METADATA QUALITY FOR FEDERATED COLLECTIONS

(Completed Paper)
IQ Concepts, Models, Case Studies

Besiki Stvilia[1], Les Gasser[1], Michael B. Twidale[1], Sarah L. Shreeves[2], Tim W. Cole[2]
[1] Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 501 E. Daniel Street, Champaign, IL 61820, USA
[2] University Library, University of Illinois at Urbana-Champaign, 1408 West Gregory Drive, Urbana, IL 61801, USA
{stvilia, gasser, twidale, sshreeves, t-cole3}@uiuc.edu

Abstract This paper presents early results from our empirical studies of metadata quality in large corpuses of metadata harvested under Open Archives Initiative (OAI)[1] protocols. Along with some discussion of why and how metadata quality is important, an approach to conceptualizing, measuring, and assessing metadata quality is presented. The approach given in this paper is based on a more general model of information quality (IQ) for many kinds of information beyond just metadata. A key feature of the general model is its ability to condition quality assessments by context of information use, such as the types of activities that use the information, and the typified norms and values of relevant information-using communities. The paper presents a number of statistical characterizations of analyzed samples of metadata from a large corpus built as part of the Institute of Museum and Library Services Digital Collections and Contents (IMLS DCC)[2] project containing OAI-harvested metadata and links these statistical assessments to the quality measures, and interprets them. Finally the paper discusses several approaches to quality improvement for metadata based on the study findings.

1. INTRODUCTION

The usability and effectiveness of any digital library is clearly affected by the quality of its metadata records. Low quality metadata can render a library almost unusable, while high metadata quality can lead to higher user satisfaction and increased use. Consequently, digital library infrastructures should include effective quality assurance mechanisms of its metadata collections.

This paper presents results from our empirical studies of metadata quality in large corpuses of metadata harvested under the Open Archives Initiative protocols. Along with some discussion of why and how metadata quality is important, we present an approach to conceptualizing, measuring, and assessing metadata quality. The approach we give in this paper is based on a more general model of information quality (IQ) for many kinds of information beyond just metadata [8]. A key feature of the general model is its ability to condition quality assessments by context of information use, such as the types of activities that use the information, and the typified norms and values of relevant information-using communities. We present a number of statistical characterizations of analyzed samples of metadata from a large corpus built as part of the IMLS DCC project containing OAI-harvested metadata. We link these statistical assessments to our quality measures, and interpret them. Finally, we discuss several approaches to quality improvement for metadata based on our findings.

1.1 Overview of Approach

This paper presents a general model for analyzing and reasoning about information quality in large aggregated metadata repositories. The model has been developed using a number of techniques such as

[1] http://www.openarchives.org
[2] http://imlsdcc.grainger.uiuc.edu

literature analysis, case studies, statistical analysis, strategic experimentation, and multi-agent modeling. The model along with the concepts and metrics presented in this paper can serve as a foundation for developing effective specific methodologies of quality assurance in various types of organizations.

Our model of metadata quality ties together findings from existing and new research in information quality, along with well-developed work in information seeking/use behavior, and the techniques of strategic experimentation from manufacturing. It presents a holistic approach to determining the quality of a metadata object, identifying quality requirements based on typified contexts of metadata use (such as specific information seeking/use activities) and expressing interactions between metadata quality and metadata value.

We include findings from the statistical analysis and experimentation with an aggregated metadata collection built as a part of the IMLS DCC project, that suggest general types and values for metadata quality metric functions. However, since quality metrics are context-sensitive, at this point these statistics provide only suggestive guidance for building such functions in general. Specific quality metrics and their value ranges can be only determined based on specific types of metadata and its local cost and value structures. Due to size limitations, the paper also does not include the full taxonomy of metadata quality dimensions, metrics and related metadata creation and use activities we developed as a part of this ongoing research.

1.2 Background and Related Research

Libraries, archives and museums are some of the oldest institutions dealing with information. Their processes of adding value to information through selection, organization and access also include, explicitly or implicitly, the reasoning and decision making about the quality of metadata and metadata tools [12,24]. Frameworks for measuring metadata quality and methodologies for identifying metadata quality priorities are proposed in [3]. Quality assurance problems of online database aggregators are discussed in [30]. In the spirit of Orr's systems theory view [18], [26] proposes an inexpensive technique of assessing and improving quality of a museum information repository through establishing a feedback link between information creators and users. There is a well developed body of research literature on data & information quality assurance in management sciences and the database world. [29] discusses impacts of poor data quality and proposes a framework of data quality dimensions. A methodology of assessing information quality through modeling information manufacturing systems is developed in [2]. [21] discusses a methodology for reasoning about information quality through process and task analysis. A comprehensive overview of information & data quality assessment tools and methodologies is given in [28]. [15] and [16] introduce techniques for quality-based evaluation of query results and source selection in a multi-database environment.

Existing research in information & data quality assurance provides a valuable knowledge base of methodologies, techniques and concepts for assessing and improving quality of information. However, for enabling effective and realistic quality assurance planning and/or simply justifying quality improvement spending, there is a need to connect the changes in an information object's quality to changes in its value and cost in some sound, systematic way.

In addition, one of the thorniest issues in the theory of information quality is how to account for the context-sensitive nature of information quality and value: the same information may have different kinds and levels of quality and value in different contexts of use. Studies of user information seeking and use behavior [14,19,22] provide the anchor for a set of generic activities that can be used to establish "typified" contexts for (and types of) information use. We can construct more generic views of the cost, value, and quality of information in relation to these typified contexts. In addition, our current research draws on insights from manufacturing [6,23] to connect information quality to the cost and value of information and to account for how changes in cost and value over time influence (and are influenced by) information quality. Using theses insights, we develop concepts, methods and software tools for modeling

quality, value and cost structures of typified metadata objects and link them consistently with typified contexts of their creation and use.

The paper is organized as follows. In the sections #2.1-2.2 we introduce concepts and foundations of our approach. In addition, using an example of a typified information seeking activity, we explain how analysis of a given activity and its actions can be used for identifying relevant quality dimensions, and important tradeoffs between those dimensions and the types of quality metric functions. In section #2.3 we present three models of quantifying the impacts of quality changes on the value of a metadata object. In section #3 we present some empirical observations and the results of statistical analysis of random samples from IMLS DCC aggregated metadata collection and interpret them through the methodology prism formulated in the earlier sections. And, finally we conclude the paper with a summary of the proposed concepts and methods and identify the directions of future research activities.

