A Conceptual Framework for Metadata Quality Assessment

Thomas Margaritopoulos
University of Macedonia, Greece
margatom@uom.gr

Merkourios Margaritopoulos
University of Macedonia, Greece
mermar@uom.gr

Ioannis Mavridis
University of Macedonia, Greece
mavridis@uom.gr

Athanasios Manitsaris
University of Macedonia, Greece
manits@uom.gr

Abstract

Metadata quality of digital resources in a repository is an issue directly associated with the repository's efficiency and value. In this paper, the subject of metadata quality is approached by introducing a new conceptual framework that defines it in terms of its fundamental components. Additionally, a method for assessing these components by exploiting structural and semantic relations among the resources is presented. These relations can be used to generate implied logic rules, which include, impose or prohibit certain values in the fields of a metadata record. The use of such rules can serve as a tool for conducting quality control in the records, in order to diagnose deficiencies and errors.

Keywords: digital repositories; metadata quality; related resources; logic rules

1. Introduction

The quality of metadata describing digital resources stored in a repository can be considered as a necessary condition for reliable and efficient operation of the repository. Metadata is considered to be the key to successfully discovering the appropriate resources. Therefore, metadata must be created and maintained according to well-defined procedures. This requirement is more important considering the vast number of available digital resources, which keeps on growing with rapid rates. Even though the requirement for quality metadata has been generally recognized, there isn't any commonly accepted approach on the definition of metadata quality, and, as a consequence, on the ways this quality can be assessed, measured and increased.

Studies conducted on the subject, represent research efforts to compute statistical indices (Najjar, Ternier & Duval, 2003; Friesen, 2004; Bui & Park, 2006), define frameworks (Moen, Stewart & McClure, 1997; Gasser & Stvilia, 2001; Bruce & Hillman, 2004), identify quality characteristics and detect quality problems (Dushay & Hillman, 2003), either directly or indirectly (by locating indicators of quality). The diversity and complexity of the proposed parameters or characteristics of metadata quality brings out the obvious need to return back to the basics and talk about the roots of the issue of quality and its fundamental components. A conceptual framework to define metadata quality by using analogies from common knowledge and experience is among the goals of this paper.

Moreover, an important conclusion drawn from studying relevant research efforts is that the majority of them assess quality of a metadata record or a metadata repository based on the syntactical level of the content and the metadata standard, but not on the semantical level. A potential source of semantical level information could be any possible interdependencies connecting the resources. Digital resources stored in a repository are not completely independent from each other; they are connected with structural or semantical relations. Especially, in digital resources constituting assemblies (like educational resources registered in a repository as collections, e.g. SCORM), or aggregations (e.g. a web page containing an image and an animation) these relations among the resources create a net of interdependencies, which affect

https://doi.org/10.23106/dcmi.952109222

their metadata records, accordingly. These interdependencies are expressed as logic rules the validity of which influences metadata quality and will be dealt with in this paper.

The rest of the paper is structured as follows: In Section 2, a literature review on the related work on the general subject of metadata quality, along with the subject of logic rules connecting metadata of related resources is conducted. In Section 3, a conceptual framework of metadata quality originating from an intuitive and empirical metaphor is proposed. Based on the framework introduced in Section 3, Section 4 presents a method of metadata quality assessment that uses logic rules connecting related resources. Section 5 provides application examples on the way such rules can be used to assess metadata quality. Finally, Section 6 draws conclusions and points out issues for future work.

2. Related Work

The related work presented in this section concerns different fields of study, which are combined for the purpose of the proposed approach; the field of metadata quality and the field of logic rules involving metadata of related resources.

