to theoretical standards, but also to information use activities, would help aggregators and the digital library community at large better understand what to prioritize for quality control. Research is needed to understand the trajectory of metadata as it travels from the initial design of the cataloging workflow to its use in a federated collection.

## Acknowledgements

This research presented here was funded through National Leadership Grant number LG-02-02-0281 from the Institute of Museum and Library Services.

## References

Barton, Jane, Sarah Currier, and Jessie M.N. Hey. 2003. Building quality assurance into metadata creation: an analysis based on the learning objects and e-prints communities of practice. In *DC-2003: Proceedings of the International DCMI Metadata Conference and Workshop.* [United States]: DCMI. http://www.siderean.com/dc2003/201_paper60.pdf.

Bruce, Thomas R., and Diane I. Hillmann. 2004. The continuum of metadata quality: defining, expressing, exploiting. In *Metadata in Practice*, Edited by Diane I. Hillmann and Elaine L. Westbrooks. Chicago: American Library Association.

Dublin Core Metadata Initiative. 2004. *Dublin Core Metadata Terms*. http://dublincore.org/documents/dcmi-terms/.

Dushay, Naomi, and Diane I. Hillmann. 2003. Analyzing metadata for effective use and re-use. In *DC-2003: Proceedings of the International DCMI Metadata Conference and Workshop.* [United States]: DCMI. http://www.siderean.com/dc2003/501_Paper24.pdf.

Gasser, Les, and Besiki Stvilia. 2001. *A new framework for information quality*. Technical report ISRN UIUCLIS--2001/1+AMAS. Champaign, Ill.: University of Illinois at Urbana Champaign.

Gertsbakh, Il´ia B. 1977. *Models of preventive maintenance*. Studies in mathematical and managerial economics. vol. 23. Amsterdam, Holland: North-Holland Publishing Company.

Greenberg, Jane. 2001. Quantitative categorical analysis of metadata elements in image applicable metadata schemas. *Journal of the American Society for Information Science and Technology* 52, no. 11: 917–24.

Hagedorn, Kat. 2003. OAIster: A 'No Dead Ends' OAI Service Provider. *Library Hi Tech* 21: 170–81.

Hillman, Diane. 2003. *Using Dublin Core*. http://dublincore.org/documents/usageguide/.

IFLA Study Group on the Functional Requirements of Bibliographic Records. 1998. *Functional Requirements of a Bibliographic Record: Final Report.* UBCIM Publications. New Series. vol. 19. Edited by Marie-France Plassard. München: K.G. Saur Verlag GmbH & Co. KG.

Palmer, Carole L., and Ellen M. Knutson. 2004. Metadata practices and implications for federated collections. In *Proceedings of the 67th Annual Meeting of the American Society for Information Science and Technology*, Edited by Linda Schamber and Carol L. Barry. Medford, N.J.: Information Today, Inc: 456–62.

Lagoze, Carl, and Herbert Van de Sompel. 2001. The Open Archives Initiative: building a low-barrier interoperability framework. In *Proceedings of First ACM/IEEE-CS Joint Conference on Digital Libraries: June 24-28, 2001: Roanoke, Virginia, USA,* Edited by Edward A. Fox and Christine L. Borgman. New York: ACM Press: 54–62.

Pipino, L., Yang W. Lee, and Richard Y. Wang. 2002. Data quality assessment. *Communications of the ACM* 45, no. 4: 211–18.

Shreeves, Sarah L., Joanne Kaczmarek, and Timothy W. Cole. 2003. Harvesting cultural heritage metadata using the OAI protocol. *Library Hi Tech* 21: 159–69.

Strong, Diane, Yang W. Lee, and Richard Y. Wang. 1997. Data quality in context. *Communications of the ACM* 40, no. 5: 103–10.

Stvilia, Besiki, Les Gasser, Michael Twidale, Sarah L. Shreeves, and Timothy W. Cole. 2004. Metadata quality for federated collections. In *Proceedings of ICIQ04—9th International Conference on Information Quality. Cambridge, MA*: 111–25.

Svenonius, Elaine. 2001. *The intellectual foundation of information organization*. Cambridge, Mass.: MIT Press.

Wand, Yair and Richard Y. Wang. 1996. Anchoring data quality dimensions in ontological foundations. *Communications of the Association for Computing Machinery* 39, no.11: 86–95.

Ward, Jewel. 2003. A Quantitative Analysis of Unqualified Dublin Core Metadata Element Set Usage within Data Providers Registered with the Open Archives Initiative. In *Proceedings: 2003 Joint Conference on Digital Libraries (JCDL 2003)*, Edited by Catherine C. Marshall, Geneva Henry, and Lois M.L. Delcambre. Institute of Electrical and Electronics Engineers, Inc., Los Alamitos, Calif.: 315–17.