2. CONCEPTS AND METHODS

The Oxford English Dictionary defines metadata as "a set of data that describes and gives information about other data." Thus, in addition to being an information object itself, a metadata object serves as a tool that provides access to (and other services for) other information. For example, [11] specifies a "typified" [13] set of information activities such as finding, identifying, and obtaining information, which bibliographic metadata records are intended to support. Differences in the quality of metadata records—assuming we could assess it---would lead us to predict to differences in the ability of that metadata to support the activities it should support. High standardization of metadata objects and availability of more or less standardized baseline representations may somewhat reduce the complexity of quality assessment, but they hardly eliminate it due to the contextual nature of information quality. In addition, metadata objects are generally complex, composite objects consisting of multiple components. Their structure ranges from schematic objects with multiple fields (e.g., Dublin Core, MARC) to complex knowledge networks (e.g., Semantic Web annotations such as DAML+OIL). As a result, we need a methodology that identifies contextual quality requirements for composite information objects and transforms them consistently into their composite (schema) and component quality requirements, as well as translating quality changes into changes of information value and cost.

We start with a framework of quality dimensions that can be used for assessing quality both at the composite and component levels. Following that, for effective reasoning about interdependencies among the quality dimensions and their impacts on aggregate quality assessments, we propose the use of quality curves modeled after the Taguchi value curves [23]. The subsection #2.2 is ended by defining a general approach/procedure for value- and cost-sensitive quality assurance. The components of the approach then are discussed in greater detail in the subsections 2.3-2.4 and we finish the section with a discussion of quality assurance challenges that an OAI metadata aggregator[3] may encounter. Note that here and throughout the paper we use a term aggregator to refer to this specific service type which collects metadata records from different OAI data providers and makes them available for gathering by others using the OAI Metadata Harvesting Protocol (OAI-MHP). The records are expected to be using a common standard schema – Simple Dublin Core (DC)[4]. Therefore, this particular aggregator does not integrate or aggregate information from different records through a mediated schema like some of the information integration services aka wrappers do, though, as we will discuss later in the paper, they still share a number of similar problems.

[3] http://www.oaforum.org/tutorial/english/page6.htm
[4] http://www.dublincore.org/

2.1 Metadata Quality Dimensions

Almost as many different taxonomies of IQ dimensions have been proposed as there are writings about IQ. While these classifications differ in the granularity, detail, and naming of IQ dimensions, there are substantial overlaps among them. In the previous research [8] we analyzed existing information, data, and metadata quality approaches, compiled a comprehensive taxonomy of IQ dimensions, and created a firmer and more unified theoretical foundation for them. This taxonomy consists of 38 dimensions divided into 3 categories:

1. **Intrinsic IQ:** Some dimensions of information quality can be assessed by measuring attributes of information items themselves, in relation to a reference standard. Examples include spelling mistakes (dictionary), conformance to formatting or representation standards (HTML validation), and information currency (age with respect to a standard index date, e.g. "today"). In general, Intrinsic IQ attributes persist and depend little on context, hence can be measured more or less objectively.
2. **Relational/Contextual IQ:** This category of IQ dimensions measures *relationships* between information and some aspects of its usage context. One common subclass in this category includes the *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 objects are always surrogates for (hence bear a relationship to) other information objects, many relational dimensions apply in measuring metadata quality (e.g. whether an identifier such as URL or ISBN actually identifies the right document; whether a title field holds the actual title). Clearly, since related objects can change independently, relational/contextual dimensions of an information item are not persistent with the item itself.
3. **Reputational IQ:** This category of IQ dimensions measures the position of an information artifact in cultural or activity structure, often determined by its origin and its record of mediation.

Due to space reasons we use only a small subset from our entire set of IQ dimensions in Section 2.2.1 to illustrate the importance of metadata quality to an information seeking activity.

2.2 Activities and Metadata Quality Metrics

Studies have repeatedly shown that information quality assessments are contextual [2,21]. In order to measure the general quality of a metadata object a) we need to understand metadata creation and use activities, b) we need to define procedures that transform attributes of metadata, its target objects (documents), and relevant contextual features into values for metadata quality, at both component and composite levels, and c) we need to condition both of these by the activity or process context [2].

Specific metadata quality problems arise when the *existing* quality level along some particular metadata dimension is lower than the *required* quality level, in the context of using that metadata to support a given activity. In general, metadata is generated based on the properties of a given information item and/or its information content by some surrogation process which itself uses a certain surrogation model or schema. The schema can be implicit or explicit like DC. However, one size does not fit all. The schema may not be complete or complex enough to match the requirements of all the activities the metadata object may be used. As a result, a metadata quality problem may occur at least at two different levels: (1) macro (schema), (2) micro (component) [21,27].To measure the quality gap we need to identify the specific information quality requirements for the activity on multiple dimensions and at multiple levels, measure the actual information quality value for the metadata objects on those dimensions, and compare them. Note that examining a metadata object and its schema alone may not be sufficient for assessing its quality on some dimensions, since all IQ dimensions except *Intrinsic* dimensions involve aspects external to a metadata object under assessment. These dimensions require analysis of the relationships and states of related objects and socio-cultural structures of user communities, and most importantly information activities of those communities. For anchoring the quality assessments of all three categories of dimensions we need to establish benchmarks/baseline representations meeting the minimum requirements of the activities and then produce an aggregate quality estimate for quality-based ranking. In [8] we proposed using activity theory [13] and information/activity genres [17] to establish generic, socially-

justified quality & value structures for different types of information objects, and using them for constructing context-sensitive baseline representations and quality requirements. In addition, for more effective reasoning about the action-specific metadata quality requirements and tradeoffs we propose the use of quality curves adapted from Taguchi's value curves [23]. In particular, we use 3 types of quality function curves. A Smaller is Better (SIB) curve is used when a decrease in certain quality dimension increases an aggregate quality of a metadata object. A Larger is Better (LIB) curve models a relationship when a larger value of the quality dimension leads to higher aggregate quality of the object. And, finally, a Nominal is Best (NIB) curve is used when the aggregate quality value for the object is highest when the quality dimension value is between too small and too large. Based on the type of a metadata object and the typified context of its use one can build a curve for each quality dimension at both macro and micro levels and define its critical and nominal values. For instance the general-case critical values of completeness for a Simple DC record can be set to 0 and 15 (the minimum and maximum numbers of distinct fields in the DC schema), while the nominal value may vary based on the types of activities the object is used in.