In the past, several research efforts related with metadata quality have been conducted. These efforts approach the subject from diverse perspectives, trying to cover most of its different aspects. Najjar, Ternier & Duval (2003), Friesen (2004) and Bui & ran Park (2006) conduct a statistical analysis on a sample of metadata records from various repositories and evaluate the usage of the standard. They designate the most frequently used fields and values attributed to these fields. While not directly associated with quality, the statistical indices produced provide an insight of the efficiency of the repositories examined. In this regard, (Greenberg et al., 2001) reports on a study that examined the ability of resource authors to create acceptable – quality metadata in an organizational setting using manual evaluation by experts. Dushay & Hillman (2003) studies the issue of quality by pinpointing deficiencies that degrade it. In the same work, the use of a graphical tool to visualize the deficiencies in a repository level is also proposed. The issue of quality assurance is treated in (Barton, Currier & Hey, 2003; Guy, Powell & Day, 2004; Currier et al., 2004) and general principles and guidelines for the creation of metadata, in order to meet the functional requirements of the application in which they are used, are provided. In the context of quality assurance, (Hillman & Phipps, 2007) discusses the contribution of application profiles as a means for exposing and enforcing metadata quality.

A more systematic and organized view of metadata quality is achieved with the introduction of generic frameworks for the evaluation of quality. In (Moen, Stewart & McClure, 1997) a procedural framework for evaluating metadata records is introduced, using a set of 23 evaluation criteria. The framework discoursed in (Gasser & Stvilia, 2001) is based on concepts and ideas of the more generic field of information quality. It identifies 32 information quality parameters classified into 3 dimensions: intrinsic, relational/contextual and reputational. (Bruce & Hillman, 2004) elaborates on 7 characteristics of metadata quality: completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness and accessibility. Using (Bruce & Hillman, 2004) as a theoretical background, (Ochoa & Duval, 2006) attempts to operationalize the measurement of quality in a set of automatically calculated metrics for the 7 parameters. Similar efforts to provide metrics for metadata quality parameters can be found in (Hughes, 2004).

Focusing on the field of logic rules connecting metadata of related resources, the review of the related literature does not reveal any attempt to use such rules as a means to evaluate metadata quality. However, they have been used for automatic metadata generation. Duval & Hodgins (2004) points out that a resource's metadata may derive from the metadata of related resources. Hatala & Richards (2003) refers to resources being parts of a collection. In this case, it is possible that these resources share common values in their metadata elements. Although the resources in the collection and their metadata records are distinct, a value set for one metadata element in one resource can propagate itself to other resources of the collection. If the assembly is organized

https://doi.org/10.23106/dcmi.952109222

hierarchically, some of the values can be inherited from the ancestor nodes or aggregated from the child nodes. In other cases, the relations connecting the resources may not be such that the metadata value propagates as it is, but the value may be the result of a mathematical or logic expression of metadata of the related resources that either imposes a certain value, or restricts the range of values. Research efforts that use logic (or inference) rules for automatic metadata generation, either explicitly, or implicitly, are included in (Bourda, Doan & Kekhia, 2002; Brase, Painter & Nejdl, 2003; Doan & Bourda, 2005; Motelet, 2005; Margaritopoulos, Manitsaris & Mavridis, 2007). These efforts make use of the LOM metadata schema (IEEE, 2002).

Based on the background of the related work, this paper proceeds to define a new conceptual framework for metadata quality and a method for its assessment that exploits logic rules expressing interdependencies of the metadata.

3. The concept of metadata quality (the court metaphor)

The purpose of metadata is to provide adequate and correct information to their user so as to obtain a true picture of the content of a resource without having to access it. Any effort to approach metadata quality must always take this purpose into account. Metadata serve as the "mirror" of the resource, therefore their quality expresses the true representation of the resource and the absence of any distortion of its picture.

In order to approach the concept of quality, we can make use of a highly intuitive metaphor from a court of law. The metaphor defines a conceptual framework which can serve as a theoretical background to support the study of metadata quality. If we represent the resources of a repository with the facts of a case in court, the assessment of the quality of metadata is a process parallel to the evaluation of the descriptions of the facts of the case, as they are testified by the witnesses (with the assumption that for every fact there is only one witness). The (one and only) metadata record describing a resource in the repository is represented by the description of a fact by a witness (his/her testimony). The testimony of the witness comprises a set of single statements for every different aspect of the fact described. These statements represent the fields of the metadata record.

