Noname manuscript No.
(will be inserted by the editor)

Automatic Evaluation of Metadata Quality in Digital Repositories

Xavier Ochoa · Erik Duval

Received: date / Accepted: date

Abstract Due to recent developments in automatic metadata generation and interoperability between digital repositories, the production of metadata is now vastly surpassing manual quality control capabilities. Abandoning quality control altogether is problematic, because low quality metadata compromise the effectiveness of services that repositories provide to their users. To address this problem, we present a set of scalable quality metrics for metadata based on the Bruce & Hillman framework for metadata quality control. We perform three experiments to evaluate our metrics: 1) the degree of correlation between the metrics and manual quality reviews, 2) the discriminatory power between metadata sets and 3) the usefulness of the metrics as low quality filters. Through statistical analysis, we found that several metrics, especially Text Information Content, correlate well with human evaluation and that the average of all the metrics are roughly as effective as people to flag low quality instances. The implications of this finding are discussed. Finally, we propose possible applications of the metrics to improve tools for the administration of digital repositories.

Keywords Metrics · Metadata Quality · Digital Libraries · Learning Object Repositories · Text Information Content

1 Introduction

The quality of metadata instances stored in digital repositories is perceived as an important issue for their operation [1] [2] and interoperability [23] [43].

The main functionality of a digital repository, to provide access to resources, can be severely affected by the quality of the metadata. For example, a learning resource indexed with the title “Lesson 1 - Course CS20”, without any description or keywords will rarely appear in a search for materials about “Introduction to Java Programming”, even if the described resource is, indeed, a good introductory text to Java. The resource will just be part of the repository but will never be retrieved in relevant searches.

Secondary functions of metadata in a digital repository can also be heavily compromised by low metadata quality. For example, the metadata instance should contain enough information, so that the user can obtain a good idea of the purpose and content of the described resource without directly accessing the resource. For example, incorrect or out-dated information about the URI of the resource could prevent the user to access the object. Also, the effectiveness of a distributed search could be degraded even if just one of the connected repositories contains mainly low quality metadata instances. Consequently, the usefulness of a digital repository is strongly correlated to the quality of the metadata that describe its resources.

Due to its importance, metadata quality assurance has always been an integral part of resource cataloging

X. Ochoa
Centro de Tecnologías de Información
Escuela Superior Politécnica del Litoral
Via Perimetral Km. 30.5 Guayaquil, Ecuador
Tel.: +593-4-2269773
Fax: +593-4-2269776
E-mail: xavier@cti.espol.edu.ec

E. Duval
Dept. Computerwetenschappen
Katholieke Universiteit Leuven
Celestijnenlaan 200 A B-3001 Leuven, Belgium
Tel: +32-16-327066
Fax: +32-16-327996
E-mail: Erik.Duval@cs.kuleuven.be

[46]. Nonetheless, most implementations of digital repositories have taken a relaxed approach to metadata quality assurance. For example, these implementations rely on the assumption that metadata were created by an expert in the field or a professional cataloguer and, as such, should have an acceptable degree of quality. In reality, experts in a given field are not necessarily experts in metadata creation, and hiring professional indexers to do the cataloging of resources is usually not feasible for most repositories due to scalability issues and the costs involved.

As repositories grow (through automatic metadata generation [7] or resource decomposition [47]) and merge (through search federation [39] or metadata harvesting [40]), quality issues become more apparent. This problem has led to the adaptation of techniques developed to review physical library instances to address the quality of digital metadata. Also, new techniques that take advantage of the ability of computers to perform repetitive calculations have been developed to assure a minimum level of quality. A review of earlier work on metadata quality evaluation for digital repositories reveals these two general approaches:

– *Manual Quality Evaluation.* The majority of approaches (see Table 1) manually review a statistically significant sample of metadata instances against a predefined set of quality parameters, similar to sampling techniques used for quality assurance of library cataloguing [8]. Human evaluations are averaged and an estimation of metadata quality in the repository is obtained. Until now, these methods are the most meaningful way to measure the metadata quality in a digital repository. However, they have three main disadvantages: 1) the manual quality estimation is only valid at sampling time. If a considerable amount of new resources is inserted in the repository, the assessment could be no longer accurate and the estimation must be redone. 2) only the average quality can be inferred with these methods. The quality of individual metadata instances can only be obtained for those instances contained in the sample. 3) obtaining the quality estimation in this way is costly. Human experts should review a number of objects that, due to the growth of repositories, is always increasing. Dushay and Hillman, in [10], propose the use of visualization tools to help metadata experts in their task, but it is still mainly a manual activity.
  Because of this last disadvantage, manual review of metadata quality is mainly a research activity with few practical implications in the functionality or performance of the digital repository.

Table 1 Review of different quality evaluation studies

Study	Approach	# of instances	Main focus of evaluation
[16]	Manual	11	Quality of non-expert metadata
[37]	Manual	140	Overall quality of instances
[43]	Manual	150	Identify quality problems
[48]	Manual	100	Quality of non-expert metadata
[27]	Manual	80	Overall quality of instances
[19]	Statistical	27,000	Completeness of instances
[29]	Statistical	3,700	Usage of the metadata standard
[6]	Statistical	1,040,034	Completeness of instances

– *Simple Statistical Quality Evaluation.* From the studies we analyzed, three follow a different approach (see Table 1). These studies collect statistical information from all the metadata instances in the repository to obtain an estimation of their quality. Hughes, in [19], calculates simple automatic metrics (completeness, vocabulary use, etc.) at repository level for each of the repositories in the Open Language Archive [20]. Bui and Park [6] perform a wide study in which more than one million instances were reviewed for completeness. Najjar et al. [28] compare the metadata fields that are produced with the metadata fields that are used in searches. This comparison provides a simple estimation of the quality of the metadata in the ARIADNE [12] repository. All these studies automatically obtained a basic estimation of the quality of each individual metadata instance without the cost involved in manual quality review. However, they do not provide a similar level of "meaningfulness" as a human generated estimation. They are mainly used as "interesting" information about the repository without any other real application.

An ideal measurement of metadata quality for fast-growing repositories should have two characteristics: to be automatically calculated for each metadata instance inserted in the repository (scalability) and to provide a useful measurement of the quality (meaningfulness). None of the approaches reviewed could claim to be scalable and meaningful at the same time. Manual evaluations are meaningful but not scalable. Simple Statistics are scalable, but are not meaningful. The main contribution of this paper is the description and evaluation of a set of metadata metrics based on the same quality parameters used by human reviewers but with the difference that they can be calculated automatically. These metrics can be used to build tools for any kind of digital repository and can provide scalable and meaningful metadata quality assurance. These kind of automated quality assurance is key to enable a true Learning Object Economy where millions of objects are published and automatically labelled throughout their lifetimes.

The structure of this paper is as follows: A review is conducted in section 2 to select a framework to measure metadata quality. Based on the selected framework, ten quality metrics are described in section 3. Three validation studies are conducted in section 4 to evaluate 1) the degree of correlation between the proposed metrics and human quality review, 2) the discriminatory power of the metrics and 3) the effectiveness of the metrics as low quality instances filters. The implications of the findings are also discussed in detail in section 4. Section 5 describes possible applications of the quality metrics. The paper closes with related work and conclusions.

2 Measuring Metadata Quality

Despite the wide agreement on the need to produce high quality metadata, there is less consensus on what high quality means and even less on how it should be measured. This paper will consider quality as the measure of fitness for a task [13]. The tasks metadata should enable in a digital repository are to help the user to find, identify, select and obtain resources [32]. The quality of the metadata will be directly proportional to how much it facilitates those tasks.

Measurements of the quality of the metadata instance do not address the quality of the metadata schema or the set of values that fields on the schema could take (we call these sets vocabularies). These measurements should be schema-agnostic, when possible. They also do not evaluate the quality of the resources themselves. This paper will provide metrics to estimate the quality of the information entered manually by indexers, generated automatically or a mixture of both.

In order to reduce subjectivity in the assessment of information quality, several researchers have developed quality evaluation frameworks. These frameworks define parameters that indicate whether information should be considered of high quality. Different frameworks vary widely in their scope and goals. Some have been inspired by the Total Quality Management paradigm [41]. Others are used in the field of text document evaluation, especially of Web documents [49]. Particularly interesting for our work, because they are focused on metadata quality, are the frameworks that have evolved from the research on library catalogs [13].

While no consensus has been reached on conceptual or operational definitions of metadata quality, there are three main references that could guide this kind of evaluation. We rely on these here as they summarize the recommendations made in previous information quality frameworks and eliminate redundant or overly specific quality parameters. Moen et al. [27] identify 23 quality parameters. However, some of these parameters (ease of use, ease of creation, protocols, etc) are more focused on the metadata schema standard or metadata generation tools. Given that the metrics should be schema-agnostic and measure only the quality of metadata instance, [27] is not considered as our base framework. Stvilia et al. [44] use most of Moen's parameters (excluding those not related with metadata quality), add several more, and group them in three dimensions of Information Quality (IQ): Intrinsic IQ, Relational/Contextual IQ and Reputational IQ. Some of the parameters (accuracy, naturalness, precision, etc) are present in more than one dimension. The Stvilia et al. framework describes 32 parameters in total. Bruce & Hillman [5], based on previous Information Quality research, condense many of the quality parameters in order to improve their applicability. They describe seven general characteristics of metadata quality: completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility. A relation between the frameworks of Bruce & Hillman and Stvilia et al. is proposed in [37] and it is summarized in Figure 1.

This analysis will use the Bruce & Hillman framework because its seven parameters are easy to understand by human reviewers and also because they capture all the dimensions of quality proposed in other frameworks. The compactness will also help to operationalize the measurement of quality in a set of automatically calculated metrics. Another advantage of this choice is that this framework is deeply rooted on well-known Information Quality parameters. There exists parallel research on how to convert these parameters into metrics for quality assurance of other types of information (for instance Web Pages [49]). However, the Bruce & Hillman framework (also Stvilia et al.) is designed with a static metadata instance in mind. These frameworks are more appropriate for library purposes than for the metadata instances of digital libraries. This kind of metadata, referred in this paper as dynamic metadata, can change each time that the resource is used or accessed. Given that to the knowledge of the author there are no frameworks to describe the quality of dynamic metadata, Bruce & Hillman will be used

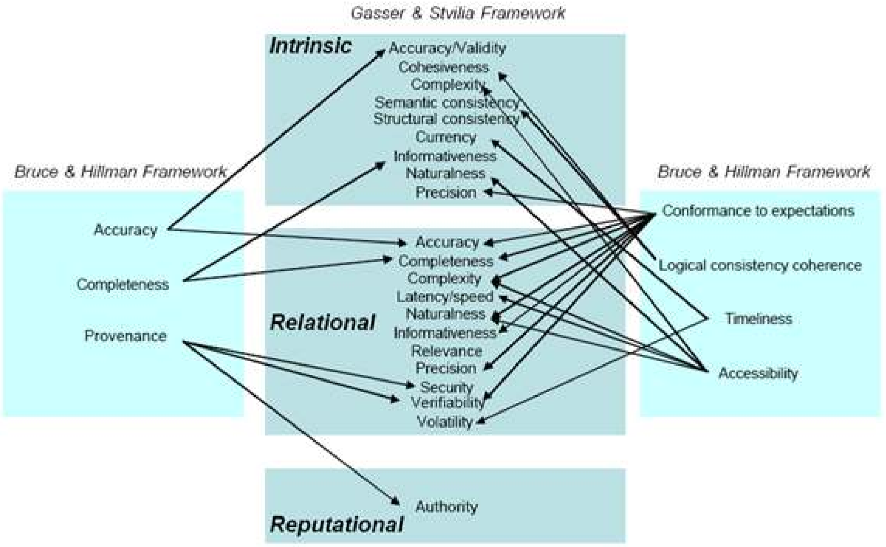


Fig. 1 Mapping between the Bruce & Hillman and the Stvilia et al. frameworks. (Taken from [37])

as a first approach, adapting the quality characteristics when needed for the particularities of dynamic instances. These adaptations are presented in section 3, where the metrics are introduced.