In summary then, a procedure for measuring metadata quality could be the following: (1) Analyze information creation and use activities: what actions are carried out within a given activity, what types of information tools, including metadata objects, are needed to provide the necessary functionalities for successfully accomplishing these actions; (2) Based on this analysis, identify macro (schema) and micro (component) functional requirements for a given type of metadata objects; (3) Based on the macro and micro requirements, construct a baseline data model; (4) Use the baseline model for measuring the quality of related metadata objects and identify quality problems; (5) Based on the measurement in the previous step assign quality rankings to the objects.

For effective metadata quality assurance, however, we also need to (6) Decompose the macro level quality problems identified in the step #4 into micro level problems; (7) Measure the value & cost changes of quality improvement or degradation along a given quality dimension (value loss, user cost, lost opportunity cost, cost saving, etc); (8) Optimize quality assurance activities based on the criticality of metadata to the activity structure and resources available.

2.2.1 Modeling metadata object quality requirements by analyzing information use activity

To develop a model of the relationship between information quality of a metadata object and the information activities that comprise its context of use, we will consider a typified activity of information discovery: finding an information entity or FIND. We have chosen FIND as a representative type from the set of activities metadata is intended to support, as indicated in [11][5]. While locating information can comprise several basic actions (e.g., select, locate, and obtain, in addition to find), for space reasons we only analyze the relationship between FIND and its quality requirements in this paper. In addition, while there are many metadata schemes and formats in use, for this paper we limit ourselves to DC metadata since it is a widely used standard for documents in electronic repositories and is the basis for the federated collections in our extended case study below.

FIND: finding an information entity: As with the other information activities not treated here, we conceptualize FIND as a set of heuristic actions in a space of possibilities. Note that FIND is an information access problem with at least two types of agents involved: information providers and information seekers. Each of these may have different metadata models, and the level of correspondence among them is an issue [4]. The FIND process involves a collection of moves that includes (possibly in repetition and in varying order) the following: a) a seeker establishes a space of possibilities to be investigated, initial criteria for success, and resource (cost) bounds; b) the seeker (heuristically) generates descriptions of a set of candidate objects from this space using its own (sometimes ad-hoc and uncertain)

[5] By "typified" we mean activities that are cognitively, culturally and/or socially generic to some process, in this case information seeking and use as defined in [11] (and elsewhere). Our general approach is to use well-understood information-seeking use models as the foundation for identifying generic activities to be used in specifying metadata needs.

metadata (e.g. structured or unstructured query terms); c) the provider uses its own procedures (e.g., search routines) to provide its own corresponding metadata (e.g., search results) for the candidate set; d) the seeker uses the provided metadata to comparatively evaluate the candidates, accepting and/or rejecting some of them, possibly accessing the entire source object (document); and e) the seeker uses information from accepted or rejected candidates to reformulate the possibility space, its own metadata, the criteria for candidate generation, and the criteria for cost and success. The process terminates when the information seeker's final success criteria and/or cost bounds are met. Note that most of the process involves search, extraction, analysis, comparison of and inference over *metadata*, rather than over actual target documents.

Metadata has three purposes in FIND. First, an object's metadata is a surrogate or model (a representation) of that object in the possibility space. For the seeker, this metadata model may necessarily be unknown, guessed, or intelligently inferred. Ideally, metadata reduces the total amount of information that must be handled (model vs. full document) in the activity of FIND. Second, metadata limits and focuses attention to specific relevant, modeled, attributes of the target object. This focus can make assessment more efficient, and can focus revisions in step e), but if inaccurate, or does not match the complexity of a task under consideration, it may also degrade quality. Third, the typified or standardized aspect of metadata makes comparison of multiple candidates (e.g., across federated repositories) more efficient.

One key idea underlying our approach here is that information quality becomes critical (only) to the degree it affects the quality of activity *outcomes* – e.g. the quality of materials generated through the FIND activity, in this example. Metadata quality impacts FIND when any of the three metadata purposes given above is enhanced or compromised by metadata quality. This is true at the level of the composite metadata record taken as a whole, as well as at the level of each metadata component (e.g., Title, Creator, Identifier, etc. in DC). These impacts can be illustrated with the following suggestive examples, which are limited for space reasons:

Intrinsic

Automated generation, assessment, or comparison of candidates can fail if the metadata lacks *consistency*. Representing similar objects with the same metadata structure and vocabulary decreases variance in matching and leads to the reduction in the number of false positives and false negatives during search. The curves of semantic, structural and vocabulary consistency dimensions are of a LIB type. In general we can estimate that the more current a metadata object is, the more accurately it reflects the current state of a given information item. Consequently, for producing a partial ordering along the *currency* dimension, while all other variables remain fixed, one may not need to examine the original object or know its properties assuming that the most recent metadata object is of the highest quality. Both the currency and consistency dimension curves are of a LIB type.

Relational/contextual

The information seeker's metadata model of information entities may not match the system or provider metadata model. The provider metadata schema or a single metadata record may not be *complete* enough to locate entities by the attributes the user intends to use. If the author name or subject word are missing in a metadata object and the user uses these attributes in search, the object will not be found and included in the set of relevant entities. A critical minimum point for the completeness dimension would be a data model containing no elements or attributes. A DC object with no elements filled is useless. The other extreme, theoretically, is an object containing an infinite number of elements which will not be useful either as *completeness* is often in conflict with *simplicity*. A completeness curve may be bounded by a simplicity curve and can be of a NIB type. Community specific nominal (optimal) values of completeness can be established using techniques from qualitative analysis.