The issue of defining the quality of the metadata of a resource can be approached by using the abstract of the oath a witness takes in the court when he/she swears to "…tell the truth, the whole truth and nothing but the truth…" for the case he/she testifies. The quality of the testimony is assessed from its distance from the true fact ("truth" – correctness of the testimony), the inclusion of all the possible aspects of the fact ("whole truth" – completeness of the testimony) and the relation of the testimony with the case under examination ("nothing but the truth" – relevance of the testimony). The representation of the resources in a repository with the facts of a case in court and the metadata describing the resources with the witnesses' testimonies, leads to defining metadata quality as the resultant of their correctness, completeness and relevance.

The correctness of metadata refers to the intellectual distance separating them from the true representation of the resource being described. Correctness can be classified into two levels: The first, lower level concerns the requirement that the values of the metadata fields must obey the grammatical and syntactical rules of the language and the metadata standard or the application profile used. Missing letters, misspelled words, inconsistent formatting or representation of the same fields, fields containing inappropriate values according to the standard, are among the problems of this level. A metadata record must strictly follow the rules and guidelines of the standard or the application profile in order to be correct, just like a witness must be able to properly use the language to communicate in order to set his/her testimony fully understandable and, thus, allow the jury to form an opinion on its truthfulness. The second, higher level of correctness requires the semantical rightness of the values of the metadata fields, that is, the true representation of the reality and the absence of any deception. In court terms, this level refers to the truthfulness of the testimony. The first level of correctness concerns objective information, and for the purposes of this paper it is considered to be resolved, for example, by using any

https://doi.org/10.23106/dcmi.952109222

relevant validation parser. The second level of correctness is more subjective and it is the one that will be dealt with in this paper.

The completeness of metadata refers to their sufficiency to fully describe a resource. In essence, completeness measures the presence or absence of values in the metadata fields. In the court metaphor, completeness of a testimony refers to the adequate coverage of all the aspects of the fact described by a witness and to the provision of answers to all the questions he/she is asked. The choice of questions addressed to the witnesses is a task performed by the judge. The choice of the metadata fields – where values are to be filled in, in order for the record to be considered complete – is a matter of the requirements of any given application. Thus, in a given application context, other fields are considered as important and others are not. The application profile plays the role of the judge by selecting certain metadata fields to be accounted as mandatory or optional. A remarkable study on the completeness of a metadata record, focusing on educational resources (learning objects), is included in (Sicilia et al, 2005).

The relevance of the metadata of a resource has to do with its context of use. A metadata record of absolute correctness and full completeness may not be of quality if the (complete and correct) values of the metadata fields do not comply with the context of use. A testimony of a witness in the court, although complete and true, might be irrelevant with the case. This means that the context of a question asked to the witness is incompatible with the context of his/her answer to the question, because of possibly different perspectives. Relevance, as a component of quality, is highly subjective and may be confused with correctness, in the sense that faulty values might be due to either incorrectness, or irrelevance. However, the discriminating factor is the context. An incorrect value is faulty regardless of the context, while an irrelevant value is associated with a particular context. For example, a faulty value for the metadata field "Date of creation" of a resource is a matter of incorrectness (either syntactical – wrong format of the real date of creation, or semantical – a syntactically correct date different from the real one) regardless of any possible context. On the other hand, (although correct) values of the metadata field "Keyword" of a digital photo may be faulty due to irrelevance regarding a given context. If the digital photo has been indexed in a museum of photography, its keywords might be irrelevant when the photo is used in an image processing course. A way to reduce subjectivity and increase the relevance of the metadata is the use of vocabularies of values. A judge in the court restricts the witness's possible answers with the use of similar vocabularies ("…please answer with a yes or no…"). In this logic, a faulty value in a metadata field with a range of values out of a vocabulary will be, more possibly, attributed to incorrectness, rather than irrelevance.