To improve the readability of this paper, a summary of the framework developed by Bruce & Hillman is presented. This framework defines seven parameters to measure the quality of metadata. These parameters are:

- *Completeness:* A metadata instance should describe the resource as fully as possible. Also, the metadata fields should be filled in for the majority of the resource population in order to make them useful for any kind of service. While this definition is most certainly based in static library instance view of metadata, it can be use to measure how much information is available about the resource.
- *Accuracy:* The information provided about the resource in the metadata instance should be as correct as possible. Typographical errors, as well as factual errors, affect this quality dimension. However, estimating the correctness of a value is in not always a "right"/"wrong" choice. There are metadata fields that should receive a more subjective judgement. For example, while it is easy to determine whether the file size or format are correct or not, the correctness of the title, description or difficulty of an object has much more levels that are highly dependent of the perception of the reviewer.
- *Conformance to Expectations:* The degree to which metadata fulfills the requirements of a given community of users for a given task could be considered as a major dimension of the quality of a metadata instance. If the information stored in the metadata helps a community of practice to find, identify, select and obtain resources without a major shift in their workflow it could be considered to conform to the expectations of the community. According to the definition of quality ("fitness for purpose") used in this paper, this is one of the most important quality characteristics.
- *Logical Consistency and Coherence:* Metadata should be consistent with standard definitions and concepts used in the domain. The information contained in the metadata should also have internal coherence, that means that all the fields describe the same resource.
- *Accessibility:* Metadata that cannot be read or understood have no value. If the metadata are meant for automated processing, for example GPS location, the main problem is physical accessibility (incompatible formats or broken links). If the metadata are meant for human consumption, for example Description, the main problem is cognitive accessibility (metadata is too difficult to understand). These two different dimensions should be combined to estimate how easy is to access and understand the information present in the metadata.
- *Timeliness:* Metadata should change whenever the described object changes (currency). Also, a complete metadata instance should be available by the time the object is inserted in the repository (lag). The lag description made by Bruce & Hillman, however, is focused in a static view of metadata. In a

digital library approach, the metadata about a resource is always increasing which each new use of the resource. The lag, under this viewpoint, can be considered as the time that it takes for the metadata to describe the object well enough to find it using the search engine provided in the repository.

- *Provenance:* The source of the metadata can be another factor to determine its quality. Knowledge about who created the instance, the level of expertise of the indexer, what methodologies were followed at indexing time and what transformations the metadata has passed through, could provide insight into the quality of the instance.

For a discussion on the rationale behind these parameters, as well as for a thoughtful analysis of what "metadata quality" means, we invite the reader to consult [5]. The following section will present calculations (metrics) that could provide a low cost estimation of some aspects of these quality parameters.

3 Quality Metrics for Metadata in Digital Repositories

Bruce & Hillman [5] devised their framework to guide human reviewers. The parameters, being domain independent, are necessarily abstract. This level of abstraction could be easily managed by metadata experts, but presents a problem for the automatic estimation of quality. This section will describe a set of calculations that work over the existing metadata information and easy-to-collect contextual data in order to "instantiate" the quality parameters into a set of quality metrics. The objective of these metrics is to provide an initial development of meaningful measurements to estimate the quality of each metadata instance for a given community of practice in a scalable way. These metrics address some of the facets of each dimension characteristic described in the Bruce & Hillman framework, but are not a perfect or comprehensive measurement of those characteristics. Also the proposed metrics do not replace more simple calculations of quality, but complement them.

3.1 Completeness Metrics

As described in section 2, Completeness is the degree to which the metadata instance contains all the information needed to have a comprehensive representation of the described resource. While easy to understand for static, library records, this concept is less clear for dynamic metadata instances, where new information is added each time that the resource is used. In the case of dynamic metadata, there is certain information, that, due its nature, should be present to enable the services of the digital library. For example, some digital libraries rely on the title of the object to present it in a list to the user. If the metadata do not contain a title, the quality of the metadata decrease. On the other hand, while reviews and ratings collected through the lifetime of the resource are highly valuable, the lack of this data does not prevent the digital library searching facilities from present the results to the user. This metric should consider the former types of metadata information to estimate its completeness.

The most direct approach to measure completeness of an instance is to use the number of filled in metadata fields as a proxy. Each metadata standard, for example Dublin Core (DC) [9] or Learning Object Metadata (LOM) [21], defines a number of possible fields (15 for Simple DC [1], 58 for LOM). In some cases, there can be more than one instance of the fields. A basic completeness metric will be to count the number of fields in each metadata instance that contain a no-null value. In the case of multi-valued fields, the field is considered complete if at least one instance exists. Equation 1 expresses how this metric can be determined.

$$Qcomp = \frac{\sum_{i=1}^{N} P(i)}{N} \tag{1}$$

Where $P(i)$ is 1 if the ith field has a no-null value, 0 otherwise. N is the number of fields defined in the metadata standard.

The maximum value of this metric is 1 (in the case all the fields contain information) and the minimum value is 0 (an empty instance). For example, if a LOM instance has 40 fields filled in, its $Qcomp$ value will be $40/58 = 0.69$.

While straightforward, the simple completeness metric does not reflect how humans measure the completeness of an instance. Not all data elements are relevant for all resources. Moreover, not all metadata elements are equally relevant to all contexts. For example, a human expert may assign a higher degree of completeness to a metadata instance that has a title, but lacks publication date than vice versa. To account for this phenomenon, a weighting factor could multiply the presence or absence of a metadata field. This factor represents the importance of the field. This weighting factor can easily be included in the calculation of the completeness metric as shown in the Equation 2.

[1] The full DC standard allows the addition of more fields for specialized purposes

$$Qwcomp = \frac{\sum_{i=1}^{N} \alpha_i * P(i)}{\sum_{i=1}^{N} \alpha_i} \quad (2)$$

Where α_i is the relative importance of the ith field.

The maximum value for *Qwcomp* will be 1 (all fields with importance different from 0 are filled) and a minimum value of 0 (all fields with importance different from 0 are empty).

The α should be any positive value that represent the importance (or relevance) of the metadata field for some context or task. This implies that each community of practice could have a different set of weighting factors to calculate the weighted completeness for different kinds of tasks. For example, α_i could represent the number of times field i has been used in queries to a given repository [29].

The Weighted Completeness measure the completeness of a metadata instance against the current needs of a community. Therefore some fields, that currently are not important for the community have no impact in the Qwcomp metric (value 0). However, to avoid to disregard fields that could become important in the future, it is recommended that the implementation of the weighting coefficient should be made adaptable to changes in the needs of the users. For example, if a field is used more frequently on queries, of its importance change for any reason, the weighted completeness metrics should change accordingly.

Consider a metadata standard that has 4 fields: Title, Description, Author Name and Publication Date. Consider also that after 5000 queries to the repository, Title has been used 5000 times, Description 2500 times, Author Name 1000 times and Publication Date 0 times, so $\alpha_1 = 5000$, $\alpha_2 = 2500$, $\alpha_3 = 1000$ and $\alpha_4 = 0$. Table 2 shows the *Qwcomp* calculation for different instances and its contrast with Qcomp. The presence of Publication Date is not relevant for the Qwcomp as it is never used in the queries. Title, on the other hand, is the most important field, and its sole presence corresponds to more than half the completeness value.

Alternatively, when measuring the completeness for the selection task, α_i could represent the score that the ith field obtained in an experiment to measure the amount of time the user expend reading each field while selecting an appropriate resource. Obtaining the alpha values from the analysis of user interaction with the digital library has the added advantage of adapting the completeness quality estimation to the changing behavior of the user community. Each time that new importance values are generated (for example, query information is collected or usability studies are performed) a more refined estimation could be obtained.

The algorithm to calculate the Completeness metric needs only access to the metadata repository. For the Weighted Completeness, also a table containing the precalculated α values should be available.

3.2 Accuracy Metrics

Accuracy is the degree to which metadata values are "correct", i.e. how well they describe the object. For objective information like file size or document format correctness could be a binary value, either "right" or "wrong". In the case of subjective information, it is a more complex spectrum with intermediate values (e.g.: a title of a picture, or the description of the content of a document). In general, correctness and, therefore accuracy, could be considered as the semantic distance between the information a user could extract from the metadata instance and the information the same user could obtain from the resource itself and its context. The shorter the distance, the higher the accuracy of the metadata instance.

While humans can assess with relative ease the accuracy of a metadata instance, computers require complex artificial intelligence algorithms to simulate the same level of understanding. Nevertheless, there exists accuracy metrics that are easy to calculate, proposed in quality evaluations presented in [20] and [27]. These metrics establish the number of easy-to-spot errors present in metadata instances. Typical examples of this type of errors are broken links, inaccurate technical properties of the digital resource, such as size or format, typographical errors in the text fields, to name a few.

This paper proposes a more complex and meaningful approach to calculate the semantic difference between the metadata instance and resources that contain textual information. Using the metadata and the resource itself helps to provide a better estimation of the accuracy of the record that just counting the number of errors in the metadata. This method builds upon Vector Space model techniques used in Information Retrieval to calculate the distance between texts [35]. A multidimensional space is constructed in which each word present in the text of the original resource defines a dimension. The number of times a word appears in the text is considered as the value of the text in that word-dimension. Following those definitions, a vector is created for the text contained in the original resource and the text present in the textual fields of the metadata instance (e.g. title, description, keywords, etc.). Finally, a vector distance metric, such as the cosine distance,

Table 2 Example of the calculation of Qwcomp for a 4-field metadata instances

instance				**Qwcomp**	**Qcomp**	**Qwcomp**
Title	*Desc.*	*Author*	*Date*			
Yes	No	No	No	(5000)/8500	0.25	0.59
No	No	Yes	Yes	(1000+0)/8500	0.50	0.12
Yes	Yes	Yes	No	(5000+2500+1000)/8500	0.75	1.0

Table 3 Example of the Qaccu values for two metadata instances

	Qaccu
Resource 1	**0.56**
Metadata Text (title+description): *SEPHYR METHODOLOGY*	
Extract of Resource Text (Word document): *Methodology of Pedagogic Segmentation Extract from the doctoral thesis by Miss M. Wentland Forte entitled: "Knowledge domain modeling and conceptual orientation in a pedagogic hypertext" What is a concept? Taking it at the level of the spontaneous mental processes (unorganized and non-verbalized), we can say that we are dealing with the realm of ideas. As soon as an idea can be named, it becomes a concept..*	
Resource 2	**0.96**
Metadata Text (title+description): *Searching for the Future of Metadata - Looking in the wrong places for the wrong things? Keynote at DC2004 conference*	
Extract of Resource Text (PowerPoint presentation): *Searching for the Future of Metadata Looking in the wrong places for the wrong things? by: Wayne Hodgins wayne.hodgins@autodesk.com Looking in the Wrong Places? Searching Helping Remembering. Looking in the Right Place? Looking in the Right Place Looking where the light is!! A few words about LEARNING Vision for learning*	

is applied to find the semantic distance between both texts. In Equation 3 the cosine distance formula is presented.

$$Qaccu = \frac{\sum_{i=1}^{N} tfresource_i * tfmetadata_i}{\sqrt{\sum_{i=1}^{N} tfresource_i^2 * \sum_{i=1}^{N} tfmetadata_i^2}} \qquad (3)$$

Where $tfresource_i$ and $tfmetadata_i$, are the relative frequency of the ith word in the text content of both the resource and the metadata. N is the total number of different words in both texts.

The minimum value (lower quality) is 0, meaning that the two texts have no words in common. The maximum value (higher quality) is 1, meaning that all the words from one of the texts appear in the other.

Table 3, as an example, presents two metadata instances with an excerpt from the text from their respective described objects. The resulting *Qaccu* value is also presented for each instance. In the first example, the word "SEPHYR" appears in the metadata description, but it cannot be found in the document. The other word in the title, "METHODOLOGY", is matched against the same and similar words in the resource text. Given that the dimensionality of the metadata text is 2, the result is approximately 0.5. In the second example, most of the words that appear in the title and description also appear in the document itself. The result of the *Qaccu* metric approaches 1.

The first example on Table 3 is also a good demonstration of how this metric could fail for some especial cases in real world applications. While the word "SEPHYR" does not appear at all in the text of the document itself, the document is indeed about the "SEPHYR METHODOLOGY". To minimize the impact of this type of omissions, a method to detect synonyms or words with close semantic relation can be used. One of the most successful is Latent Semantic Analysis (LSA) [22]. This algorithm can be used to reduce the dimensionality of the space before the distance calculation. With the reduction of dimensionality, the noise introduced by semantic similar words is reduced.

To implement this metric, the LSA algorithm needs to be trained with corpora taken from the text resources present in the repository. Afterwards, the lower-dimensional matrices resulting from the Single Value Decomposition (SVD) [22] calculation could be used to compute the semantic distance between any arbitrary pair of texts.

3.3 Conformance to Expectation Metrics

Conformance to expectations measures the degree to which the metadata instance fulfills the reqpurppurposeuirements of a given community of users for a given task. As mentioned previously, metadata in digital repositories is mainly used to find, identify, select and obtain resources. The usefulness of a metadata instance for the first three tasks (find, identify and select) depends heavily on the amount of useful (unique) information contained in the instance. Instances with non-common words are easier to find. Users can differentiate resource more easily if their metadata instances are not similar. Users can make better selections if the instances provide better descriptions of the resource. A method that could measure the amount of unique information in the metadata instance can be used to estimate its conformance to the expectation of a community. The method proposed in this paper is the calculation of the information content of the metadata instance.

In Information Theory, the concept of entropy is used as a measure of the information content of a message [36]. Entropy is the negative logarithm of the probability of the message. Intuitively, the more probable a message is, the less information it carries. For example, if all the metadata instances in a repository have the field "language" set as "English", a new instance with that field set to "English" carries few information, meaning that it does not help to distinguish this particular resource from the rest. On the other hand, if the new instance has the "language" field set to "Spanish", it is highly improbable (based on the content of the repository) and this value helps to differentiate this new resource from the others.