High *naturalness* of metadata increases the chances of the overlap between the user and provider data models and consequently the effectiveness of search. For instance, using typified words in the surrogate

title of an image increases the probability of the user choosing the same words in search. Naturalness is culture- and community-specific. Its curve is of a LIB type.

Inaccurate metadata can make an information object "disappear" and not be found in search. In cases when partially accurate metadata is still useful for reducing a search space, the accuracy curve is of a LIB type. Indeed, when correcting all information errors is too expensive or impossible, Web search engines and other information systems take a "coping approach" and develop special procedures/algorithms to utilize partially accurate information [26]. However, some attributes can be less robust or tolerant with regard to inaccuracy than others. For instance, even small inaccuracy can render an identifier element useless. For these types of metadata the accuracy dimension takes only binary type values: accurate or not accurate. The last example, however, points to the important tradeoff which may exist between *redundancy* and *robustness*. In certain cases when achieving high accuracy is too expensive, some redundancy can prove to be useful to increase robustness and reduce chances of search failure. For instance, if metadata is misspelled in one element of a DC record, but repeated in a correct form within the same or a different element, the item still will be found [3]. Therefore, for FIND the redundancy and robustness curves are of a NIB type. Increase in *volatility* increases the chance of metadata become invalid. For instance, 9 digit social security numbers (SSN) may not be valid identifiers once the US population exceeds 1 billion. Similarly, the relatively high probability of a female person changing her last name after marriage or divorce, in comparison to her DNA sequence, makes the DNA sequence a better quality identifier. However, there can be a tradeoff between *volatility* and *simplicity*. A personal name can be less complex to use than a SSN and a SSN itself is less complex than a DNA sequence according to the information theoretic measure of descriptive complexity [7]. Therefore, the volatility dimension curve bounded by the simplicity curve is of a NIB type.

If the metadata object or any of its elements is not accessible for any reasons, it is of no use. There can be a tradeoff, however, between *accessibility* and *certainty*. If an increase in accessibility is achieved at the expense of security reduction, certainty of the information conveyed by metadata may suffer. Therefore, the accessibility curve is of NIB type.

Reputational

Low *certainty* or *believability* of metadata objects can lead to poor identification. The uncertainty about the provenance of a metadata object or its content may reduce its use or the collection as whole. This is especially true for an aggregated metadata collection where two or more records may refer to a same object (redundancy), but conflicting information in the date elements or the author elements may not allow the seeker to establish the correct identity of the object. On the other hand, if the information content conveyed by redundant metadata elements or records is confirmatory or "monotonic", an increase in redundancy may help in increasing certainty. For instance, if there is no contradiction among the duplicate records, certainty or believability of the metadata gets increased. The certainty curve is of a LIB type.

Thus, the quality requirements for FIND may involve multiple quality dimensions. In addition, the interactions and interdependencies between these dimensions may result in significant tradeoffs. The following tradeoffs have been identified for FIND activity: *completeness vs. simplicity*; *robustness vs. simplicity*; *volatility vs. simplicity*, *robustness vs. redundancy* and *accessibility vs. certainty*.

2.3 Value & Cost Sensitive Quality Aspects

In the previous section, using an example of the FIND activity we showed the importance of reasoning about metadata quality needs through a prism of the typified activities and actions in a domain. To come up with realistic quality assurance plans and targets it is essential that one is aware about the activity- and action-specific behavior of the quality dimensions and the tradeoffs among them. In addition, however, it is often necessary to quantify the effects of poor or good quality for justifying or making more effective investments in quality improvement.

In manufacturing, improving a product's quality means finding an optimal mean value for each quality-related attribute of the product and reducing the variance around the mean of each of these attributes. The target values for these attributes are calculated from the product's cost and user value structure. That is, the cost of improving quality of a given system attribute has to be met with an increase of product value, exemplified by an increase in cash flow [6]. In digital library settings monetary metrics may not be directly applicable to metadata objects, however, one still can treat them as products and measure the increase in value from increased quality based on user satisfaction or increased use. Or, alternatively, one can measure the value loss caused by inferior quality [23].

2.3.1 Poor Quality as Value Loss

Consider a metadata object as an organizational asset and for simplicity assume that its value is equal to the cost of its creation measured by cataloging staff time and administrative overhead, even though the real value of the object may vary considerably depending of its context of use. If the average cost of an OAI DC record to an institution is $8, the record is comprised of 4 DC elements and the *identifier* information is inaccurate, then we may say that the value loss due the inaccurate identifier is $8 * 1/4 = $2. Obviously, if the *identifier* is the only element that may allow the user to find the corresponding information object, then the value loss due to poor quality is $8.

Likewise, one can design a function and calculate the total loss from all the quality deviations found in the collection. However, as we argued earlier, nominal (target) values as well as the ranges of tolerance may vary from one type of metadata to another. For instance, the allowable amount of inaccuracy for a DC identifier element can be 0, while a user may still find useful a *description* element with intrinsic quality deficiencies such as spelling errors. Similarly, the target values can be different from one domain to another. For instance, the target value of completeness for metadata objects used for academic library information may not be the same ones used in public libraries due to the different levels of complexity of typified activities carried out in these institutions. If for academic libraries the target number of distinct elements in a Simple DC record is N_d, the target total number of elements N_t and the collection shows the mean deviations (d_d, d_t) from these targets, then we can construct a loss function (F_{ca}) for completeness at the composite (schema) level as follows:

$$F_{ca} = k_a \times (d_d \times d_t)^2 / 2$$

Here k is similar to Taguchi's proportionality constant and can be measured as

$$k_a = (L_d + L_t)/(\Lambda_d + \Lambda_t)^2$$

where L_d and L_t are the amounts of loss associated with the values of schema level completeness being outside of the tolerance limit; $?_d$ and $?_t$ are the tolerance limits of schema completeness with regard of distinct and total number of elements used. Similar functions can be constructed at the component level as well as for different types of organizations by using their domain/type specific values for the above variables.

2.3.2 Quality as Amount of Use

Compounded value and quality of a metadata collection, object or element can be also assessed by the number of transactions performed against it. According to [10] value of information may increase in use. More frequently used metadata objects and metadata elements can become more valuable than less used ones. The amount of use on the other hand is determined by the needs of a user community and by the quality of the metadata. For instance, increasing completeness by adding additional metadata elements may enable new uses and increase the number of transactions. Consequently, the effects of quality changes can be quantified based on the decrease or increase in the number of uses.