The concept of quality is approached in the proposed conceptual framework by identifying the fundamental components and explicitly stating a solid definition which is domain and method independent. This definition targets the notion of metadata quality, directly. In a different sense, several of the related studies of metadata quality referenced above (Stvilia, 2001; Hillman, 2004; Moen, Stewart, & McClure, 1997) try to locate characteristics of metadata indicating quality or to detect deficiencies indicating its absence. Since no researcher claims to have found an exhaustive list of such characteristics, although this list is necessary to have quality, being not sufficient, it cannot guarantee its existence. Conversely, only if quality exists, all of the proposed characteristics are considered to be present. Such characteristics include parameters or dimensions of quality, like "accuracy", "precision", "naturalness", "informativeness", etc. Some other characteristics constitute signs and trails implying quality and not indicators assessing quality itself. For example, the parameter "provenance" (Bruce & Hillman, 2004) corresponds to the level of reliability and expertise of the metadata record creator. However, although the value of the creator of metadata is a good starting point to assume quality of his/her product, it cannot serve as a proof for quality, for the same reason a testimony of a witness cannot be considered to be true, only because of his/her high social acceptance and respectability. One could say that provenance assesses the probability of having quality in metadata. Other parameters in this category include "timeliness", "currency", "conformance to expectations", "volatility", "authority".

https://doi.org/10.23106/dcmi.952109222

The conceptual framework for metadata quality presented in this section provides the necessary background to support methods and techniques for assessing quality. A method for quality assessment exploiting logic rules that correlate metadata fields and records will be introduced in the next section.

4. A Method for Metadata Quality Assessment Using Logic Rules

Keeping on the court metaphor, one can say that the verdict of the court for the case under examination is based on the assessment of the quality (i.e. correctness, completeness and relevance) of all the witnesses's testimonies. With the already stated assumption that for each fact there is one and only witness account (each resource in the repository is described by only one metadata record), a method for assessing the quality of a testimony is to check for the presence of inconsistencies; on the one hand to check for contradicting descriptions regarding the aspects of the fact and on the other hand to check for contradictions when the testimony is examined in comparison with other testimonies describing related facts. Any such contradictions violate implied logic rules and cause the testimonies to be considered unreliable. Compliance of a testimony with these rules classifies it as reliable.

In this sense, in order for a metadata record to be of quality, it has to comply with similar rules expressing logic dependencies, both among fields inside the record and among fields of records of related resources. A method to assess metadata quality is to check for the validity of logic rules expressing these dependencies.

4.1. Dependencies of Metadata Fields

In some cases, the fields of a metadata record are not completely independent from each other denoting intra-record dependencies. They present some sort of correlation, which is implicitly (if not explicitly) imposed by the specifications of the standard. The degree to which the values of correlated fields inside the record conform to the logic dictated by the relation between the fields is an indication of the record's quality. For example, the fields «1.7 General.Structure» and «1.8 General.Aggregation Level» of LOM are directly interdependent, as it is dictated by the LOM specification (IEEE, 2002), according to which "a learning object with Structure="atomic" will typically have AggregationLevel=1". The violation of this rule indicates degraded quality of the record.

Of course, the existence of relations between the fields of a metadata record indicates a "weakness" of the metadata schema, since, "…an efficient metadata system strives to have as nearly independent dimensions as possible…" (Wason & Wiley, 2001). However, the exclusion of such interdependences between the fields of a record is not always possible; hence, this fact is exploited for the evaluation of quality by examining the existence of certain combinations of values in the related fields inside the record (Ochoa & Duval 2006).

The dependencies of metadata fields are not restricted to fields inside a single record. They may concern fields of records of related resources denoting inter-record dependencies. Resources, related to each other with some kind of relation, create together a whole and therefore, it is possible that several of their metadata fields are influenced by each other. The influence of the values of the metadata fields is done on the basis of logic rules which constitute a set of validation principles that quality metadata fields must conform to. The definition of logic rules is an intellectual task, which has to take into account the semantics of the relations and the metadata. A methodology to create logic rules stemming from relations between metadata fields among records has been proposed in (Margaritopoulos, Manitsaris & Mavridis, 2007) for the purpose of metadata generation. The concepts and ideas presented in this work will serve as the starting point for defining logic rules to be used as validation rules for quality assessment of metadata, in the next subsection.