The information content of categorical fields (those that can only take a value from a defined and finite vocabulary) can be easily calculated using the entropy method described in the previous paragraph. To obtain the Categorical Information Content for a given instance, the entropy values of the categorical fields can be averaged. This calculation is presented in Equations 4 and 5.

$$infoContent(cat\_field) = -\log(f(value)) \tag{4}$$

Where $f(value)$ is the relative frequency of $value$ in the categorical field for all the current instances in the repository. This relative frequency is equivalent to the probability of $value$.

$$Qcinfo = \frac{\sum_{i=1}^{N} infoContent(field_i)}{N} \tag{5}$$

Where N is the number of categorical fields.

Table 4 shows the Qcinfo calculation for some categorical fields of real metadata instances in the ARIADNE repository [12]. The Qcinfo is calculated by averaging only the entropy values of two fields: "Main Discipline" and "Difficulty Level". From these two categorical fields, Resource 1 seems to be an average instance, similar to the majority of the other instances in the repository, therefore it has a low Information Content (Qcinfo). On the other hand, Resource 2 is highly atypical, leading to a high Information Content value.

In order to normalize the Information Content value, so it will vary from a minimum of 0 (lowest quality) to a maximum of 1(highest quality), the formula in Equation 4 should be changed as to the one presented in Equation 6.

$$infoContent(cat\_field) = 1 - \frac{\log(times(value))}{\log(n)} \tag{6}$$

Where $times(value)$ is the number of times that the value is present in that categorical field in the whole repository. n is the total number of instances in the repository. When $times(value)$ is 0 (the value is not present in the repository), the infoContent is 1. On the other hand, if $times(value)$ is equal to n (all the instances have the same value), the infoContent is 0.

For free text fields the Information Content calculation is not as straight forward as for categorical fields. Each word in the text can be considered as a possible carrier of information. To calculate the total information content of textual fields we have to estimate the contribution of every word in each field. In the field of Information Retrieval, the importance of a word is calculated with the Term Frequency-Inverse Document Frequency (TFIDF) [34] value. The importance of a word in a document is directly proportional to how frequently that word appears in the document and inversely proportional to how frequently documents in the corpora contain that word. More concretely, the frequency in which a word appears in the document is multiplied by the negative log of the relative frequency in which that word appears in all the documents in the corpora. This calculation could be considered as a weighted entropy measurement for each word. To get the Information Content of the text field, the TFIDF value of each word is added. The Information Content of an instance can be calculated by adding the Information Content of its text fields. Equation 7 provides a description of the Information Content calculation.

$$infoContent(freetext\_field) = \sum_{i=1}^{N} tf(word_i) * \log\left(\frac{1}{df(word_i)}\right) \tag{7}$$

Where $tf(word_i)$ is the term frequency of the ith word, $df(word_i)$ is the document frequency of the ith word. N is the number of words in the field.

A common method to normalize TFIDF values is to divide the sum of the TFIDF values by the total number of words in the text. This division gives a measure of the information density. However, the Qtinfo metric attempts to estimate the total information content of

Table 4 Example of the calculation of Qcinfo for 2 metadata instances

Field	Value	f(value)	Entropy	Qcinfo
Resource 1				
Main Discipline	Computer Science	1220/4460	0.59	0.31
Difficulty Level	Medium	4124/4460	0.03	
Resource 2				
Main Discipline	Mechanical Engineering	314/4460	1.15	1.36
Difficulty Level	High	120/4460	1.57	

the instance, not its density. A way to reduce the range of the Information Content value, while preserving the length of the text as a component, is to obtain its logarithm. Therefore, the final formula for the Qtinfo is the logarithm of the sum of the Information Content of the textual fields (Equation 8).

$$Qtinfo = \log\left(\sum_{i=1}^{N} infoContent(field_i)\right) \tag{8}$$

Where $field_i$ is the ith textual field. N is the number of textual fields in the metadata standard.

Intuitively, texts composed mainly of common words in a language (for example: “the”, “a”, “for”, etc.) and words that are common in a given repository (for example, for a learning object repository: “course”, “lesson”, “material”, etc.) carry less information to identify the resource than more specialized words. Also, longer texts contain more information than shorter texts. Table 5 present the calculation of the Textual Information Content for four different texts extracted from metadata instances of the ARIADNE Repository. Although Resource 1 has fewer words (17), it obtains a slightly higher value than Resource 2 (19 words). The reason for the score difference is that the words “Metadata” and “Searching” are quite common in the ARIADNE repository. However, when one of the texts has considerably more words than other the difference is clearly represented in the Information Content values. This is the case of Resource 4 (291 words) which obtain a higher Qtinfo score than Resource 3 (34 words).

The information needed to calculate Qcinfo and Qtinfo could be extracted from the target repository. Precalculated probabilities for the categorical fields, such as Document Frequencies (DF) for existing words, can be stored in temporal database tables that are refreshed in periodical intervals. With this tables, the Qcinfo and Qtinfo calculation will only involve few mathematical operations.

Table 5 Example of the calculation of Qtinfo for text of different words and lengths

R.	Text	infoContent	Qtinfo
1	Gap Report Identified consequences of the developments of the other WPs for design of knowledge work management.	85	1.93
2	Searching for the Future of Metadata - Looking in the wrong places for the wrong things? Keynote at DC2004 conference	83	1.91
3	Traveling salesman This is a quick implementation of the traveling salesman problem, written in Java.It shows a frame with the execution of the backtracking algorithm for some citiesusage: java Practicum2 ’SHORTEST’—’ANY’	162	2.20
4	Control of the transfer channel Control of the transfer channel: First step towards a competence map This deliverable is based on the performance indicators developed in D8.4. Controlling the activities via the transfer channels by performance indicators is the main issue outlined in this deliverable. The monitoring process is focussed on major activities made by visitors... [235 WORDS MORE]	1557	3.19

3.4 Consistency & Coherence Metrics

3.4.1 Consistency

The logical consistency of a metadata instance can be estimated as the degree to which it matches the metadata standard definition. There are three ways in which this consistency can be broken: 1) instances include fields not defined in the standard or do not include fields

Table 6 Recommendations for values in the LOM Standard (v.1.0)

Field 1	Field 2	Recommendation
Structure	Aggregation Level	Structure=atomic :: Aggregation Level=1
Interactivity Type	Interactivity Level	Interactivity Type=active :: high values of Interactivity Level
Semantic Density	Difficulty	high values of Semantic Density :: high values of Difficulty
Resource Type	Interactivity Level	Resource type=narrative text :: Interactivity Level=expositive
Context	Typical Age Range	Context=higher education :: Age Range at least 17 years

that the community sets as mandatory. 2) Categorical fields, that should only contain values from a fixed list, are filled with a non sanctioned value 3) The combination of values in categorical fields is not recommended by the standard definition. In the case of isolated repositories, problems of type 1 and 2 are heavily reduced by the use of a common indexing tool. For distributed or aggregated repositories, problems of type 1 and 2 should be expected as the result of different indexing practices [37]. Problems of type 3 are more subtle and affect all types of repositories. They can be directly associated with violation of consistency rules at indexing time. An example of such rules is defined in the LOM Standard (v.1.0): If the value of the “Structure” field is set to “atomic”, the “Aggregation Level” field should be set as “1 (Raw media)”. Other Structure values could be paired with any other value of Aggregation Level except 1. Table 6 presents more of these rules for the LOM standard.

An estimation of the Consistency of the metadata instance should be inversely proportional to the number of problems found in the instance. Firstly, the amount of possible problems of type 1, 2 or 3 is obtained by the examination of the metadata standard and indexing rules of a community. Secondly, the number of problems present in the instance is counted. The number of problems of type 1 or 2 can be calculated with a simple validation parser. For problems of type 3, a set of ”If...Then” rules could be used instead. Finally, the Consistency metric will be equal to 1 minus the average of fraction of problems found, for each type of problem. Equations 9 and 10 present the calculation for the type 3 consistency. The minimum value for the Consistency metric is 0 (all possible errors were made) and the maximum value is 1 (there were no consistency problems).

$$brokeRule_i = \begin{cases} 0; \text{ if instance complies with } i\text{th rule} \\ 1; \text{ otherwise} \end{cases} \tag{9}$$

$$Qcons = 1 - \frac{\sum_{i=1}^{N} brokeRule_i}{N} \tag{10}$$

Where N is the number of rules in the metadata standard or community of use.

3.4.2 Coherence

The Coherence of the instance, on the other hand, is more related to the degree to which all the fields describe the same object in a similar way. The Coherence metric can be estimated analyzing the correlation between text fields. A procedure similar to the one used in the Accuracy metric (Section 3.2) can be implemented. The semantic distance is calculated between the different free text fields. The average semantic distance is used as a measure of the coherence quality (Equations 11 and 12). This method is commonly used to establish the internal coherence of a text piece [14]. To cope with synonyms, a LSA algorithm could be applied before the semantic distance is calculated.

$$\text{distance}(f1, f2) = \frac{\sum_{i=1}^{N} tfidf_{i,f1} * tfidf_{i,f2}}{\sqrt{\sum_{i=1}^{N} tfidf_{i,f1}^2 * \sum_{i=1}^{N} tfidf_{i,f2}^2}} \tag{11}$$

Where $tfidf_{i,field}$ is the Term Frequency Inverse Document Frequency of the ith word in the textual field f. N is the total number of different words in the field 1 and 2.

$$Qcoh = \frac{\sum_{i}^{N}\sum_{j}^{N} \begin{cases} \text{distance}(field_i, field_j); \text{ if i} < \text{ j} \\ 0; \text{ otherwise} \end{cases}}{\frac{N*(N-1)}{2}} \tag{12}$$

Where N is the number of textual fields that describe the object.

Table 7 Example of Qcoh calculation for 2 metadata instances

Field	Text	Qcoh
Resource 1		
Title	Infrastructure for (semi-)automatic generation of Learning Object Metadata	0.95
Description	The Month 6 deliverable for D4.1 is a "functional prototype" for an "infrastructure that supports (semi-) automatic generation of LOM metadata". As such, this deliverable consists of software: this report documents the design and status of the software. The actual software is also deposited on http://ariadne.cs.kuleuven.ac.be/amg	
Resource 2		
Title	Searching for the Future of Metadata - Looking in the wrong places for the wrong things?	0.0
Description	Keynote at DC2004 conference	

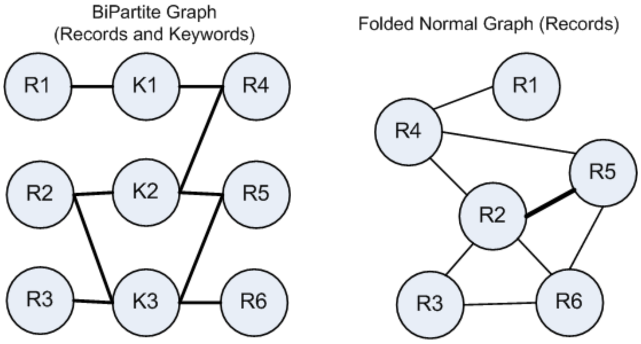
Table 7 presents the calculation of the Coherence metric for the Title and Descriptions belonging to Learning Objects in the ARIADNE Repository. If the Title and Description have semantically similar words, the Qcoh is close to 1, otherwise, as in the case of Resource 2, where there are no words in common, the QCoh approaches 0.

The second example in Table 7 presents a possible problem of this metric in real world scenarios. While the Title and Description are completely different, they are indeed describing the same resource. This problem will make this metric not very informative for individual instances. However, the value of Qcoh can provide some information if applied to a whole repository. A low value of Qcoh for a considerable number of instances could be the signal of poor titles or descriptions.

3.5 Accessibility Metrics

Accessibility implies the level to which a metadata instance can be found and later understood. It should not be confused with the more common meaning of accessibility, "design for all". One way to estimate the logical accessibility or "findability" of a metadata instance could be to count the number of times that the instance has been retrieved during searches. While these kind of studies carry a lot of information about the quality of the metadata, the"findability" of the objects does not only measures the intrinsic properties of the metadata instance, but also the capabilities of the searching tool and preferences of the users. Because is the objective of the proposed metrics to isolate the metadata properties, the calculation should measure the potential accessibility of the object independently of the method used for its retrieval. In Network Science, the logical accessibility of a node in the network is calculated as the number of links from the node to other nodes [30]. Borrowing this idea, this work proposes the use of the linkage of an instance as an intrinsic accessibility value. A link can be explicit (for example "is-related-to" or "is-version-of" fields) or it can be implicit (for example objects of the same author, on the same subject, etc.). An easy way to visualize how implicit linking could be calculated is to create a bipartite graph where Partition 1 contains the instances and Partition 2 contains the concept through which the linking will take place (authors, categories, etc.). Then the graph is folded over Partition 2, leaving a normal graph with linking between resources. An example of this procedure is shown in Figure 2.

Fig. 2 Procedure to establish the linking between instances, based on classifying concept



The linkage metric can be calculated by adding all links pointing from or towards an instance and dividing that number by the number of links of the most connected object (Equation 13).

$$Qlink = \frac{links(instance_k)}{\max_{i=1}^{N}(links(instance_i))} \tag{13}$$

Where $links(instance)$ represent the number of pointers to or from the metadata instance. N is the number of resources in the repository.