The value of a metadata element can be a function of the probability distribution of the operations/transactions using the element. Likewise, the value of a metadata object can be assessed as an aggregated value of its individual element values. The ability of assessing and/or predicting the amount of use of an existing component and/or object, or the number of additional transactions enabled by the addition of a new component, can help in reasoning about quality - value - cost tradeoffs and spending

quality improvement resources more effectively. For instance, the use statistics from Appendix (Table 3) rank *identifier* and *title* as the most highly used elements in the metadata records. Assuming that the providers of these records are at the same time one of the main consumers the metadata and their needs are reflected in the composition of the records, we may argue that *identifier* and *title* are the most valuable elements in this particular aggregated metadata collection. Another important source of establishing the value of metadata objects and elements can be a transaction log of component and object uses by the end users.

2.3.3 Quality as Effectiveness

To complete the picture of metadata quality assessment and reasoning, changes in a metadata object's quality need to be linked to changes in the cost of its generation and use. And, there are certain cost generating factors that can be assessed objectively and linked systematically to the changes in quality. One of the simplest "cost drivers" which can be evaluated automatically is the number of metadata elements used and the length of each element. Creating a metadata object with many elements containing values comprising many bytes of information will require more cataloger time than generating a sparse object with a smaller amount of metadata in it. At the micro or component level the cost of using a DC *description* element in the object will be higher than the cost of using *identifier* or *type* elements (see Appendix, Table 3). In addition, creating metadata of a complex information object or an information object in a foreign language may require extra cognitive effort and time from the cataloger and result in a cost increase. Similarly, on the use side, the cost of using a metadata object can be affected by the number of elements it contains, the lengths of these elements and the complexity of metadata. Complexity or simplicity can be calculated as a normalized number of the words not recognized by a general purpose spellchecker dictionary. Thus, assuming that metadata elements do not convey duplicate information, one of the metrics of the cost sensitive quality assessment or effectiveness of a metadata object can be as follows:

$$E = \sum_{1}^{n} e_i / \sum_{1}^{n} t_i$$

where e_i-s are the elements used; t_i stands for the average times needed to create or evaluate & comprehend the element and n is the total number elements in the object. The formula reflects the completeness vs. simplicity tradeoff discussed previously and suggests that reducing complexity along with maintaining the necessary level of completeness can be a target for cost sensitive quality improvement activities.

2.4 Quality Assurance

2.4.1 Improving quality through process control and preventive quality maintenance

Effective quality planning involves exploring and analyzing the metadata production chain, that is, identifying processes, people, input/output interdependences, and quality controls [2]. [12] lists the following surrogation operations: selection, analysis, descriptive cataloging, subject indexing, classification, abstracting and editing. Modeling and analyzing quality assurance and value creation processes throughout these operations and linking them consistently with the use activities and actions, can help in enactment of appropriate control checks and triggers to prevent certain types of quality problems from occurring and make quality assurance decisions more sound and effective. Methods of metadata quality assurance through process control and preventive quality maintenance deserve their own separate treatment and won't be discussed further here.

2.4.2 What a metadata aggregator can do to improve metadata quality inexpensively

In the federated information environments of interest in our research, metadata from a number of individual sources or collections is gathered together and normalized so it can be used uniformly for analysis and retrieval operations on the distributed collections under federation. Such federated or aggregated metadata repositories can pose quality challenges that may not be typical or pervasive for standalone metadata collections. These problems include: loss of context or loss of information - when the local contextual information from a local collection is lost or mapped ambiguously into a global schema;

relational/representational problems such as "link-rot" and differences in update frequencies; changes in concept definitions and controlled vocabularies; misuse or disuse of standard metadata schema elements; variations in record completeness and use of controlled vocabularies; different domain models and definitions, abstraction levels and representation conventions; duplicate records and overlapping collections. These problems are not entirely unique to federated metadata collections; studies of them have been reported to some degree in the general data quality literature where the focus is on large integrated corporate databases and warehouses with more or less strong centralized attempts at quality control. However, federated metadata collections, such those in the IMLS/DCC project, rely heavily on volunteer cooperation among multiple metadata providers who are affected by local cultural, economic, and knowledge factors.

The quality of a metadata object is a product of the "quality assurance chain" which goes in parallel with the information production, supply and use chain [30]. The scope of an aggregator's quality assurance activities is usually limited by its role in this process of division of labor, and the quality of an aggregated collection is largely determined by the metadata quality supplied by individual data providers. Nonetheless, the aggregator, informed by information seeking behavior studies, can still perform a number of actions to improve the metadata quality inexpensively.

The aggregator may not have leverages to influence the metadata creation processes of individual providers, but it still can influence the ways metadata is selected, aggregated, and presented. Earlier in this section we identified 3 models of how the effects of quality changes can be connected to value changes. The aggregator can use these models to improve the value of the aggregated collection. Specifically, the aggregator can (1) reduce the value loss due to poor quality; (2) increase the collection value by increasing the number of its uses and (3) increase the effectiveness of metadata objects.
The aggregator can significantly reduce value loss by simple data scrubbing operations such as spellchecking or duplicate detection, which often can be done automatically. In addition, collection statistical profiles and data mining tools can be used for identifying and correcting misplaced or invalid metadata entries inexpensively.

The aggregator can increase the pool of metadata users by generating supplemental metadata and enabling additional perspectives and contexts of uses. One can visualize the value of a metadata object as the sum of its values to local and global users: $V_m = V_l + V_g$. Often, individual data providers create metadata objects with only local users in mind and the presence of some shared, *a-priori* knowledge K is assumed due to their membership in the community. As a result the utility of including a metadata element E_i containing the same K or a subset of K can be zero or even negative if the cost of the inclusion of E_i is positive [20]: $U(E_i)$= Value (E_i | K) – Cost (E_i). If Cost(E_i) > 0 and Value(E_i | K) = 0 then $U(E_i)<0$. For global users, who are not the members of the particular metadata-generating community and may not share the community knowledge/information, Value(E_i | K) can be positive. However, since marginal information users for the local data provider are the local users, the local provider might lack incentives to improve quality that only benefit the global users and not the local ones, even in a short run. As a result, local resource providers may under-provide quality necessary for attracting non-local users. Therefore, the challenge the aggregator may face is to increase the value of a metadata object for the non-local users by inexpensive means and without decreasing its value to the local users.