The core concept in the proposed methodology is the interrelated properties of the resources connected with a relation. These properties are called "connection features" and are specified on

https://doi.org/10.23106/dcmi.952109222

the basis of similarities or differences of the related resources. Connection features may be stated explicitly in the definition of the semantics of a relation. However, in other cases, connection features may be implied. For example, the definition of the semantics of the relation "IsVersionOf" of Dublin Core (DCMI Usage Board, 2008) clearly highlights the connection features "Format" and "Creator", because the related resources have the same format and the same creator, whereas, one can presume the connection feature "Topic area" because different versions of a resource belong to the same topic area. Another example of a connection feature is "Intellectual content" deriving from the relation "IsFormatOf" of Dublin Core, since resources related with this relation have the same content. Apart from relations referring to semantic characteristics of the resources they connect, structural relations (part – whole relations) connecting the related resources are also included in the definition of connection features ("part" or "subset, "whole" or "superset" connection features).

The connection features, thought as properties of resources, can be mapped to certain metadata fields of the schema used for describing the resources. For example, the connection feature "Intellectual content" maps to metadata fields which express concepts and properties of learning objects exclusively influenced by their intellectual content. For the LOM standard, in these fields, «1.2 General.Title», «1.4 General.Description», «1.5 General.Keyword», «5.2 Educational.Learning resource type», «5.4 Educational.Semantic density», «5.6 Educational.Context», «5.7 Educational.Typical age range» are included.

The interrelation of the connection features of two resources (through the relation they are connected with) is translated into the interrelation of their respective metadata fields. These interrelations form a set of logic rules the violation of which indicates metadata records of degraded quality. An example of such rule for the metadata field "1.5 General.Keyword" of LOM can be derived from the connection feature "Intellectual content" of the relation "IsFormatOf". "Intellectual content" feature can be mapped to this field because keywords are determined by the content of an object. The rule can be expressed as "learning objects that differ only in their format (they have the same content), must have the same keywords".

4.2. Quality Assessment Rules

The logic rules, used for assessing quality of metadata utilizing related resources, can be distinguished into three major categories:

- *Rules of Inclusion*: the resource's metadata field values must include the values of the same metadata field of records of related resources. Rules of inclusion apply only on metadata fields with cardinality greater than 1.
- *Rules of Imposition*: the resource's metadata field values must be equal to the result of a mathematical or logic expression of metadata field values of the records of related resources (or of metadata field values of the same record, resulting from intra-record dependencies).
- *Rules of Restriction*: the range of a resource's metadata field values is not the complete value space defined by the specification of the standard used, but a proper subset of it computed from the values of the same metadata field of records of related resources (or of another metadata field of the same record, resulting from intra-record dependencies).Values not belonging to this subset are prohibited.

In order to come up with a complete set of such rules, the semantics of relations connecting the resources and the semantics of metadata have to be taken into account. The rules influence the values of the metadata fields according to the category they belong to. It is obvious that the rules are metadata standard (or application profile) specific. For example, in the LOM standard a rule of inclusion dictates that the field "1.3 General.Language" of a learning object must include the values of the same metadata field of its parts (relation "HasPart"). Additionally, a rule of imposition imposes the value of the field "4.1 Technical.Format" of a learning object to be the same with the corresponding value of another learning object connected to the first one with the

https://doi.org/10.23106/dcmi.952109222

relation "IsVersionOf". Moreover, a rule of restriction restricts the range of values of the field "5.7 Educational.Typical age range" of a learning object to be greater than the maximum typical age range of the objects it "Requires".

A comprehensive list of rules is a matter every community of practice should deal with in the context of the application profile used. An important issue that remains open for consideration is the matter of conflicts. A conflict may come up when the value of the metadata field of a resource is influenced by two or more rules, according to the resource's relations, yielding contradicting values. In this case a conflict policy must be defined.

5. Application of Quality Assessment Logic Rules

Given the definition of quality presented in Section 3, the logic rules deriving from relations among digital resources can be applied to their metadata in order to assess the fundamental components of their quality, i.e. their correctness, completeness and relevance.