Cognitive accessibility measures how easy it is for a user to understand the information contained in the metadata instance. Librarians measure this characteristic [17] with several simple metrics: measuring spelling errors, conformance with the vocabulary, etc. Nonetheless, a better way to measure the accessibility will be to assess the overall difficulty of the text. However, this task requires human evaluation. The difficulty assessment could be estimated by automatic means using one of the available readability indexes, for example the Flesch Index [25]. This metric could be applied especially to analyze long text fields of instances (e.g. description). Readability indexes in general count the number of words per sentence and the length of the words to provide a value that suggest how easy it is to read a text. For example, a description where only acronyms or complex sentences are used will receive a lower score (lower quality) than a description where normal words and simple sentences are used. The text difficulty of the metadata is not necessarily related to the difficulty of the referred object itself. A complex book can be easily described and vice versa. What this metric try to estimate is how difficult would be to the user to understand the text contained in the metadata when it is presented to her.

Table 8 presents the calculation of the Flesch index for descriptions taken from descriptions of learning object metadata instances from ARIADNE. Short sentences and words lead to high values of Readability. On the other hand, long sentences, lack of punctuation, numbers and heavy use of acronyms reduce that value. The approximate maximum value of the Flesh index is 100 (easy to read text) while the minimum approximate value is 0 (hard to read text).

The readability metric is the normalized average of the Flesh Index of all the text fields in the instance. This calculation is presented in Equation 14.

$$Qread = \frac{\sum_{i}^{N} Flesch(fieldtext_i)}{100 * N} \tag{14}$$

Where N is the number of textual fields and $Flesch()$ is the calculation of the Flesch readability index.

3.6 Timeliness Metrics

The timeliness in digital repositories mainly relates to the degree to which a metadata instance remains current. The currency of a metadata instance could be measured as how useful the metadata remains with the pass of time. For example, if an instance describing a resource was created 5 years ago, and users could still find, identify, select and obtain the resource correctly, the metadata can be considered current. On the other hand, if the metadata instance misleads users, because the referred resource has changed to the point where the description in the metadata differed from the resource, the metadata instance is obsolete and must be changed or replaced.

The currency of the instance at a given time can be equated with its overall quality. Following this reasoning, the average value of previously presented metrics could be used as an estimation of the instantaneous currency of an instance (Equation 15). However, the instantaneous currency does not offer any information on how long the instance will continue to be current. For example, knowing that the currency of the description of a web page is high at the moment of the creation of the description does not guaranty that it will stay current after a year. Also, different objects change at different paces. A better estimation of the timeliness of an instance could be obtained measuring the rate of change of the instantaneous currency over a period of time. In more concrete terms, the timeliness of an in-

Table 8 Example calculation of the Flesch Index for different texts

Resource	Description	Flesch Index
1	This deliverable is based on the performance indicators developed in D8.4. Controlling the activities via the transfer channels by performance indicators is the main issue outlined in this deliverable. The monitoring process is focussed on major activities made by visitors and registered users in the Virtual Competence Centre. By organizing and managing the community of practice we focussed in the interpretation on two aspects...	80
2	This deliverable reports on the LOMI seminars - a series of virtual seminars on Learning Objects, Metadata and Interoperability (LOMI). The basic intent of the seminars is to facilitate exchange of opinions, ideas, plans and results on the overall theme of learning objects, metadata and interoperability. .... o 03 May, 15:00-16:30 CEST o 23 May, 16:00-17:30 CEST o 07 June, 15:00-16:30 CEST o 21 June, 16:00-17:30 CEST o 05 July, 15:00-16:30 CEST There is no cost involved for the participants. ...	30
3	Analysis of future professional training needs in Europe It is the same document as D6.1 joint report on economical approaches, user needs and market requirements for technology enhanced learning already submitted by WP6.	14

stance will be equal to its change of average quality per unit of time (Equation 16). Following the example of the web page descriptor, if after a year, the currency of the instance has been reduced by half, it is logical to expect that after another year it will be degraded to one quarter of its original currency. This metric can also measure positives changes in currency, for example, if the metadata instances are constantly enriched through user tagging and usage information. In those cases, the timeliness metric could be used to estimate how much better the instance will be after a defined period.

$$Qcurr = Qavg = \frac{\sum_{i=1}^{N} \frac{(Q_i - minQ_i)}{(maxQ_i - minQ_i)}}{N} \tag{15}$$

Where Q_i is the value of the ith quality metric (for example Qcomp, Qtinfo or Qread), $minQ_i$ and $maxQ_i$ are the minimum and maximum value of the ith metric for all the instances in the repository. N is the total number of metrics considered in the calculation. Q_avg is then the average of the different quality metrics for a given instance.

$$Qtime = \frac{Qcurr_{t2} - Qcurr_{t1}}{Qcurr_{t1} * (t2 - t1)} \tag{16}$$

Where $t1$ is the time when the original currency ($Qcurr_{t1}$) was measured and $t2$ is the current time with is corresponding value of instantaneous currency ($Qcurr_{t2}$).

The sign of $Qtime$ indicates if the change in quality has been positive (increase in quality) or negative (decrease in quality). The absolute value represents the rate of currency change per unit of time used (years, months, days, etc.). Equation 17 can be used to estimate the currency ($Qcurr$) of the instance in a future time. Table 9 presents some example calculations of $Qtime$. The lower bound for $Qcurr$ is 0 while it does not have an upper bound (a metadata instance could always be improved). Given that we are working with rates, this formula is identical to the one used to calculate the future value knowing the present value with compound interest.

$$Qcurr_{t3} = \left(\left(1 + Qtime_{(t2-t1)}\right)^{(t3-t2)} \right) \bullet Qcurr_{t2} \tag{17}$$

Where $Qtime_{(t2-t1)}$ is the calculation of the Qtime metric during the interval between $t1$ and $t2$. $t3$ is the time to which the Qcurr estimation is desired.

This metric can only be calculated if there are at least two values for Qavg taken at two different known times. In case that there are no previous measurements of $Qtime$ is 0 (no change). On the other hand, if three or more values of $Qavg$ exist, the $Qtime$ is calculated pairwise and then averaged in order to obtain a more representative value of change. Two consecutive measurements of $Qavg$ can be stored into the instance itself. Some metadata standards provide annotation fields where this information can be included. In case that the metadata standard or repository policies do not allow on-instance storage, a simple database could be implemented as part surrounding technological infrastructure.

3.7 Provenance Metrics

Provenance quality measures the trust that a given community has in the source of the metadata instance. For example, a metadata instance from the Library of Congress could be considered to have a higher Provenance quality than one generated in a local library. This higher level of provenance quality is not estimated by any intrinsic property of the metadata, but from the reputation that the Library of Congress has (quality assurance methods, expert staff, etc.) among the library community.

The main problem in converting the Provenance parameter into a metric is in obtaining information about the perception of the users about the metadata produced by a given source. This information can be captured explicitly, for example surveying the user about how useful the metadata has been to select the resources. The explicit collection of this type of information is bound to distract the user from her normal workflow. Given that the metrics proposed in this section should be an approximated measurement of the quality of the instance, a more scalable way to obtain the perceived quality of a source of metadata is to combine the metric values of its instances. The more straightforward way to combine those values is to first, obtain an Average Quality (Qavg) for each instance (Equation 15), and afterwards, average the Qavg of all the instances produced by the source (Equation 18). Once the quality of the source has been obtained, it is assigned to each one of its objects. This process is graphically explained in Figure 3.

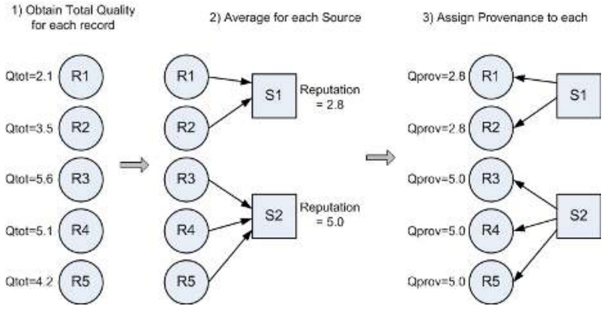
$$Qprov = Reputation(S) = \frac{\sum_{i=1}^{N} Qavg_i}{N} \tag{18}$$

Where $Qavg_i$ is the Average Quality of the ith instance contributed by the source S. N is the total num-

Table 9 Example calculation of Qtime

t1-t2	Qavg(t1)	Qavg(t2)	Qtime	Qcurr in 1 year
1 year	0.8	0.5	−37.5% per year	0.31
1 year	0.5	0.8	+60% per year	1.28
1 month	0.95	0.85	−26% per month	0.22

Fig. 3 Calculation of the Source Reputation and the Provenace of each instance (R represent the instances and S the sources)



ber of instances produced by S. The *Qprov* of an instance is equal to the reputation of its source.

The Qprov metric can be calculated once the other quality metrics has been calculated and assigned to each instance. As it can be distilled from the calculations, each time a new instance is entered in the repository, the reputation of its source should be recalculated and the Qprov of all its instances could change. This is a desired effect given that the provenance of a source is not static. A previously good source could diminish its reputation if all its recent instances have low quality. In order to compromise between having an up-to-date value of reputation against the calculation load in the system, the recalculation could be performed at fixed interval of time or instances inserted.

3.8 Metrics Consideration and Limitations

The proposed metrics are not presented as a comprehensive or definite set, but should be seen as a first step for the automatic evaluation of metadata quality. As such, they have several common characteristics and limitations:

- The metrics are standard-agnostic and can be used for a wide range of digital repositories such as digital libraries, learning object repositories or museum catalogs. These metrics are easy to implement in real environments and fast enough to be applied to each metadata instance at indexing or transformation time. Most of these metrics only need the information contained in the metadata instance to be calculated.
- The metrics calculations are also independent of the specific community of practice being served. However, the parameters needed to initialize the calculations heavily depends of the particularities of each group of users, because quality itself is context dependent.
- The proposed metrics are mainly designed to work over text and numbers. Given that most metadata is some form of alphanumeric value, this metrics could be applied ”as is” for the majority of metadata formats currently in use. However, if multimedia information is added to the metadata record, for example, the thumbnail of an image, new approaches based on Multimedia Information Retrieval should be used to extract a similar level of information from those multimedia fields.
- The proposed metrics are designed to estimate the quality of instances that conform with a relatively stable metadata schema. They are not suited to measure the quality of ad-hoc collections as metadata, such as the ones expressed in RDF for Semantic Web collections.
- While desirable, the normalization of the metrics is sometimes not possible. There is not a maximum value of quality for some metrics. For example, Qcomp has a natural maximum value (all the fields are filled) that can be normalized to 1, however, Qtinfo measure the amount of information present in the metadata instance. It is difficult to assign a maximum to the amount of information, as you can always find an instance with more information. The difficulty of un-normalized metrics is just important for humans. Once the metrics are included in machine learning calculation model such as RankNet [33], the coefficients self-adjust to weight the contribution of each metric, even if the values are not normalized.
- The mix of these quality parameters generate the general quality of the metadata instance. However, how they actually mix is not currently known. There exist several tradeoffs between different quality characteristics. For example, a record made complete filing it with default values could decrease its accuracy. This topic, however, is outside the scope of this paper. Further discussion on this issue can be found in [15] and [45].

4 Evaluation of the Quality Metrics

Three validation studies were conducted in order to evaluate the metrics proposed in the previous section. The first study measures the correlation between the value of the quality metrics and the quality assessment by human reviewers. The second study applies the metrics to two different sets metadata in order to establish their discriminatory power. The third study tests the metrics in a more realistic application: filtering bad quality metadata instances. These studies, along with the analysis of their results, are presented in the following subsections.

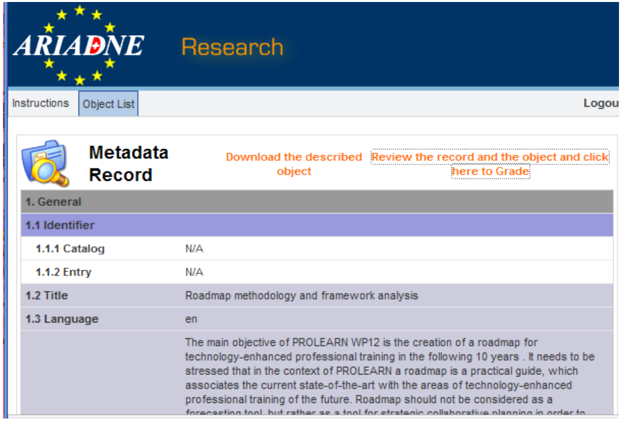
4.1 Quality Metrics correlation with Human-made Quality Assessment

A validation study was designed to evaluate the level of correlation between the quality metrics presented above and the quality assessment scores provided by human reviewers. During the study, several human subjects graded the quality of a set of instances sampled from the ARIADNE Learning Object repository [12]. We selected metadata instances about objects on Information Technologies that were available in English. From this universe (425 instances), we randomly selected 10 instances that were manually generated and 10 with metadata generated by an automated indexer. Each manual instance was produced by the author of the object (in this study, each metadata instance had a different author). The automatic metadata instances were produced by SAmgI [26]. The original objects, from which these metadata were automatically generated, are a set of Project deliverables that explain internal technologies of ARIADNE. An example of the sampled instances has been presented as examples in section 3.