Whenever possible, collection statistics and structural regularities of metadata objects can be used by the aggregator to infer preexisting knowledge K, which was not included in metadata objects intentionally by local providers or was lost in conversion between different metadata standards. In that way the aggregator can mitigate some of the clarity/context loss problems mentioned earlier. The missing or new metadata can be also obtained from the original objects, if these objects are available to the aggregator and the metadata can be generated/extracted inexpensively, that is, automatically. Furthermore, the collection statistics can be used for generating or supplementing the metametadata on individual collections, which combined with typified scenarios of metadata uses can be a substantial added value to the existing user communities, enable new uses and bring additional users to the aggregated collection.

The aggregator can increase metadata effectiveness by reducing unnecessary complexity or hide it and make it a lesser burden to the user by offering the view of the metadata that matches the user's cognitive capabilities and level of experience. At the information discovery activity level, one of the indirect measures of the quality of a metadata artifact can be average time the user spends to finish all the actions in the activity. Each metadata element can perform more than one function and each information action can be accomplished by utilizing more than one metadata element. However, some metadata elements are optimized towards certain functionalities and the cost of element use can be also different. Studies of information seeking behavior show that information seeking action trajectories and the choices of metadata elements for achieving a same goal may vary from one type of user to another [22]. It has been found that novice users used fewer subject keywords and did more evaluation and comparison than expert users when searching for relevant articles [14,22]. Consequently, some users may utilize a subject element more extensively while others may rely more on a description element for sensemaking and comparison.

The different types of users choosing different metadata elements to accomplish the same goal can be explained by the different cognitive cost structures they may have for the same metadata element. For instance, the keyword vocabulary of an aggregated repository comprised of specialized vocabularies of the individual collections can be quite large and diverse (see Table 1). To memorize and use this vocabulary can require a significant cognitive effort (or incur a significant cost) from users doing cross-disciplinary research [19] or the novice users. Abstraction and reduction of complexity may help in reducing this effort by better aligning the user and the system models of information objects [5]. However, it may come at the expense of a metadata object losing its discriminatory power and become less effective in reducing the search space. Certain techniques can make the tradeoff less crisp. For instance, by adding thesauri and ontology mappings, the aggregator can reduce the complexity of metadata and make it easier to use for novice users, while at the same time retaining the discriminatory power of domain specific subject keywords, which can be enjoyed by expert searchers or subject experts.

Element	Total #	Unique #	Entropy
identifier	205,719	184,769	0.98
title	133,108	87,689	0.88
subject	304,661	80,702	0.71
source	29,537	11,008	0.68
description	153,088	59,523	0.67
creator	84,829	18,385	0.65
date	189,661	11,068	0.62
coverage	12,103	1,738	0.59
contributor	16,813	2,882	0.54
relation	80,629	3,115	0.35
publisher	114,305	3,347	0.35
rights	68,228	341	0.33
type	124,853	191	0.15
format	111,647	2,308	0.13
language	85,397	95	0.10

Table 1. DC elements ordered by their average information content or uncertainty (entropy) culculated from a collection of 154,782 OAI DC records, not normalized (The standardized entropies are calculated as follows $entropy = -\sum_{i=1}^{n} p_i Log(p_i) / Log(N)$ where p_i is the probability of the i-th unique value for a given element and n is the total number of the unique values for that element).

Finally, users can play an important role in the aggregator's quality improvement activities. Adding user feedback/error reporting capabilities to the system can be an effective and efficient tool in increasing the collection's quality inexpensively by communicating the user dynamic quality needs to the aggregator

[26]. While the producer/provider model of metadata quality can be more process oriented, the user's perception of quality tends to be more product experience oriented mainly focusing on usability [21] – that is, whether a given information object and/or a system as whole is usable or not. Harnessing the user community power for error reporting can help in aligning these two models better by (1) identifying metadata creation process errors, and (2) identifying user metadata quality needs at both composite and component levels. The aggregator can serve both as a quality improvement agent and as an intermediary by forwarding relevant feedback to an appropriate metadata provider [30].

3. CASE OF IMLS DDC AGGREGATED METADATA COLLECTION

In this section we present some of the results of statistical analysis of random samples from an aggregated collection of OAI Simple DC records harvested as a part of IMLS DCC project, and interpret them using the concepts and procedures proposed in the earlier sections.

At the time of the analysis the collection contained a total of 154,782 records that had been harvested from 16 OAI data and service providers – academic and public libraries, museums and historical societies. A manual inspection of a random sample of 150 OAI Simple DC records identified six major types of quality problems: (1) lack of completeness, (2) redundant metadata; (3) lack of clarity; (4) incorrect use of DC schema elements or semantic inconsistency; (5) structural inconsistency and (6) inaccurate representation (see Table 2). All of the 150 examined records were incomplete. None of them used all 15 DC elements. 94% of the records contained elements with duplicate metadata. In addition, most of the date elements were ambiguous. Removed from the context, it was not clear to what date the information referred to: was it the date when a given photo was taken, donated, microfilmed or digitized. Incorrect use of the DC elements was also common. The most frequent misuse (or a possible workaround [9]) was putting date information in a source element or using a description element for format information. Almost a half of the sample had consistency problems. Using different date formats was the most common problem as well as inconsistent structuring of records and the lack of using controlled vocabularies. And, finally, 24% of the records had representational problems in forms of broken identifier links and partially inaccurate representation of original object content.

Problem type	Incomplete	Redundant	Unclear	Incorrect Use of Elements	Inconsistent	Inaccurate Representation
%	100	94	78	73	47	24

Table 2.. Metadata quality problems

The results of statistical analysis corroborates to the existence of quality tradeoffs mentioned in the section #2. The analysis of a random sample of IMLS OAI DC records shows a strong negative correlation between Log(number of distinct DC elements per record) and simplicity (sample size =1,000 records, population size = 154,782, the Spearman correlation coefficient -.684, significant at the 0.01 level, two tailed). Simplicity, in this case, is calculated as 1 minus a ratio of the number of the words not recognized by MS Word spellchecker dictionary over the size of a record. Thus, there can be a significant tradeoff between completeness and simplicity/complexity.