For each metadata record in a repository, all the rules affecting the value of metadata fields are applied. Thus, according to whether a rule is valid or not, we infer the following:

- *Validity of a rule of inclusion:* If a rule of inclusion is valid, i.e. the metadata field of a resource under consideration includes values of corresponding metadata fields of its related resources, there is a clear indication of quality of all the involved fields. On the contrary, if such a rule does not hold, it is an indication either of reduced completeness of the field under examination, or of reduced correctness or relevance of its related fields. For example, as stated in the previous Section, in the LOM standard, a rule of inclusion dictates that the field "1.3 General.Language" of a learning object must include the values of the same metadata field of its parts (relation "HasPart"). Examining the validity of this rule for a learning object by comparing the values of its "1.3 General.Language" field against the value, e.g. "en", of the same field of a learning object that is part of the first one, can lead to two results: If the rule holds, that is, the value "en" is included in the values of the field of the learning object under examination, then there is a clear indication of quality of the two involved fields. If the rule does not hold (the English language is not included in the values of the field of the learning object under examination), there are two cases: a) There is an indication of reduced completeness of the field of the first learning object. b) There is an indication of reduced correctness, either on the first, or on the second learning object (or on both). While in this example, concerning field "1.3 General.Language" of LOM, the problem of reduced quality in case b is, clearly, correctness, there might be situations where the faulty values derive from the context, so relevance might be the problematic component of quality.
- *Validity of a rule of imposition:* If a rule of imposition is valid, that is the resource's metadata field values are equal to the result of the mathematical or logic expression of metadata field values of related resources suggested by the rule, then there is a clear indication of quality of all the involved fields. On the contrary, if the rule does not hold, there are two cases corresponding to this: a) The metadata field under examination does not have any value. The absence of value is a matter of reduced completeness. b) The metadata field under examination has a different value than the one dictated by the rule. In the case of a field with cardinality 1, the inequality of its value with the value dictated by the rule is an indication of absence of correctness (or reduced relevance) for the set of the involved fields. If the field under examination is of cardinality greater than 1, then the inequality of its (multiple) value with the value dictated by the rule, implies either completeness or correctness – relevance deficiencies (or both) for the set of the involved fields, depending on the relation between the set of values of this field and the set of values dictated by the rule. For example, a rule of imposition in the LOM standard, dictates that the value of the field "5.11 Educational.Language" of a learning object must be equal to the value of a learning object related to the first one with the relation

https://doi.org/10.23106/dcmi.952109222

"IsRequiredBy", in the sense that if a learning object is required by another one, then the human language used by the typical intended user of this object will be the same with the corresponding language of the object that requires it. Considering several combination of values, we have: If learning object x "IsRequiredBy" learning object y and "5.11" of x = "en", while "5.11" of y = "en", then the rule holds, and there is a clear indication of quality of the two involved fields. If "5.11" of x does not have any value, while "5.11" of y = "en", then the rule does not hold, and there is indication of reduced completeness. If "5.11" of x = "en", while "5.11" of y = "en", "fr", then the rule does not hold and there is indication of reduced completeness, as well. If "5.11" of x = "en, "fr", while "5.11" of y = "en", "it", then the rule does not hold. This situation might imply either a problem of reduced correctness ("fr" has been mistakenly taken for "it"), or a problem of both correctness and completeness ("fr" has by error been included in the values of "5.11" of x, while at the same time "it" has been omitted from the set of the values). In the last case, the indicated quality problems do not concern only learning object x, but both related objects as a pair, since the quality of y has not been taken for granted.

- *Validity of a rule of restriction:* The validity of a rule of restriction, that is, the presence of a value of the metadata field under examination within the restricted range dictated by the rule, is an indication of quality. On the contrary, if the value of this field is outside the dictated range, it is a case of absence of correctness for the involved fields. For example, a rule of restriction in the LOM standard, dictates that the value of the field "1.8 General.Aggregation level" of a learning object must be less than the minimum aggregation level of its parts (the learning object is related with its parts with the relation "IsPartOf"). If such is the case, then the validity of the rule indicates quality of the involved objects. If the value of "1.8" of the learning object is not less than the minimum aggregation level of its parts, then the rule does not hold and there is an indication that the values of the involved fields are not correct.