Following a common practice to reduce the subjectivity in the evaluation of the quality of metadata, we used the same evaluation framework described by Bruce & Hillman on which the metrics are based. A brief explanation of this framework can be found in section 2. The reviewers had to grade the completeness, accuracy, provenance, conformance to expectations, consistency and coherence, timeliness and readability of the metadata instances.

The study was carried out online using a web application. After being trained in how to use the quality framework, each reviewer was presented with a list of the 20 selected objects in no specific order (automatic and manual generated instances were mixed). When the user selected an object, a representation of its IEEE LOM instance was displayed. The user then downloaded the referred object for inspection. Once the user had reviewed the metadata and the object, he was asked to provide grades in a 7-point scale (From "Extremely low quality" to "Extremely high quality") for each one of the seven parameters. A screen capture of the application can be seen in Figure 4.

Fig. 4 Screen were the reviewer is presented with the metadata of the object, the option to download and to rate its quality



Metadata Record

Download the described object

Review the record and the object and click here to Grade

1. General	
1.1 Identifier	
1.1.1 Catalog	N/A
1.1.2 Entry	N/A
1.2 Title	Roadmap methodology and framework analysis
1.3 Language	en
	The main objective of PROLEARN WP12 is the creation of a roadmap for technology-enhanced professional training in the following 10 years . It needs to be stressed that in the context of PROLEARN a roadmap is a practical guide, which associates the current state-of-the-art with the areas of technology-enhanced professional training of the future. Roadmap should not be considered as a forecasting tool, but rather as a tool for strategic collaborative planning in order to

Only participants that graded all the objects were considered in the study. The online application was available for 2 weeks. During that time, 22 participants completed successfully the review of all the 20 objects. From those 22, 17 (77%) work with metadata as part of their study/research activities; 11 (50%) were undergraduate students in their last years, 9 (41%) were postgraduate students and 2 (9%) had a Ph.D. degree. The participants belong to 3 different, and geographically distant, research & development groups. All of them had a full understanding of the nature and meaning of the examined objects and their metadata, and had a working knowledge of the evaluation framework.

Parallel to the human evaluation, an implementation of the quality metrics described earlier was applied to the same set of data that was presented to the reviewers. The metrics used in the study were:

- *Completeness metric (Qcomp):* It was implemented taking as a base the complete LOM instance, as described in Equation 1.
- *Weighted Completeness metric (Qwcomp):* The alphas needed in Equation 2 were obtained from the frequency of use of the fields in searches to the ARIADNE repository as reported in Najjar et al. [28].
- *Accuracy metric (Qaccu):* It was calculated using Equation 3 to measure the semantic distance between the text extracted from the object and the title and description of the metadata instance. A LSA algorithm (SVD with S=2) was applied before obtaining the distance.

- *Categorical Information Content metric (Qcinfo):* The probability of each one of the values for different fields was extracted from all the metadata information in the ARIADNE repository. Equations 4 and 6 were used to compute the final metric.
- *Textual Information Content metric (Qtinfo):* The Inverse Document Frequency (IDF) values needed to compute Equation 8 were extracted from the corpora made with all the text from the instances of the ARIADNE repository.
- *Coherence metric (Qcoh):* The title and description of the LOM instances were contrasted to measure their semantic distance as described in Equation 12.
- *Readability metric (Qread):* Equation 14 was applied to text contained in the title and description of metadata instances.
- *Provenance metric (Qprov):* The Qav (Equation 15) was obtained from all the previous calculated metrics (Qcomp, Qwcomp, Qaccu, Qcinfo, Qtinfo, Qcoh and Qread) for each instance. Qprov was equal to Qavg for all the manual generated instances because they were created by different sources. In the case of the automatic generated instances, they all were assigned to the same source and they were assigned the same Qprov.

A limitation of the study was the constant result of some metrics. The Consistency metric (Qcons) always returned 1 because the instances did not violate any of the community or LOM rules. The Linking metric (Qlink) always returned 0 because there were no explicit nor implicit linking between the objects in the study set. Finally, the Timeliness metric (Qtime) was not calculated because there were no previous registers of the average quality (Qavg). Those metrics were excluded from the study.

Because of the inherent subjectivity in measuring quality, the first step in the analysis of the results was to estimate the reliability of the human evaluation. In this kind of study, the evaluation could be considered reliable if the variability between the grades given by different reviewers to an instance is significantly smaller than the variability between the average grades given to different objects. To estimate this difference, we use the Intra-Class Correlation (ICC) coefficient [38] which is commonly used to measure the inter-rater reliability. We calculate the average measure of ICC using the two-way mixed model, given that all the reviewers grade the same sample of objects. In this configuration, the ICC is equivalent to another widely used reliability measure, the Cronbach's alpha.

The ICC was calculated for each one of the quality parameters. The results can be seen in Table 10.

Table 10 Inter Class Correlation values for the rates provided by the human reviewers. 0.7 is the critical point for ICC

Quality Parameter	**ICC**
Completeness	0.881
Accuracy	0.847
Provenance	0.701
Conformance to Expectations	0.912
Consistency & Coherence	0.794
Timeliness	0.670
Accessibility	0.819

The results of all the parameters, except for Timeliness, are higher than the recommended threshold of 0.7. This result suggests that reviewers provided similar quality scores and that further statistical analysis may be performed with those values. Given the near miss of the Timeliness evaluation, it will only be used to calculate the average quality score, but not in further statistical analysis. Table 11 presents the average value for each parameter of the human review for 6 of the 20 instances in the sample. Higher values represent higher quality.

Table 12 presents the metrics values for the same objects presented in Table 11. For all these metrics, higher values represent higher quality. While metadata instances with high quality review present roughly higher values of the metrics, it is difficult to evaluate from these tables if the metrics are a good estimation of the manual quality review of the metadata instances. In order to provide a more appropriate evaluation of the effectiveness of the metrics, the next step in the analysis was to correlate the human quality score for each parameter with the metrics. The results are presented in Table 13. The main insight obtained is that, in general, the quality metrics do not correlate with their expected quality parameters as human rate them. For example, the Qcomp metric has a low and insignificant correlation with the completeness value. On the other hand, Qaccu has a slightly significant correlation with completeness. Moreover, Qtinfo correlates with all the human parameters. The default assumption with this kind of results should be to reject the hypothesis that the proposed metrics produce an estimation of the quality parameters proposed by Bruce & Hillman. Nevertheless, before the hypothesis is rejected, the unusual correlation of all the human scores with Qtinfo deserves a closer examination.

In a previous study by Zhu et al. [49], it was found that the Information Content of text is highly correlated to the quality of web pages as perceived by human reviewers. In this paper (Section 3.4), Qtinfo measures the Information Content of the text fields of the metadata instance. A longer, more specialized text receives

Table 11 Example of the average quality value assigned to 6 of the 20 sampled instances. The first 3 were obtained from manually generated metadata, the last 3 from automatic generated metadata

Parameter	**Manual**			**Automatic**		
	R1	R2	R3	R4	R5	R6
Completeness	2.59	3.86	3.14	3.27	2.14	3.27
Accuracy	3.36	4.27	3.86	3.73	3.23	3.86
Provenance	2.95	3.77	3.73	3.18	3.14	3.55
Conformance to Expectations	1.95	4.14	3.23	3.50	2.14	3.64
C & C	3.59	4.14	3.64	4.23	3.59	3.77
Timeliness	2.91	3.41	3.36	3.77	3.27	3.91
Accessibility	3.14	4.00	3.36	3.73	2.77	3.68
Average	**2.93**	**3.94**	**3.47**	**3.63**	**2.90**	**3.67**

Table 12 Example of the metric values assigned to 6 of the 20 sampled instances. The first 3 were obtained from manually generated metadata, the last 3 from automatic generated metadata

Metric	**Manual**			**Automatic**		
	R1	R2	R3	R4	R5	R6
Completeness (Qcomp)	0.33	0.35	0.29	0.29	0.29	0.30
Weighted Completeness (Qwcomp)	0.81	0.81	0.81	0.48	0.48	0.48
Accuracy (Qaccu)	0.96	0.93	0.97	0.97	0.99	0.98
Categorical Info Content (Qcinfo)	0.32	0.32	0.20	0.20	0.22	0.22
Textual Info Content (Qtinfo)	1.49	2.21	1.92	3.34	1.93	2.46
Coherence (Qcoh)	0.0	0.27	0.13	0.90	0.80	0.35
Readability (Qread)	32	15	40	0	30	3
Provenance (Qprov)	0.56	0.57	0.54	0.35	0.35	0.35

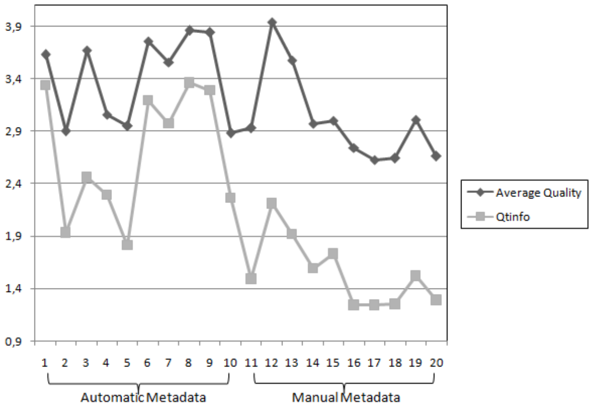
Table 13 Correlation between the human quality evaluation and the quality metrics. Bold font represents that the correlation is significant at the 0.01 level (2-tailed). Italic font represents that the correlation is significant at the 0.05 level (2-tailed).

	Qcomp	Qwcomp	Qaccu	Qcinfo	Qtinfo	Qcoh	Qread	Qprov
Completeness	.247	.537	*.519*	.011	**.787**	.282	.241	.152
Accuracy	-.370	-.421	*.492*	-.170	**.761**	.098	.270	.033
Conformance	-.290	-.533	.345	-.159	**.752**	.460	.191	-.022
C & C	-.393	-.453	*.470*	-.170	**.805**	-.083	.178	-.037
Accessibility	-.328	-.371	-.430	-.177	**.770**	.103	.334	.027
Provenance	-.437	-.473	.392	-.272	**.798**	.045	.397	-.101
Average	-.395	- *.457*	*.461*	-.182	**.842**	.225	.257	-.022

a higher score than a shorter, common one. Given that this value correlates highly with all the average human scores provided for each one of the quality parameters and that the ICC between reviewers was high, it can only be concluded that the human review was biased. This bias consists in rating instances with good textual fields with high values, even when that was not an indicated aspect of the framework quality parameter. Taking into account the diversity of the reviewer group, their knowledge in the field of metadata and that they have received instruction on how to apply the framework (and also had access to the descriptions while rating), the results suggest that non certified-expert evaluation of metadata is not a reliable method to estimate the quality of the instance in all its different dimensions.

While it can be concluded that this study is not suited to establish the "quality" of the quality metrics, it can be turned around and used to extract more information about what the reviewers took into account when rating the quality of the metadata. Firstly, a deeper analysis of the components that affect the human evaluation will be conducted. Figure 5 presents in the first 10 positions the objects with automated generated metadata. In the following 10, the objects that have their metadata manually generated. The average value for the human review is represented by the line at the top. The Qtinfo values are represented by the bottom line. The Qtinfo has higher values for the automatic generate learning objects. This result is expected because during the automatic generation process text segments contained in the objects are added to the description field. Manually generated instances, on the other hand, have small and sometimes not descriptive descriptions. Nevertheless, the quality value of human

Fig. 5 Comparison between the average quality score and the textual information content metric values)



evaluations does not decrease as sharply for manually generated metadata. There seem to be other factors that determine the human review.

A multivariate regression analysis (Stepwise) was performed including all the metrics and the origin of the metadata (1 for manual, 0 for automatic) to find possible explanations to the variability of each one of the parameters considered in the human review (except Timeliness). The results of the analysis are shown in Table 14 and explained in the following lines:

- *Completeness:* The rating behavior for the Completeness is almost fully explained ($R^2 = 0.824$) by the addition of Qtinfo (62%) and Qcomp (22%). In other words, when assigning the value for completeness of the instance, the reviewers took into account the amount and quality of text fields and the total number of filled fields.
- *Accuracy:* The rating of Accuracy is only partially explained (58%) by the Qtinfo. The Qaccu was not relevant in the model. While textual information is good to establish the general quality of the object, textual similarity cannot explain how the reviewer rated the accuracy. Factors, not considered in the calculated metrics, seem to play a major role in the rating behavior of the reviewers.
- *Conformance to Expectations:* Qtinfo seems to explain part (57%) of this parameter. As mentioned in section 2.3, Qwcomp also seems to play a role in how reviewers perceive this dimension of quality. The relatively low adjusted R^2 value (0.681) suggests that there are other factors that influence the reviewers.
- *Logical Consistency and Coherence:* Again, Qtinfo explains more than half (65%) of the variability of this parameter. It is interesting to find out that the origin of the metadata also play a small role in the model (9%). This result suggests that users found manually generated instances more coherent.
- *Accessibility:* Apart from Qtinfo contribution (59%), unexpectedly from previous discussion but logical in retrospective, the presence of some fields, measured by Qwcomp, seems to affect (14%) the accessibility rate of the metadata.
- *Provenance:* This parameter can only be partially explained (61%) by the Qtinfo. Qprov was not related with the reviewers' score.
- *Average Quality:* As it can be inferred from Figure 5, if all the parameters are averaged, the final result could be mostly estimated (80%) by the Qtinfo metric in combination with the origin of the metadata. This is consistent with the high level of correlation of Qtinfo with the value of all the quality parameters.