# Appendix 1

| Table 5. Gasser & Stvilia Information Quality Categories, Dimensions, and Definitions | | |
|---|---|---|
| Category | **Dimension** | **Definition** |
| **Intrinsic** | Accuracy/ Validity | the extent to which information is legitimate or valid according to some stable reference source such as a dictionary, standard schema and/or set of domain constraints and norms |
| | Cohesiveness | the extent to which the content of an object is focused on one topic |
| | Complexity | the extent of cognitive complexity of a information object measured by some index/indices |
| | Semantic consistency | the extent to which the same values and elements are used for conveying the same concepts and meanings in an information object. |
| | Structural consistency | the extent to which similar attributes or elements of an information object are represented with the same structure & format |
| | Currency | the age of an information object |
| | Informativeness | the amount of information contained in an information object: the ratio of the size of the informative content (measured in word terms which are stemmed and stopped) to the overall size of an information object. |
| | Naturalness | the extent to which an information object's model/schema and content are expressed by conventional, typified terms and forms according to some general purpose reference source |
| | Precision | the granularity or precision of an information object's model or content values according to some general purpose IS-A ontology such as WordNet. |
| **Relational/ Contextual** | Accuracy | the degree to which an information object correctly represents another information object or a process in the context of a particular activity. |
| | Completeness | the degree to which an information object model matches the ideal representation model of a given activity. |
| | Complexity | the degree of cognitive complexity of a information object relative to a particular activity |
| | Latency/speed | the speed of access to an information object relative to the context of a particular activity |
| | Naturalness | the degree to which an information object's model and content are semantically close to the objects, states or processes they represent in the context of a particular activity (measured against the activity/community specific ontology) |
| | Informativeness | the extent to which the information is new or informative in the context of a particular activity/community |
| | Relevance (aboutness) | the extent to which information is applicable and helpful in a given activity |
| | Precision | the extent to which an information object matches the precision and granularity needed in the context of a given activity |
| | Security | the extent of protection of information from harm |
| | Verifiability | the extent to which the correctness of information is verifiable and/or provable |
| | Volatility | the amount of time the information remains valid |
| Reputational | Authority | the degree of reputation of an information object in a given community |

**Table 6. Collection 1–Use and non-use of Dublin Core elements**

| Dublin Core element | % of institutions using element at least once | No. of records containing element | Total times element used | % of total records containing element | Average times used per record | Average element length (in characters) | Mode | Mode Frequency in % |
|---|---|---|---|---|---|---|---|---|
| <title> | 100 | 27,442 | 31,765 | 100 | 1 | 42 | 1 | 88 |
| <creator> | 93 | 16,215 | 17,363 | 59 | 1 | 16 | 1 | 55 |
| <subject> | 100 | 26,610 | 112,189 | 97 | 4 | 24 | 3 | 22 |
| <description> | 100 | 26,326 | 77,531 | 96 | 3 | 92 | 2 | 54 |
| <publisher> | 100 | 27,444 | 53,872 | 100 | 2 | 36 | 2 | 67 |
| <contributor> | 47 | 2,267 | 5,581 | 8 | 2 | 23 | 0 | 92 |
| <date> | 97 | 23,955 | 40,828 | 87 | 2 | 9 | 2 | 46 |
| <type> | 73 | 19,342 | 26,598 | 70 | 1 | 6 | 1 | 44 |
| <format> | 70 | 16,174 | 24,633 | 59 | 2 | 19 | 0 | 41 |
| <identifier> | 100 | 27,440 | 33,344 | 100 | 1 | 68 | 1 | 80 |
| <source> | 57 | 13,955 | 28,257 | 51 | 2 | 34 | 0 | 49 |
| <language> | 67 | 12,220 | 12,416 | 45 | 1 | 5 | 0 | 55 |
| <relation> | 100 | 27,214 | 47,125 | 99 | 2 | 27 | 2 | 60 |
| <coverage> | 17 | 646 | 778 | 2 | 1 | 22 | 0 | 98 |
| <rights> | 87 | 24,927 | 31,371 | 91 | 1 | 92 | 1 | 81 |

**Table 7. Collection 1 - Statistical characterization of use of Dublin Core elements**

| | N | Minimum | Maximum | Mean | Standard Deviation | Mode | Mode Frequency |
|---|---|---|---|---|---|---|---|
| **Total number of elements per record** | 2000 | 4 | 34 | 20 | 4.52 | 18 | 0.15 |
| **Number of distinct elements per record** | 2000 | 4 | 15 | 11 | 1.56 | 11 | 0.2 |