The quality problems located by examining the validity of logic rules provide valuable hints to the administrators of the metadata repository. Although the method cannot locate the problematic component of quality, exactly on a single record or element, it restricts the field of interest and focuses on a reduced set of resources with degraded quality. This is evident, since the conclusions one can draw by examining the validity of the rules concern more than one (related) fields, where no field is considered to be of high quality in advance. In the general case, where no such assumptions deriving from the context of use or the specific application are made, the set of the related fields with problematic quality is the limit of the quality assessment's "granularity". However, this method combined with other methods of metadata quality assessment can be of valuable contribution. For example, metrics referenced in Section 3, or manual inspection by experts can be applied to the set of the fields not following a certain rule, in order to pinpoint the problematic ones. This is much more feasible and efficient compared to the usage of these methods over the whole repository.

The logic rules, which in this paper are proposed to be used as a means for quality assessment of the metadata, can also be used to enhance quality when chosen to be applied and modify the values of the involved fields. Used as metadata generation rules, they can increase completeness by populating empty fields, as well as correctness or relevance by replacing faulty values. Especially, the increase of relevance can be considered as a method to preserve the context in the metadata records, in cases where the records are created by various indexers with diverse backgrounds. Of course, all these benefits are a result of well established application policies on the fields to be considered of high quality as reference.

https://doi.org/10.23106/dcmi.952109222

6. Conclusion and Future Work

In this paper, a new framework for conceptualizing metadata quality was defined using analogies from common knowledge and experience. The framework was inspired from entities and procedures involved in a court of law and aims at setting a solid, simplified theoretical background by defining the fundamental components of metadata quality, namely: correctness, completeness and relevance. Then, metadata quality assessment is performed by assessing these three components. Hence, the paper proposes a method for assessing metadata quality by exploiting structural and semantic relations among digital resources in a repository. Such relations create logic rules connecting the metadata of the related resources. Examining the validity of the rules serves as a means to conduct quality control on the metadata of the involved resources.

The conclusions deriving from this process can form the basis for a metric system to measure the components of metadata quality. Possible factors to be taken into account in the design of the metric system might be the number of non-valid rules at record or repository level, the number of the involved fields in a rule, the number of faulty or missing values in a field, the number of the resources participating in a problematic set, the number of problematic sets a single resource participates in, and so on. The design of such metrics is a step forward following this work. The method proposed in this paper can be combined with other metadata quality assessment methods and techniques in an integrated quality assurance system for the metadata of a digital repository.

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. *Proceedings of Dublin Core Conference 2003: Supporting Communities of Discourse and Practice - Metadata Research and Applications, 2003,* (pp. 39-48).

Bourda, Yolaine, Bichlien Doan, and Walid Kekhia. (2002). A semi-automatic tool for the indexation of learning objects. *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, 2002,* (pp. 190-191).

Brase, Jan, Mark Painter, and Wolfgang Nejdl. (2003). Completion Axioms for learning object metadata - Towards a formal description of LOM. *3rd IEEE International Conference on Advanced Learning Technologies (ICALT 2003).*

Bruce, Thomas R., and Diane I. Hillmann. (2004). The continuum of metadata quality: Defining, expressing, exploiting. In Dianne. I. Hillmann and Elaine L. Westbrooks (Eds.), *Metadata in practice,* (pp. 238-256). Chicago: ALA.

Bui, Yen, and Jung-ran Park (2006). An assessment of metadata quality: A case study of the national science digital library metadata repository. In Haidar Moukdad (ed.), *CAIS/ACSI 2006 Information Science Revisited: Approaches to Innovation* from http://www.cais-acsi.ca/proceedings/2006/bui_2006.pdf.

Currier, Sarah, Jane Barton, Rónán O'Beirne, and Ben Ryan. (2004). Quality assurance for digital learning object repositories: Issues for the metadata creation process. *ALT-J, Research in Learning Technology, 12*(1), 5-20.

DCMI Usage Board. (2008). *DCMI Metadata Terms.* Retrieved March 18, 2008, from http://dublincore.org/documents/dcmi-terms/.