A final analysis that could be performed with the results of the study will be to establish whether the origin of the metadata can be deduced from the metrics. To find out, a multivariate regression (Stepwise) is performed with the metrics as independent variables and the origin as the dependent. It was found that the origin of the data can be completely deduced (Adjusted $R^2 = 0.99$) from the values of Qwcomp (90%) and Qcinfo(10%). As was found after manual inspection of the metadata instances used in the study, manual instances provide more important fields (higher Qwcomp), while automatic instances have a low variability in their categorical values, being the same for most of the objects. As a result, the origin variable used to explain some quality parameters (Consistency & Coherence and Average Quality) could be replaced by a sum of Qwcomp and Qcinfo.

The main, serendipitous, conclusion from this study is that non-expert evaluation of metadata instances, even when guided with a multidimensional metadata quality framework, is biased toward considering metadata as content. The most measurable consequence of this bias is the application of one-dimensional assessment shortcuts (in this case, quality as amount of text) as the main factor for quality estimation. While a biased human evaluation of quality could not be used to establish how the proposed metrics correlate with the different quality parameters as described by Bruce & Hillman, it offered the opportunity to measure the usefulness of the metrics to explain the rating behavior of the reviewers. Even the origin of the metadata could be deduced from the metric values.

These results also add information to the discussion about the usefulness of presenting users with complete metadata record as the main way to interact with the system. For example, Web Search engines use metadata

Table 14 Multivariate regression analysis of the quality parameters in function of the quality metrics. The Explanatory metrics specify which metrics where selected in the model (Stepwise) and their explanation power.

Parameter	Explanatory metrics	Adjusted R^2	Std. Error
Completeness	Qtinfo(62%) + Qcomp(22%)	0.824	0.2366
Accuracy	Qtinfo (58%)	0.555	0.3570
Conformance	Qtinfo (57%) + Qwcomp (14%)	0.681	0.4025
C & C	Qtinfo (65%) + origin (9%)	0.705	0.2162
Accessibility	Qtinfo (59%) + Qwcomp (14%)	0.702	0.2563
Provenance	Qtinfo (64%)	0.617	0.2501
Average Quality	Qtinfo (71%) + origin (10%)	0.798	0.2062

internally to improve the efficiency of the system, but these metadata are never shown to the user. On the other hand, most Digital Libraries try to present the user with the most complete metadata instance. While it is not the main objective of the evaluation, the results seem to indicate that this action is possibly detrimental [11]. However, more research need to be done in the area of Human Computer Interaction of Information Systems before there are strong conclusions on what is the best practice.

4.2 Quality Metrics comparison between two metadata sets

In the second study, the quality metrics were applied to two different sets of metadata to evaluate their ability to discriminate key properties of the sets. Given that there are no publicly available metadata sets of known quality, this paper select two metadata sets, that to the criteria of the authors, present a very different level of quality. The first set was composed of 4426 LOM instances corresponding to an equal number of PDF Learning Objects provided in 135 courses in Electrical Engineering and Computer Science at the MIT Open Courseware site. These metadata have been manually generated by expert catalogers in the MIT OCW team [24]. The metadata downloading was performed on January 12nd 2008. The second set of metadata was composed by LOM instances automatically generated from the same 4426 PDFs described in the first metadata set. The metadata was generated only using the Text Content Indexer of SAmgI [26] that extracted and analyzed the text from the PDF files in order to fill the LOM fields. This setup was created in order to compare the value of the metrics for a set composed of expected good quality metadata (manual metadata created by experts) against a set of expected bad quality metadata (automated metadata only based on the text of the learning object). Also, the fact that both instances refer to the same object enables the use of statistical tools to establish whether the difference between the average metric values for both sets is significant.

This study uses the same metrics used in the previous study with three important changes. Firstly, the Qlink metric was added because both the manual and the automatic metadata instances contained keywords. These keywords were used to link the instances using the procedure proposed in the first part of section 4.3.6. A considerable amount of links (130 per object in average) were obtained. Secondly, to reduce the computational time for the 8852 instances, the SVD algorithm used to calculate the semantic distance between words was replaced by the Random Projection algorithm [4] in the calculation of Qaccu and Qcoh. The Random Projection produces similar results to SVD at a fraction of the computational time [4]. Thirdly, Qprov was not calculated because the MIT OCW metadata set does not specify the author of the metadata. Moreover, the automatic generated metadata set just have one source, thus having a constant value for Qprov. Instead of Qprov, the Qavg value was obtained for each instance.

The metrics were applied to each metadata instance in both sets. Once the values were obtained, a Paired T-Test was applied to measure whether the difference between the average values was statistically significant. The average value of metrics for each metadata set as well as the result of the Paired T-Test are reported in Table 15. All the metrics have a statistically significant different average value for the two sets. Also, the values obtained for metadata instances referencing the same learning object in the manual and automatic sets are not correlated. This independence let us discard the influence that the object itself have in the metadata quality measurement.

From the Qavg values in Table 15, it can be concluded that, in general, the metrics found that the manual metadata set has higher quality than the automatic metadata set. This corroborates the hypothesis raised at the setup. A closer examination of the average of each quality metric reveals more information about the differences between both sets. The Completeness (Qcomp) and Weighted Completeness (Qwcomp) metrics point

Table 15 Metric values for the Manual and Automatic metadata sets, the correlation between the values for a same instance and the result of the comparison of means using the Paired T-Test. In bold, the highest quality average for each metric.

	Average Metric Value			
Metric	**Manual**	**Automatic**	**Correl.**	**Paired T-Test (2-tailed)**
Qcomp	**0.49**	0.38	0.073	t=344,df=4425,Sig=.000
Qwcomp	**0.75**	0.41	0.182	t=232,df=4425,Sig=.000
Qaccu	0.59	**0.90**	0.191	t=107,df=4425,Sig=.000
Qcinfo	**0.93**	0.16	0.142	t=432,df=4425,Sig=.000
Qtinfo	**6.14**	5.9	0.029	t=10,df=4425,Sig=.000
Qcoh	**0.40**	0.26	-0.024	t=8,df=4425,Sig=.000
Qlink	0.22	**0.24**	0.103	t=3.5,df=4425,Sig=.001
Qread	**0.26**	0.11	-0.014	t=4.5,df=4425,Sig=.000
Qavg	**0.66**	0.47	0.115	t=210,df=4425,Sig=.000

that human experts filled more fields (and also more important fields) that the SamgI Text Content Indexer. This is an expected result given the limited amount of information that can be extracted by simple text analysis algorithms.

The automatic set has a better average value of the Accuracy (Qaccu) metric. This, however, does not mean that automatic metadata is more accurate that the manual one, but it is attributable to a measuring artifact. Qaccu is calculated measuring the semantic distance between the text in the metadata instance and the text in the original object. The fact that all the text in the automatic metadata instances is directly extracted from the object's text explains the high value of Qaccu for the automated metadata set.

Another expected result is that humans tend to select a richer set of categorical values than the simple automated algorithm. This is reflected in the average values of the Categorical Information Content (Qcinfo) metric. For example, where the Learning Resource type value for all the learning object is set to "narrative text" in the automated instances, the human experts classify the same objects as "problem statement", "lecture", "questionnaire", "slide", etc. When all the objects in the set have the same value, Qcinfo tend to be low.

An interesting result from the comparison is that the Textual Information Content (Qtinfo) of both sets is high and very similar. That means that both instances, manual and automatic, contain long (and useful) descriptions. The manual ones were generated by humans; the automatic ones were obtained from text fragments of the original document. This finding implies that both metadata sets could have a similar level of performance (or quality) in learning object search engines that are based on text search in the metadata content. Also, as found in the previous studies, humans will be satisfied with the automatic metadata instances, given that they provide good text descriptions.

The Coherence (Qcoh) and Readability (Qread) metrics are higher also for the manual metadata sets. Text written by humans is bound to be easier to read and more coherent than text fragments automatically obtained from the learning object itself. Also, the coherence between the title and the description in the automatic set is expected to be low because the automatic algorithm takes the title contained in the PDF metadata as the value for the Title field. Normally this title in the PDF metadata is just the name of the file.

Finally, another interesting result is the almost tie in the Linkage metrics (Qlink). That result implies that the keywords manually added to the instances, and the keywords automatically generated have the same capability to link instances among them. This capability could be useful in search engine that use keywords as a way to discover new material (similar to the practice to use tags to link content).

This study comparing the quality metrics values ofdatasets two different metadata sets confirms the ability of the metrics to measure quality characteristics in the instances. Differences expected from a-priori knowledge of the origin of the datasets were discovered as differences in the quality metrics values. Also, the study served to test the feasibility of applying the quality metrics to a relatively large set of instances.

4.3 Quality Metrics as automatic low quality filter

A final study was setup to test the metrics in a more realistic task. This task is to automatically filter or identify low quality instances inside a collection. It is expected that lower quality instances get lower metric values. To test this hypothesis, human reviewers were asked to select the lowest quality instance (according to different quality dimensions) from a given set of instances. These instances belong to different ranges of the corresponding metric value. At the end the human selections were compared with the metric value to es-

Fig. 6 Range explanation. 4 ranges were selected from the quality metric value to indicate 4 groups (R1, R2, R3 and R4) of increasing metric value

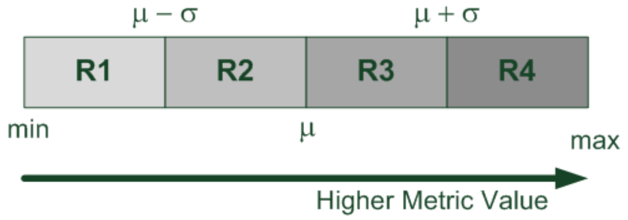


Table 17 Consistency percentage for each Comparison Sets

Comparison Set	Review Consistency
Completeness (Qcomp)	100%
Weighted Completeness (Qwcomp)	90%
Accuracy (Qaccu)	70%
Categorical Info Content (Qcinfo)	70%
Textual Info Content (Qtinfo)	100%
Coherence (Qcoh)	70%
Readability (Qread)	80%
Total (Qavg)	90%

tablish whether the ones with the lowest values were selected as the worst instances.

The instances for this study were selected from the manual and automatic sets used in the previous study. Four ranges were created for each metric value. The ranges were delimited by the mean, the mean minus one standard deviation and the mean plus one standard deviation. Figure 6 represents graphically these ranges. R1 represents the instances with the lowest metric value, while R4 contains the instances with the highest metric value.

All the metrics (Comp, Qwcomp, Quack, Qcinfo, Qcoh, Qread) used in the previous study with exception of Qlink were considered also for this study. Qlink was removed because humans cannot evaluate how connected an instance without access to the whole repository. Qavg was again used, and was calculated as the combination of all the proposed metrics (including Time).

For each metric 10 comparisons were generated. Five comparisons were drawn from the manual metadata set and five from the automated metadata set. Each comparison contains four instances. Each of those four instances was selected randomly from a different range of the metric. Once generated, each comparison has an instance from each one of the four ranges. Each comparison was presented to four human reviewers. The four reviewers were assigned to each comparison in alternate order to balance the effect of subjective review. Eight reviewers participated in the study. All of them were graduated research assistants that work with metadata as part of their research. When presented with a comparison, the reviewer had to select the lowest quality instance according to a given directive, different for each metric. The directives for each metric are presented in Table 16.

Once the results of the comparisons were collected, the first step was to determine whether there was consistency in the selections performed by the reviewers. An object was consistently selected if at least three of the four reviewers had selected it. Table 17 shows the consistency percentage for each one of the metrics. Given that for all the metrics obtained more than 70% of consistency (only in three or less comparisons there was not a majority for a given instance) it can be concluded that the noise present due to different criteria of each reviewer was low and the results could be used as a good approximation of what human reviewers would select as low quality instances.