[1] suggests that the size and complexity of metadata can be inversely related to its quality. Indeed, our analysis of the random sample of 150 OAI Simple DC records too shows a significant negative correlation between an aggregate quality problem rate, which is a simple normalized linear combination of the above mentioned quality problem category scores for a given metadata object divided by the object's length, and a simplicity score (Spearman correlation coefficient -.434, significant at the 0.01 level, two tailed). To our surprise, however, no significant correlation has been found between the aggregate quality problem rate and the normalized length of a metadata object (Spearman correlation coefficient .043, not significant at the 0.01 level, two tailed). Interestingly enough, similar observations have been made in some of the software quality research projects. For instance, [25] found a strong

correlation between the cyclomatic software complexity index and the number of defects/bugs in software. Although intuitive, the above observations once again emphasize the importance of reasoning about the quality tradeoffs. Generating completeness - simplicity and simplicity - quality problem tradeoff curves may help in predicting quality of metadata objects as well determining provider specific optimal levels of their completeness and complexity.

We mentioned earlier that the nominal or target values of a given quality dimension can vary from one domain to another and from type of metadata to another. Indeed, our analysis of a 2,000-record random sample from the aggregated collection shows that the number of distinct elements used in the records varies from a minimum of 2 elements to the maximum 14 with the mean equal to 7.62 and the standard deviation equal to 2.93. The total number of elements used per record also varies from the minimum 2 to the maximum 82 with the standard deviation 5.73 (see Appendix, Table 4). The results of the statistical analysis of the data also identified a significant correlation between the consistency of element use and the type of metadata objects and the type of data providers. Indeed, grouping by type of the same metadata records made the standard deviation of the total number of elements used to drop significantly (from 5.73 to 3.6), even for binary type values: photos and non-photos. Likewise, grouping the sample by the type of providers (public libraries, academic libraries, and museums) reduced the variance of the total number elements used per a metadata record from 5.73 to 4.5. Not surprisingly, academic institutions on average use more distinct elements per record: 13 versus 8 used by public libraries. The metadata records generated by academic libraries on average are larger in size than the records generated by the other types of institutions. Indeed, clustering by the use of distinct DC elements, using the K-means clustering technique with 2 clusters, almost perfectly discriminates the public library records from the academic library records. Most of the public library records go into a cluster #1 with a center of 8 distinct elements (title, subject, description, publisher, date, type, identifier, rights) while the academic library records are placed in a cluster #2 with a center of 13 elements (title, creator, subject, description, publisher, date, type, format, identifier, source, language, relation, rights). Museum records in our sample are split almost equally between these two clusters. Thus, implicitly or explicitly, public and academic libraries may use different subsets of the DC schema when generating metadata records based on the needs of their marginal users and cost structures. A realistic loss function needs to reflect these differences and use organization type specific baseline values. For instance, if for academic libraries the target number of distinct elements used in a DC record is 13, the target total number of elements 16 and the collection shows mean deviations (d_d, d_t) from these targets - 1.7 and 4.6 respectively - then we can construct a loss function (F_{ca}) for completeness as follows: $F_{ca} = k_a \times (d_{da} \times d_{ta})^2 / 2 = k_a \times (1.7 \times 4.6)^2 / 2 \approx 30.6 \times k_a$ where k and N are defined in the same way as earlier. If bringing distinct element and total element completeness up to the nominal values at academic libraries costs on average $8 per record and the tolerance limits are 10 and 13, then $k_a = (8+8)/(10+13)^2 \approx 0.03$ and $F_{ca} = 30.6 \times k_a = 30.6 \times 0.03 = \$0.918$ per record. Thus, if we assume that 100% of the records are incomplete and there are 50,000 records from academic libraries, then the value loss to the collection due to incomplete records can be estimated as $45,900.

Inexpensive data quality improvement actions such as date, type and language element normalization can substantially reduce the element variance (noise) and make metadata object evaluation and comparison easier [5]. Indeed, we found that simple spellchecking and normalization can reduce the variance of these elements more than in half. In addition, to further reduce cognitive cost of using the metadata records, we experiment with the use of subject thesauri allowing automatic query expansion during search.

4. CONCLUSION

In this research we studied the problems of quality assurance in large federated collections of metadata objects. Random samples from the collection of 154,782 OAI Simple DC records have been examined to identify the types and severity of metadata quality problems. A general framework of information quality assessment proposed earlier in [8] has being used for developing metadata quality dimensions and metrics

and linking them consistently with the quality problems encountered in typified information seeking and processing activities.

We found that developing an inventory of the mappings among metadata elements, metadata creation and use activities, quality dimensions and tradeoffs can allow more robust reasoning about metadata quality and apply limited quality improvement resources more effectively.

In addition, the results of the statistical analysis of the data has shown significant correlation between some of the quality dimensions such as completeness and the type of metadata objects and the type of data providers. However, whether these regularities are the results of the data providers using implicit genre/type based schemas when generating metadata objects or the byproducts of particular cataloging software conventions are open to question.

In future research we will also focus on user valuations of metadata quality. We intend to use focus groups and surveys for establishing the quality & value structures for different types of metadata objects and developing functions for measuring and translating objectively the changes of a metadata object quality into the changes of its value and cost.

REFERENCES

[1] Bade, D. (2002). The creation and persistence of misinformation in shared library catalogs: language and subject knowledge in a technological era: Vol. 211. GSLIS Occasional Papers. Champaign, IL: GSLIS, University of Illinois.

[2] Ballou, D., Wang, R., Pazer, H., Tayi, G. (1998). Modeling Information Manufacturing Systems to Determine Information Product Quality. Management Science, 44(4), 462-484.

[3] Basch, R. (1995). Introduction: An overview of quality and value in information services. In: R. Basch (Ed.), Electronic information delivery. (pp. 1-13). , Brookfield, VE: Gower.

[4] Buckland, M. (1999). Vocabulary as a central concept in library and information science. In: Proceedings of the Third International Conference on Conceptions of Library and Information Science . Dubrovnik, Croatia.