Doan, Bich-lien, and Yolaine Bourda. (2005). Defining several ontologies to enhance the expressive power of queries. *Proceedings on Interoperability of web-based Educational Systems, WWW'05 conference, Chiba, Japan.*

Dushay, Naomi., and Diane I. Hillmann. (2003). Analyzing metadata for effective use and re-use. *DCMI Metadata Conference and Workshop, Seattle, USA.*

Duval, Erik, and Wayne Hodgins. (2004). Metadata matters. *Proceedings of the International Conference on Dublin Core and Metadata Applications, 2004,* (pp. 11-14).

Friesen, Norm. (2004). *International LOM Survey: Report (Draft).* Retrieved March 18, 2008, from http://dlist.sir.arizona.edu/403/01/LOM%5FSurvey%5FReport2.doc.

Gasser, Les, and Besiki Stvilia. (2001). A new framework for information quality. *Technical report, ISRN UIUCLIS–2001/1+AMAS, 2001*.

https://doi.org/10.23106/dcmi.952109222

Greenberg, Jane, Maria Cristina Pattuelli, Bijan Parsia, and W. Davenport Robertson. (2001). Author-generated Dublin Core metadata for web resources: A baseline study in an organization. *Proceedings of the International Conference on Dublin Core and Metadata Applications, 2001,* (pp. 38–46). National Institute of Informatics.

Guy, Marieke, Andy Powell, and Michael Day. (2004). Improving the quality of metadata in Eprint archives. *Ariadne 38.*

Hatala, Marek, and Griff Richards. (2003). Value-added metatagging: Ontology and rule based methods for smarter metadata. In Michael Schroeder and Gerd Wagner (Eds.), *RuleML 2003,* (pp. 65–80).

Hillmann, Diane. I., and Jon Phipps. (2007). Application profiles: Exposing and enforcing metadata quality. *Proceedings of the International Conference on Dublin Core and Metadata Applications, 2007,* (pp. 52-62).

Hughes, Baden. (2004). Metadata quality evaluation: Experience from the open language archives community. *Digital Libraries: International Collaboration and Cross-Fertilization,* (320–329).

IEEE. 1484.12.1 (2002). Draft Standard for Learning Object Metadata*. Learning Technology Standards Committee of the IEEE*. Retrieved 18 March, 2008, from http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf.

Margaritopoulos, Merkourios, Athanasios Manitsaris, and Ioannis Mavridis. (2007). On the Identification of Inference Rules for Automatic Metadata Generation. *Proceedings of the 2nd International Conference on Metadata and Semantics Research (CD-ROM), 2007.* Ionian Academy.

Moen, William E., Erin L. Stewart, and Charles L. McClure. (1997). Assessing metadata quality: Findings and methodological considerations from an evaluation of the US Government information locator service (GILS). *Proceedings of the Advances in Digital Libraries Conference, 1998,* (p. 246). IEEE Computer Society

Motelet, Olivier. (2005). Relation-based heuristic diffusion framework for LOM generation. *Proceedings of 12th International Conference on Artificial Intelligence in Education AIED 200.* Amsterdam, Holland: Young Researcher Track.

Najjar, Jehad, Stefaan Ternier, and Erik Duval. (2004). User behavior in learning object repositories: An empirical analysis. *Proceedings of the ED-MEDIA 2004 World Conference on Educational Multimedia, Hypermedia and Telecommunications, AACE, 2004,* (pp. 4373–4379).

Ochoa, Xavier, and Erik Duval. (2006). Quality Metrics for Learning Object Metadata. In Elain Pearson and Paul Bohman (Eds.), *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, 2006,* (pp. 1004-1011). Chesapeake, VA: AACE.

Sicilia, Miguel-Angel, Elena García-Barriocanal, Carmen Pagés, José-Javier Martínez, and José-María Gutiérrez. (2005). Complete metadata records in learning object repositories: Some evidence and requirements. *International Journal of Learning Technology, 1*(4), 411-424.

Wason, Thomas D., and David Wiley. (2001). Structured Metadata Spaces. In Jane Greenberg (ed.), *Metadata and Organizing Educational Resources on the Internet,* (pp. 263-277). New York: Haworth Press.

https://doi.org/10.23106/dcmi.952109222