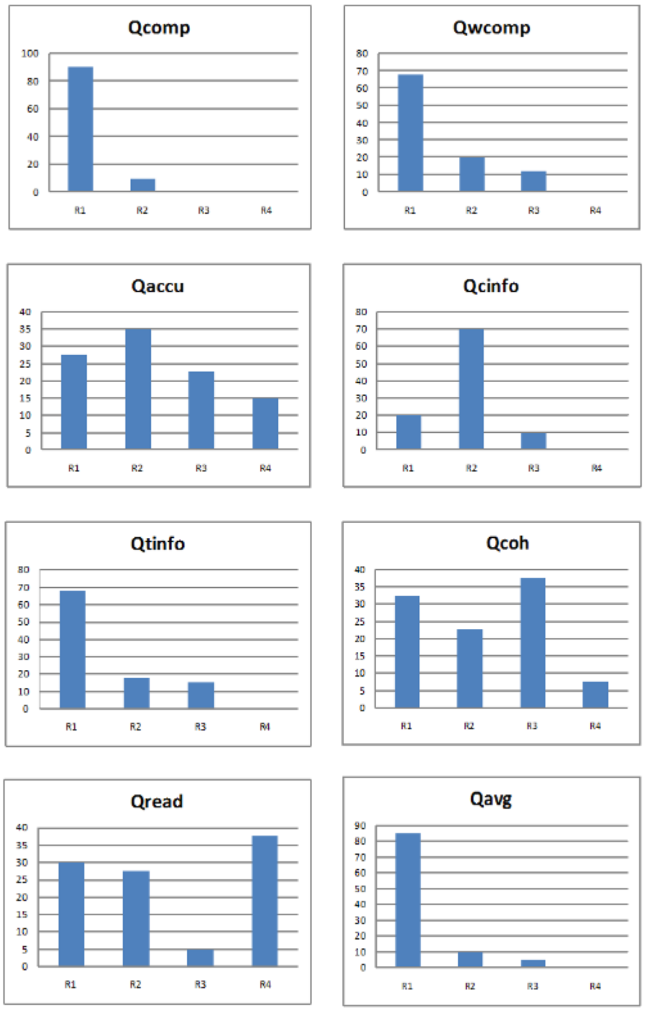
The next step in the result analysis was to compare the selections performed by the reviewers against the value of each metric. Figure 7 presents the percentage of times that an object in each range was selected as the lower quality instance for each one of the metrics. Three metrics (Qcomp, Qwcomp and Qtinfo) seem to agree with the human selection. The majority of the reviewers selections for these metrics took place in the R1 range. A decreasing number of selections can still be seen in R2 and R3. No instances with a high value (R4) in those metrics were selected by any of the human reviewers. On the other hand, Qaccu, Qcoh and Qread do not seem to correlate well with human selections. A non-conclusive distribution could be seen across the four ranges. This is an indication that the metrics are not measuring the same quality characteristics as humans interpret from the given directives. An exceptional case is Qcinfo. Here, there is a clear preference of the reviewers for the R2 range. A deeper analysis of the metrics and human selection suggests that those instances in R1 do miss several categorical fields while the ones in R2 have those fields, but are filled with very common values. It seems that Human reviewers seem do not take into account missing values when evaluating the descriptive power of the instance. Finally, the combination of the metrics, Qavg, seems to be well related to what human reviewers consider as the general quality of the instance.

As a final analysis, Table 18 presents the effectiveness percentage (percentage of times that a value in the R1 range won the reviewers vote in a comparison). This value amounts to the percentage of times that the metric would have agreed with the human selection if it was meant to be an automatic filter to discard low

Table 16 Directives given to reviewers to select the lowest quality instance according to each metric

Metric	Directive
Completeness (Comp)	Select the instance that presents less information
Weighted Completeness (Qwcomp)	Select the instance that present less useful information
Accuracy (Qaccu)	Select the less accurate instance (original object supplied)
Categorical Info Content (Qcinfo)	Select the less descriptive instance
Textual Info Content (Qtinfo)	Select the less descriptive instance
Coherence (Qcoh)	Select the instance with less internal coherence
Readability (Qread)	Select the instance that is less readable
Total (Qavg)	Select the lowest quality instance

Fig. 7 Distribution of human selection of lowest quality instances among Ranges of the Quality Metrics. R1 are the lowest metric values and R4 are the highest metric values.



quality instances. The most important finding in this analysis is that the Qavg metric (the combination of all other metrics) would have flagged 9 out of 10 instances selected as the lowest quality by the majority of human reviewers. Table 18 also presents the effectiveness percentage considering only the manual and the automated set. From these values, it can be concluded that the source of the metadata does not affect the effectiveness of the metrics.

The results of this study strongly suggest that some of the metrics (Qcomp, Qwcomp and Qtinfo), as well as the combination of all the proposed metrics (Qavg), can be used to build an automated quality filter system for metadata instances. This system could also take the form of a metadata expert assistant that flag the most problematic instances to guide cleaning or enrichment processes. This type of system is presented as an application of the metrics in the next section.

4.4 Studies Conclusions

From the three validation studies performed to evaluate the quality metrics, several conclusions could be drawn:

- Human reviewers tend to agree when evaluating the quality of metadata. However, it is no so clear which dimensions of quality they evaluate, even if they are guided by a framework as in the first study or guidelines as in the third one. From the results of the first study it seems that when confronted with the metadata, the reviewers evaluated it as content.
- Some metrics correlate well with human reviews while others seems to be completely orthogonal. From all the proposed metrics, the Textual Information Content (Qtinfo) seems to be a good approximation of the human perceived quality of an instance (the metadata as content effect). In a surprising result, given that half of the metrics did not correlate with human evaluation, Qavg, the combination of all the proposed metrics does seem to agree with the reviewers selections.
- There are quality characteristics that human reviewers are not able to evaluate. The variability of the categorical values or the level of connection of the instances, where the reviewer needs to have information about the whole universe of instances, are

Table 18 Effectiveness percentage for each metric. This indicate the percentage of times that the metric agreed with the human most voted instance. It also presents the percentage disaggregated for the Manual and Automated Metadata sets.

	Effectiveness		
Metric	**General**	**Manual**	**Automated**
Completeness (Qcomp)	90%	100%	80%
Weighted Completeness (Qwcomp)	70%	80%	60%
Accuracy (Qaccu)	30%	20%	40%
Categorical Info Content (Qcinfo)	20%	40%	0%
Textual Info Content (Qtinfo)	80%	80%	80%
Coherence (Qcoh)	40%	40%	40%
Readability (Qread)	30%	20%	40%
Total (Qavg)	90%	100%	80%

specially difficult to evaluate manually. In this sense, the quality metrics, even the ones that did not correlate well with the human evaluation, were able to measure characteristics related to the quality of the two different metadata sets in the second study.

- The usefulness of the combination of the proposed quality metrics in at least one practical application, low quality metadata filtering, was strongly suggested by the results of the third study. This set of metrics is indeed a step forward the automatic evaluation of metadata quality in digital repositories.

5 Implementation and Applications of Metadata Quality Metrics

The most important aspect of the metrics proposed in this work is that they can be automatically calculated from the metadata present in the repository and the digital objects being described. The result of the metrics can then be used in tools that generate metadata (manually or automatically) to provide an automatic quality estimation of each metadata instance that is produced. Also, the value of the metrics for a whole repository, or federation of repositories, can be used in quality assurance applications that allow an administrator to identify quality problems in order to take corrective actions.

The metrics can be used by applications that generate metadata (to provide a quality control over each produced instance), applications that analyze the indexing behavior of metadata producers or applications that search for low quality instances in order to correct them. Some examples of that type of applications that can benefit from the quality metrics include:

- *Automatic validation and correction of metadata.* While previous research suggests that automatic metadata generation has a similar quality level as human generated metadata [26], the main objection against automatic generation of metadata is how to provide it with some degree of quality assurance [31]. Metadata extraction mechanisms work most of the time, but sometimes they produce useless instances. Without quality assurance, those low quality instances will be mixed with the whole repository, decreasing its overall value. Manually reviewing the output of an automatic generator is an unfeasible task. The metadata quality metrics proposed in this paper could be used to implement an automatic evaluator of metadata that can flag low quality instances. For example, instances that do not contain a meaningful description or whose title is not coherent with the description can be flagged before they are inserted into the repository. On the other hand, if the automatic evaluator of metadata is run over human generated metadata, it could guide an automatic generator of metadata to improve the content of low quality instances. For example, metadata instances that lack a description could be improved with an automatic summary created by an automatic generator from the textual content of the resource.
- *Visualization of repository-wide quality.* The metrics values can be used to create visualizations of the repository in order to gain a better under-standing of the distribution of the quality problems. For example, a treemap visualization [3] could be used to find answers to different questions: Which authors or sources of metadata cause quality problems? How has the quality of the repository evolved over time? Which is the most critical problem of the metadata in the repository?, etc. An example of such visualization is shown in Figure 8. The treemap represents the structure of the ARIADNE repository. The global repository contains several local repositories and different authors publish metadata in their local repository. The boxes represent the set of learning objects metadata instances published by a given author. The color of the boxes represents the average of the Qtinfo metric score of that set of instances. The

color scale goes from red/dark grey (low quality) to yellow (medium quality), to green/light grey (high quality). This visualization helps to easily spot authors that provide good textual descriptions to their objects. Figure 9 shows the same type of visualization, but indicating incomplete instances (Qwcomp) in the human generated metadata from MIT OCW used during the studies. In this case, finding the incomplete instances would have been difficult without the help of the visualization tool.

Fig. 8 Visualization of the Textual Information Content of the ARIADNE Repository. Red (dark) boxes indicate authors that produce low quality descriptions.

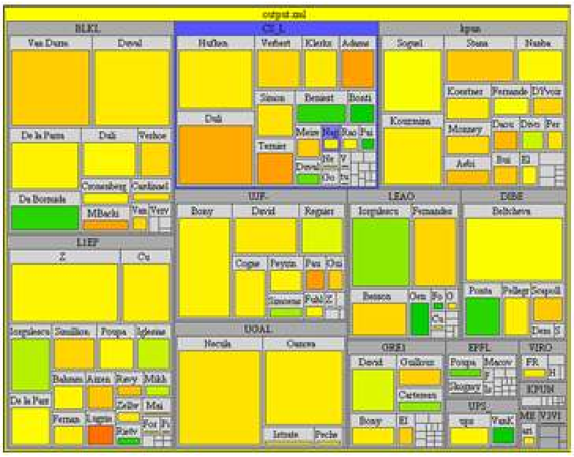
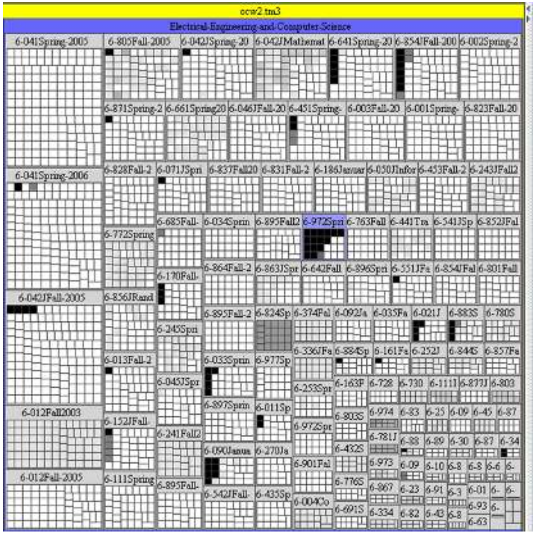


Fig. 9 Visualization of the Completeness of the Manual Metadata set extracted from MIT OCW. Dark boxes represents instances that are incomplete.



– *(Automatic) Selection of repositories for federated search.* If the repositories belonging to a federation publish their results for the quality metrics, that information can be used by federated search engines to automatically select repositories with a quality similar or superior to the local repository. Also, depending on the task to perform, the engine could choose to return only instances that have a good textual description of the resource. An initial implementation of this kind of application has already been devised by Hughes [19] to provide a "star-ranking" for repositories of the Open Language Archive but based mostly on completeness metrics.

6 Related Work

As shown in sections 1 and 2, there is extensive conceptual research in Information Quality and more specifically Metadata Information Quality. On the other hand, automatic calculation of metrics to estimate quality of metadata is much rarer. To our knowledge, only Stvilia et al. in [42] and [43] seriously address the issue of multidimensional metadata quality estimation based on automatic calculations. Their metrics are also based on a 9 quality parameter framework: Intrinsic Precision, Intrinsic Redundancy, Intrinsic Semantic Consistency, Intrinsic Structural Consistency, Relational Accuracy, Relational Completeness, Relational Semantic Consistency, Relational Structural Consistency and Relational Verifiability. In [42], Stvila presents 13 quality metrics. While most of them (11) are simple counts of errors or defects (for example: # of broken links, # of words not recognized by MS word dictionary over the total number of words, etc.), the remaining two, Information Noise and Kullback - Leibler Divergence have some relation with our Qtinfo and Qcinfo metrics respectively. Due to the lack of reference quality information, Stvilia was not able to evaluate his metrics directly. What he found is that the metrics correlate with a-priori knowledge of two different sources of metadata (similar to what have been done in the second study of this paper). A comparison analysis over a common set of metadata could be an interesting subject for further research.