| Table 8. Collection 2–Use and non-use of Dublin Core elements | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Dublin Core element** | **No. of records containing element** | **Total times element used** | **% of total records containing element** | **Average times used per record** | **Average element length (in characters)** | **Mode** | **Mode Frequency in %** |
| <title> | 14,346 | 29,172 | 99 | 2 | 38 | 2 | 82 |
| <creator> | 14,425 | 14,425 | 100 | 1 | 34 | 1 | 100 |
| <subject> | 14,421 | 115,628 | 100 | 8 | 12 | 6 | 13 |
| <description> | 3,767 | 4,863 | 26 | 1 | 17 | 0 | 74 |
| <publisher> | 14,425 | 28,850 | 100 | 2 | 47 | 2 | 100 |
| <contributor> | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| <date> | 14,407 | 14,407 | 100 | 1 | 10 | 1 | 100 |
| <type> | 14,425 | 45,481 | 100 | 3 | 12 | 3 | 80 |
| <format> | 14,425 | 28,850 | 100 | 2 | 10 | 2 | 100 |
| <identifier> | 14,425 | 43,275 | 100 | 3 | 35 | 3 | 100 |
| <source> | 14,425 | 14,425 | 100 | 1 | 59 | 1 | 100 |
| <language> | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| <relation> | 14,425 | 14,425 | 100 | 1 | 57 | 1 | 100 |
| <coverage> | 14,339 | 15,039 | 99 | 1 | 47 | 1 | 95 |
| <rights> | 14,425 | 14,425 | 100 | 1 | 57 | 1 | 100 |

| Table 9. Collection 2. Statistical characterization of use of Dublin Core elements | | | | | | | |
|---|---|---|---|---|---|---|---|
| | **N** | **Minimum** | **Maximum** | **Mean** | **Standard Deviation** | **Mode** | **Mode Frequency** |
| Total number of elements per record | 2000 | 15 | 38 | 26 | 3.48 | 31 | 0.2 |
| Number of distinct elements per record | 2000 | 9 | 13 | 12 | 0.55 | 12 | 0.99 |

**Table 10. Collection 3 - Use and non-use of Dublin Core elements**

| Dublin Core element | No. of records containing element | Total times element used | % of total records containing element | Average times used per record | Average element length (in characters) | Mode | Mode Frequency in % |
|---|---|---|---|---|---|---|---|
| <title> | 1,599 | 1,599 | 100 | 1 | 45 | 1 | 100 |
| <creator> | 909 | 909 | 57 | 1 | 41 | 1 | 57 |
| <subject> | 919 | 919 | 57 | 1 | 140 | 1 | 57 |
| <description> | 915 | 915 | 57 | 1 | 390 | 1 | 57 |
| <publisher> | 92 | 92 | 6 | 1 | 42 | 0 | 94 |
| <contributor> | 705 | 705 | 44 | 1 | 65 | 0 | 56 |
| <date> | 909 | 909 | 57 | 1 | 9 | 1 | 57 |
| <type> | 914 | 914 | 57 | 1 | 4 | 1 | 57 |
| <format> | 785 | 785 | 49 | 1 | 4 | 0 | 51 |
| <identifier> | 1,599 | 2,305 | 100 | 1 | 85 | 1 | 56 |
| <source> | 907 | 907 | 57 | 1 | 137 | 1 | 57 |
| <language> | 914 | 914 | 57 | 1 | 4 | 1 | 57 |
| <relation> | 916 | 916 | 57 | 1 | 55 | 1 | 57 |
| <coverage> | 711 | 1,412 | 44 | 2 | 7 | 0 | 56 |
| <rights> | 919 | 919 | 57 | 1 | 121 | 1 | 57 |

**Table 11. Collection 3 - Statistical characterization of use of Dublin Core elements**

| | N | Minimum | Maximum | Mean | Standard Deviation | Mode | Mode Frequency |
|---|---|---|---|---|---|---|---|
| Total number of elements per record | 1,599 | 0 | 16 | 9 | 6.51 | 2 | 0.5 |
| Number of distinct elements per record | 1599 | 0 | 15 | 9 | 5.72 | 2 | 0.5 |

| Table 12. Collection 4-Use and non-use of Dublin Core elements | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Dublin Core element** | **No. of records containing element** | **Total times element used** | **% of total records containing element** | **Average times used per record** | **Average element length (in characters)** | **Mode** | **Mode Frequency in %** |
| <title> | 35 | 64 | 100 | 2 | 27 | 2 | 49 |
| <creator> | 29 | 30 | 83 | 1 | 15 | 1 | 80 |
| <subject> | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| <description> | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| <publisher> | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| <contributor> | 6 | 6 | 17 | 1 | 14 | 0 | 83 |
| <date> | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| <type> | 35 | 35 | 100 | 1 | 4 | 1 | 100 |
| <format> | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| <identifier> | 35 | 35 | 100 | 1 | 72 | 1 | 100 |
| <source> | 35 | 35 | 100 | 1 | 66 | 1 | 100 |
| <language> | 35 | 35 | 100 | 1 | 3 | 1 | 100 |
| <relation> | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| <coverage> | 0 | 0 | 0 | 0 | 0 | 0 | 100 |
| <rights> | 0 | 0 | 0 | 0 | 0 | 0 | 100 |

| Table 13. Collection 4 - Statistical characterization of use of Dublin Core elements | | | | | | | |
|---|---|---|---|---|---|---|---|
| | **N** | **Minimum** | **Maximum** | **Mean** | **Standard Deviation** | **Mode** | **Mode Frequency** |
| Total number of elements per record | 35 | 6 | 8 | 6.79 | .692 | 7 | 0.117 |
| Number of distinct elements per record | 35 | 6 | 6 | 6.00 | 0 | 6 | 1 |