[5] Cole, T., Kaczmarek, J., Marty, P., Prom, C., Sandore, B., Shreeves, S. (2002). Now that we've found the "hidden web," what can we do with it? The Illinois Open Archives Initiative Metadata Harvesting Experience. In: Proceedings of the Museums and the Web 2002. 63-72.

[6] Cook, H. (1997). Product management: Value, Quality, Cost, Price, Profits, and Organization. Amsterdam, Netherlands: Chapman & Hall.

[7] Cover, T., Thomas, J. (1991). Elements of Information Theory. New York, NY: Wiley.

[8] Gasser, L., Stvilia, B. (2001). A New Framework for Information Quality. Technical Report ISRN UIUCLIS--2001/1+AMAS. Champaign, IL: University of Illinois at Urbana Champaign.

[9] Gasser, L. (1986). The integration of computing and routine work. ACM Transactions on Office Information Systems, 4(3), 225-250.

[10] Glazer, R. (1993). Measuring the value of information: The information - intensive organization. IBM Systems Journal, 32(1), 99.

[11] IFLA Study Group on the Functional Requirements for Bibliographic Records (1998). Functional requirements for bibliographic records : final report. München : K.G. Saur. Available as: www.ifla.org/VII/s13/frbr/frbr.pdf or as html: www.ifla.org/VII/s13/frbr/frbr.htm

[12] Landau, H. (1969). The cost analysis of document surrogation: A literature review. American Documentation, 20(4), 302-310.

[13] Leontiev, A. (1978). Activity, Consciousness, Personality. Englewood Cliffs, NJ: Prentice Hall.

[14] Marchionini, G., Lin, X., Dwiggins, S. (1990). Effects of search and subject expertise on information seeking in a hypertext environment. In: Proceedings of the 53rd Annual Meeting of the American Society for Information Science. Toronto, Canada 129-142.

[15] Motro, A., Anokhin, P., Acar, A. (2004). Utility-based Resolution of Data Inconsistencies . In: Proceedings of the SIGMOD IQIS 2004 Workshop. 35-43.

[16] Naumann, F., Leser, U., Freytag, J. C. (1999). Quality-Driven Integration of Heterogeneous Information Systems. In: Proceedings of the International Conference on Very Large Databases. Edinburgh, UK..

[17] Orlikowski, W., Yates, J. (1994). Genre Repertoire: The Structuring of Communicative Practices in Organizations. Administrative Science Quarterly, 39, 541 - 574.
[18] Orr, K. (1998). Data quality and systems theory. Communications of the ACM, 41(2), 66-71.
[19] Palmer, C. (1998). Structures and Strategies of Interdisciplinary Science. Journal of the American Society for Information Science , 50(3), 242-253.
[20] Radner, R., Stiglitz, J. (1984). A nonconcavity in the value of information. In: M. Boyer, R. Kihlstrom (Eds.), Bayesian models in economic theory. (pp. 33-52). , New York, NY: Elsevier.
[21] Strong, D., Lee, Y., Wang, R. (1997). Data Quality in Context. Communications of the ACM, 40(5), 103-110.
[22] Sutcliffe, A., Ennis, M., Watkinson, S. (2000). Empirical studies of end-user information seeking. Journal of the American Society for Information Science, 51(13), 1211-1231.
[23] Taguchi, G., Elsayed, E., Hsiang, T. (1989). Quality engineering in production systems. McGraw-Hill.
[24] Taylor, R. (1986). Value-added processes in information systems. Norwood, NJ: Ablex Publishing.
[25] Troster, J. (1992). Assessing design-quality metrics on legacy software. In: Proceedings of Proceedings of the 1992 conference of the IBM Centre for Advanced Studies on Collaborative research. 113-131.
[26] Twidale, M., Marty, P. (1999). Investigation of data quality and collaboration. Champaign, IL: GSLIS, University of Illinois at Urbana - Champaign.
[27] Wand, Y., Wang, R. (1996). Anchoring data quality dimensions in ontological foundations. Communications of the ACM, 39(11), 86-95.
[28] Wang, R., Allen, T., Harris, W., Madnick, S. (2003). An information product approach for total information awareness. In: Proceedings of IEEE Aerospace Conference.
[29] Wang, R., Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12(4), 5-35.
[30] Williams, M. (1990). Highlights of the online database industry - the quality of information and data. In: Proceedings of the National Online Meeting. Medford, New Jersey.

APPENDIX:

Dublin Core element	% of repositories 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	100.0	124,304	133,108	80.3	1.1	39.9	1	75.8
Creator	87.5	78,402	84,829	50.7	1.1	21.5	0	49.3
Subject	93.8	112,875	304,661	72.9	2.7	110.4	2	37.1
Description	81.3	73,298	153,088	47.4	2.1	104.1	0	52.6
Publisher	75.0	94,791	114,305	61.2	1.2	38.5	1	50.9
Contributor	62.5	10,158	16,813	6.6	1.7	47.0	0	93.4
Date	81.3	66,514	77,175	43.0	1.2	10.9	0	57.0
Type	81.3	118,419	124,853	76.5	1.1	6.6	1	72.5
Format	56.3	107,381	111,647	69.4	1.0	8.3	1	66.6
Identifier	100.0	154,113	205,719	99.6	1.3	84.4	1	71.5
Source	50.0	23,012	29,537	14.9	1.3	68.3	0	85.1
Language	75.0	85,201	85,397	55.0	1.0	3.3	1	54.9
Relation	43.8	48,356	80,629	31.2	1.7	95.6	0	68.8
Coverage	37.5	9,136	12,103	5.9	1.3	21.0	0	94.1
Rights	62.5	63,435	68,228	41.0	1.1	151.7	0	59.0

Table 3. Use of Dublin Core elements (16 providers, 154,782 records)

	N	Minimum	Maximum	Mean	Std. Deviation	Mode	Mode Frequency
Total number of elements per record	2000	2	82	10.45	5.73	9	0.2
Number of distinct elements per record	2000	2	14	7.62	2.93	7	0.3

Table 4. Descriptive Statistics (a random sample of 2000 records from the population of 154,782 records)