7 Further Work

This work is a first step towards the automatic evaluation of digital repositories metadata. Some open and interesting research topics not addressed in this paper are:

– *New metadata quality frameworks oriented to automatic processing and completely dynamic metadata*

Current metadata quality frameworks are deeply rooted in traditional, analog metadata. This metadata was meant to be consumed by humans and thus the quality characteristics considered in the frameworks were the ones that humans found important. Now, the metadata is mainly consumed and processed by automated software systems. It could be created or modified with each human interaction (corrections, annotations, tags, reviews, etc.) with the system. While it preserves some relation with its analog counterpart, digital metadata could not be measured with the same standards. Also, new metadata approaches, such as Semantic Web, eliminate the idea of a formal metadata schema. Current frameworks needs the idea of schema in order to work. New quality frameworks oriented to digital metadata, automatic processing and more dynamic metadata schemas should be developed.

– *Establishing a common data set.* Borrowing the idea that initiate the TREC conference [18] and in order to provide a better "measurement" of the quality of different metrics, the quality metrics should be applied to a known set of test metadata instances with an established and known value for different dimensions of quality. This is especially important to provide common ground to metrics proposed by different researchers. When applied to a common set, the prediction power of the metrics could be objectively compared and its progress measured.

8 Conclusions

Although quality of metadata for digital repositories is a very difficult concept to measure as a whole, when divided into more concrete parameters, as the ones proposed by several quality frameworks, quality can be operationalized in the form of quality metrics. These metrics, while simple to calculate, could be effective estimators of quality. In this work, some of the proposed metrics, especially Textual Information Content metric and the combination of metrics (Qavg), were able to explain the quality rating behavior of human reviewers, discriminate between different sets of metadata and even automatically flag low quality instances as good as any human reviewer.

The development of quality metrics will enable metadata quality researchers to not only obtain snapshots of the quality of a repository, but also to constantly monitor its evolution and how different events affect it without the need to run costly human-involving studies. This could lead to the creation of innovative applications based on metadata quality that would improve the final user experience.

The proposed metrics are not presented as an optimal solution to the problem of automatic evaluation of quality, but they can be used as a baseline against which new, better, metrics could be compared. While a lot more research and experimentation in metadata quality metrics is needed, this paper shows that automatic quality assurance based on metrics is possible. Moreover, automatic evaluations have to be provided in order to sustain the increase in the metadata production. That is the only way for current digital repositories to avoid degradation of their functionality.

References

1. Barton, J., Currier, S., Hey, J.M.N.: Building quality assurance into metadata creation: an analysis based on the learning objects and e-prints communities of practice. In: S. Sutton, J. Greenberg, J. Tennis (eds.) Proceedings 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice - Metadata Research and Applications, pp. 39–48. Seattle, Washington (2003)
2. Beall, J.: Metadata and data quality problems in the digital library. JoDI: Journal of Digital Information **6**(3), 20 (2005)
3. Bederson, B.B., Shneiderman, B., Wattenberg, M.: Ordered and quantum treemaps: Making effective use of 2D space to display hierarchies. ACM Trans. Graph. **21**(4), 833–854 (2002)
4. Bingham, E., Mannila, H.: Random projection in dimensionality reduction: applications to image and text data. In: F. Provost, R. Srikant (eds.) Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 245–250. ACM Press, New York, NY, USA (2001)
5. Bruce, T.R., Hillmann, D.: Metadata in Practice, chap. The continuum of metadata quality: defining, expressing, exploiting, pp. 238–256. ALA Editions, Chicago, IL (2004)
6. Bui, Y., ran Park., J.: An assessment of metadata quality: A case study of the national science digital library metadata repository. In: H. Moukdad (ed.) Proceedings of CAIS/ACSI 2006 Information Science Revisited: Approaches to Innovation, p. 13 (2006)
7. Cardinaels, K., Meire, M., Duval, E.: Automating metadata generation: the simple indexing interface. In: WWW '05: Proceedings of the 14th international conference on World Wide Web, pp. 548–556. ACM Press, New York, NY, USA (2005)
8. Chapman, A., Massey, O.: A catalogue quality audit tool. Library Management **23**(6-7), 314–324 (2002)
9. DCMI: Dublin Core Metadata Innitiative, http://dublincore.org, retrieved 2/04/2007 (1995)
10. Dushay, N., Hillmann, D.: Analyzing metadata for effective use and re-use. In: S. Sutton, J. Greenberg, J. Tennis (eds.) DCMI Metadata Conference and Workshop, p. 10. Dublin Core Metadata Initiative, Seattle, USA (2003)
11. Duval, E., Hodgins, W.: Making metadata go away: Hiding everything but the benefits. In: Proceedings of the DCMI 2004 conference, pp. 29–35. Dublin Core Metadata Initiative, Shanghai, China (2004)
12. Duval, E., Warkentyne, K., Haenni, F., Forte, E., Cardinaels, K., Verhoeven, B., Van Durm, R., Hendrikx, K., Forte, M., Ebel, N., et al.: The ariadne knowledge pool system. Communications of the ACM **44**(5), 72–78 (2001)

13. Ede, S.: Fitness for purpose: The future evolution of bibliographic records and their delivery. Catalogue & Index **116**, 1–3 (1995)
14. Foltz, P.W., Kintsch, W., Landauer, T.K.: The measurement of textual coherence with latent semantic analysis. Discourse Processes **25**, 285–307 (1998)
15. Foulonneau, M.: Information redundancy across metadata collections. Information Processing and Management: an International Journal **43**(3), 740–751 (2007)
16. Greenberg, J., Pattuelli, M.C., Parsia, B., Robertson, W.D.: Author-generated dublin core metadata for web resources: A baseline study in an organization. In: K. Oyama, H. Gotoda (eds.) DC '01: Proceedings of the International Conference on Dublin Core and Metadata Applications 2001, pp. 38–46. National Institute of Informatics (2001)
17. Guy, M., Powell, A., Day, M.: Improving the quality of metadata in eprint archives. Ariadne **38**, 5 (2004)
18. Harman, D.: Overview of the first trec conference. In: R. Korfhage, E.M. Rasmussen, P. Willett (eds.) SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 36–47. ACM Press, New York, NY, USA (1993)
19. Hughes, B.: Metadata quality evaluation: Experience from the open language archives community. In: Z. Chen, H. Chen, Q. Miao, Y. Fu, E. Fox, E. Lim (eds.) Digital Libraries: International Collaboration and Cross-Fertilization: Proceedings of the 7th International Conference on Asian Digital Libraries, ICADL 2004, pp. 320–329. Springer Verlag, Shangay, China (2004)
20. Hughes, B., Kamat, A.: A metadata search engine for digital language archives. D-Lib Magazine **11**(2), 6 (2005)
21. IEEE: IEEE 1484.12.1 Standard: Learning Object Metadata, http://ltsc.ieee.org/wg12/par1484-12-1.html, retrieved 2/04/2007 (2002)
22. Landauer, T., Foltz, P., Laham, D.: An introduction to latent semantic analysis. Discourse Processes **25**(2-3), 259–284 (1998)
23. Liu, X., Maly, K., Zubair, M., Nelson, M.L.: Arc - an oai service provider for digital library federation. D-Lib Magazine **7**(4), 12 (2001)
24. Lubas, R., Wolfe, R., Fleischman, M.: Creating metadata practices for mit's opencourseware project. Library Hi Tech **22**(2), 138–143 (2004)
25. McCallum, D.R., Peterson, J.L.: Computer-based readability indexes. In: W.J. Burns, D.L. Ward (eds.) ACM 82: Proceedings of the ACM '82 conference, pp. 44–48. ACM Press, New York, NY, USA (1982)
26. Meire, M., Ochoa, X., Duval, E.: Samgi: Automatic metadata generation v2.0. In: C.M..J. Seale (ed.) Proceedings of the ED-MEDIA 2007 World Conference on Educational Multimedia, Hypermedia and Telecommunications, 1195-1204. AACE, Chesapeake, VA (2007)
27. Moen, W.E., Stewart, E.L., McClure, C.R.: Assessing metadata quality: Findings and methodological considerations from an evaluation of the u.s. government information locator service (gils). In: T.R. Smith (ed.) ADL '98: Proceedings of the Advances in Digital Libraries Conference, pp. 246–255. IEEE Computer Society, Washington, DC, USA (1998)
28. Najjar, J., Ternier, S., Duval, E.: The actual use of metadata in ariadne: an empirical analysis. In: E. Duval (ed.) Proceedings of the 3rd Annual ARIADNE Conference, pp. 1–6. ARIADNE Foundation (2003)
29. Najjar, J., Ternier, S., Duval, E.: User behavior in learning objects repositories: An empirical analysis. In: L.C..C. McLoughlin (ed.) Proceedings of the ED-MEDIA 2004 World Conference on Educational Multimedia, Hypermedia and Telecommunications, pp. 4373–4378. AACE, Chesapeake, VA (2004)
30. Newman, M., Watts, D., Barabsi, A.L.: The Structure and Dynamics of Networks. Princeton University Press (2006)
31. Ochoa, X., Cardinaels, K., Meire, M., Duval, E.: Frameworks for the automatic indexation of learning management systems content into learning object repositories. In: P. Kommers, G. Richards (eds.) Proceedings of the ED-MEDIA 2005 World Conference on Educational Multimedia, Hypermedia and Telecommunications, pp. 1407–1414. AACE, Chesapeake, VA (2005)
32. O'Neill, E.T.: Frbr: Functional requirements for bibliographic records; application of the entity-relationship model to humphry clinker. Library Resources & Technical Services **46**(4), 150–159 (2002)
33. Richardson, M., Prakash, A., Brill, E.: Beyond pagerank: machine learning for static ranking. In: C. Goble, M. Dahlin (eds.) Proceedings of the 15th international conference on World Wide Web, pp. 707–715. ACM Press, New York, NY (2006)
34. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management: an International Journal **24**(5), 513–523 (1988)
35. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**(11), 613–620 (1975)
36. Shannon, C., Weaver, W.: The Mathematical Theory of Communication. University of Illinois Press (1963)
37. Shreeves, S.L., Knutson, E.M., Stvilia, B., Palmer, C.L., Twidale, M.B., Cole, T.W.: Is "quality" metadata "shareable" metadata? the implications of local metadata practices for federated collections. In: H.A. Thompson (ed.) Currents And Convergence: Navigating the Rivers of Change: Proceedings of the Twelfth National Conference of the Association of College and Research Libraries, pp. 223–237. ALA, Minneapolis, USA (2005)
38. Shrout, P., Fleiss, J.: Intraclass correlations: uses in assessing rater reliability. Psychol Bull **86**, 420–428 (1977)
39. Simon, B., Massart, D., van Assche, F., Ternier, S., Duval, E., Brantner, S., Olmedilla, D., Miklos, Z.: A simple query interface for interoperable learning repositories. In: D. Olmedilla, N. Saito, B. Simon (eds.) Proceedings of the 1st Workshop on Interoperability of Web-based Educational Systems, pp. 11–18. CEUR, Chiba, Japan (2005)
40. Van de Sompel, H., Nelson, M., Lagoze, C., Warner, S.: Resource harvesting within the oai-pmh framework. D-Lib Magazine **10**(12), 1082–9873 (2004)
41. Strong, D.M., Lee, Y.W., Wang, R.Y.: Data quality in context. Communications of the ACM **40**(5), 103–110 (1997). URL citeseer.ist.psu.edu/strong97data.html
42. Stvilia, B.: Measuring information quality. Ph.D. thesis, University of Illinois at Urbana - Champaign, Urbana, IL (2006)
43. Stvilia, B., Gasser, L., Twidale, M.: Information quality management: theory and applications, chap. Metadata quality problems in federated collections, pp. 154–18. Idea Group, Hershey, PA (2006)
44. Stvilia, B., Gasser, L., Twidale, M.: A framework for information quality assessment. Journal of the American Society for Information Science and Technology **58**(12), 1720–1733 (2007)
45. Stvilia, B., Gasser, L., Twidale, M.B., Shreeves, S.L., Cole, T.W.: Metadata quality for federated collections. In: I.N. Chengalur-Smith, L. Raschid, J. Long, C. Seko (eds.) IQ, pp. 111–125. MIT (2004)
46. Thomas, S.E.: Quality in bibliographic control. Library Trends **44**(3), 491–505 (1996)
47. Verbert, K., Jovanovic, J., Gasevic, D., Duval, E.: Repurposing learning object components. In: R. Meersman, Z. Tari,

P. Herrero (eds.) On the Move to Meaningful Internet Systems 2005: OTM Workshops, *Lecture Notes in Computer Science*, vol. 3762, pp. 1169–1178. Springer Berlin / Heidelberg, Agia Napa, Cyprus (2005)

48. Wilson, A.J.: Toward releasing the metadata bottleneck - a baseline evaluation of contributor-supplied metadata. Library Resources & Technical Services **51**(1), 16–28 (2007)
49. Zhu, X., Gauch, S.: Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In: E. Yannakoudakis, N.J.B.M.K. Leong, P. Ingwersen (eds.) Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 288–295. ACM Press, New York, NY (2000)