Towards Automatic Evaluation of Learning Object Metadata Quality

Xavier Ochoa[1] and Erik Duval[2]

[1]Escuela Superior Politcnica del Litoral (ESPOL), Via Perimetral Km. 30.5, Guayaquil, Ecuador
xavier@cti.espol.edu.ec
[2]Computerwetenschappen Dept., Katholieke Universiteit Leuven, Celestijnenlaan 200 A, B-3001 Leuven, Belgium
erik.duval@cs.kuleuven.be

Abstract. Thanks to recent developments on automatic generation of metadata and interoperability between repositories, the production, management and consumption of learning object metadata is vastly surpassing the human capacity to review or process these metadata. However, we need to make sure that the presence of some low quality metadata does not compromise the performance of services that rely on that information. Consequently, there is a need for automatic assessment of the quality of metadata, so that tools or users can be alerted about low quality instances. In this paper, we present several quality metrics for learning object metadata. We applied these metrics to a sample of records from a real repository and compared the results with the quality assessment given to the same records by a group of human reviewers. Through correlation and regression analysis, we found that one of the metrics, the text information content, could be used as a predictor of the human evaluation. While this metric is not a definitive measurement of the "real" quality of the metadata record, we present several ways in which it can be used. We also propose new research in other quality dimensions of the learning object metadata.

1. Introduction

In the first years of learning object metadata [1] production, most of the metadata were created manually by the author of the learning object, a field expert or a librarian and inserted it in a repository where other users could search for it. This wasn't that much of a problem as long as the number of objects present in most repositories was small (less than 10000). The main research effort among the learning object community was to find ways to increase the number of available objects.

As a result of these efforts, several approaches have been developed to deal with the small number of learning objects indexed in a single repository. Information extraction techniques are now used to (semi-) automatically extract the metadata values from the learning object itself or from the context where it is created or published. Examples of systems that employ these techniques could be found at [2] and [3]. Moreover, federated search facilities across learning object repositories have been standardized [4]. In federated searches, the aggregated content of all participant re-

positories can be accessed from any one of them. The GLOBE partnership of repositories is a clear example of this tendency [5]. A third approach, disaggregating existing learning objects in their components, has been successfully researched and workable prototypes are available [6]. These new technologies have lead to an exponential growth in the number of learning objects with metadata. While this increase is solving the lack of available learning objects, it creates a new and different problem: there exists no feasible way to assure the quality of metadata, either newly produced or accessed from a repository.

The quality of the metadata record that describes a learning object affects directly the chances of the object to be found, reviewed or reused [7]. For example, a learning object automatically indexed from a Learning Management System (LMS) with the title "Lesson 1 - Course CS201", without any description or keywords will hardly appear in a search for materials about "Introduction to Java", even if the described object is, indeed, a good introductory text to Java. The object will just be part of the repository but will never be retrieved in relevant searches.

The traditional approach to evaluate the quality of learning object metadata has been the same as in the digital library world [8][9]: Manually review a statistical significant sample of records by comparing the values with those provided by metadata experts. While this human review could be useful for small-sized and slow-growing repositories, it becomes impractical for large or federated repositories or in the case of automatic indexers as humans "do not scale" [10]. Also learning object metadata is different from a library record as the first is not unique (there could be several instances for the same learning object) and evolves as the object itself is reused in different contexts.

To deal with the exponential grow of metadata records available and at the same time to be able to retain some sort of quality assurance for the information contained in the metadata record, we propose automating the quality assessment of learning object metadata. This automated evaluator will assess intrinsic characteristics of the metadata itself, measured through the use of one or more synthetic metrics. This paper builds on a previous work [11] where several quality metrics for learning object metadata are presented. In the following sections, we select and describe a group of those metrics. The values obtained from the metrics are contrasted with evaluations by human reviewers to a sample of learning object metadata from a real repository, and the results are evaluated. We propose some applications for the metrics that correlate highly with the human evaluation. Finally, in the last sections, we propose more experiments to evaluate metrics for different functions of the metadata and present related works.

2. Quality Metrics

The quality metrics are small calculation performed over the values of the different fields of the metadata record in order to gain insight in a quality characteristic. The quality characteristics in which these metrics are loosely based are the ones proposed by [12]: completeness, accuracy, provenance, conformance to expectations, consistency & coherence, timeliness and accessibility. For example, we can count the number of fields that have been filled with information (metric) to assess the completeness of the metadata record (quality characteristic). These measurements are only concern

with the quality of the learning object metadata record, not the quality of the learning object itself. [13] gives a more holistic view of the quality of the learning object as a whole (metadata information included). A more formal description of the metrics that we will use and their *rationale* has been discussed in detail in [11]. We have selected the metrics that are feasible to implement with the information available in real repositories. We present here a brief description and an example of calculation of each one of the selected metrics.

- **Simple Completeness:** It tries to measure the Completeness of the metadata record. This metric counts the number of fields that content a no-null value. In the case of multi-valued fields, the field is considered complete if at least one instance exists. The score could be calculated as a percentage of possible fields and divided by 10 to be in a scale from 0 to 10. For example, according to this metric, a record with 80% of its fields filled has a higher score (higher quality) (q=8) than one in which only the 40% has been filled (q=4).
- **Weighted Completeness**: It tries to measure the Completeness in a more meaningful way than the "Simple Completeness" metric as not all the fields are equally important for a given application. This metric not only counts the number of filled fields, but assigns a weight value to each of the fields. This weight value should reflect the importance of that field for the application. To be comparable with the Simple Completeness the obtained value should be divided by sum of all the weights and multiplied by 10. For example, if the main application of the metadata will be to provide information about the object to a human user, the title, description and annotation fields are more important than the identifier or metadata's author fields. The more important fields could have a weight of 1 while the not important records could have a weight of 0.2. In this case, a record with information for title and description only will receive a higher score (higher quality) (q=2/2.4*10=9) than a record with information for title, identifier and metadata author (q=1.4/2.4*10=6).
- **Nominal Information Content:** One of the main requirements of a metadata record is that it contains enough information to describe uniquely its referred learning object. Unique information helps the user to distinguish and evaluate different objects. This metric tries to measure the Amount of Information that the metadata posses in its nominal fields (fields that can only be filled with values taken from a fixed vocabulary). For this kind of field, the Information Content can be calculated as 1 minus the entropy of the value (the entropy is the negative log of the probability of the value in a given repository) [14]. This metric sums the information content for each categorical field of the metadata record. For example, if the difficulty level of a metadata record is set to "high", where the majority of the repository is set to "medium", it will provide more unique information to the record and, thus, a higher score (high quality). On the other hand, if the record's nominal fields only content the "default values" used in the repository, they will provide less unique information to the record and a lower score.
- **Textual Information Content:** This metric also tries to measure the Amount of Information contained in the record. It measures how much relevant and unique words are contained in the record's text fields (the fields that can be filled with free text). The "relevance" and "uniqueness" of a word is directly proportional to how often that word appears in the record and inversely proportional to how many re-

cords contain that word. This relation is handled by the TF-IDF (Term Frequency-Inverse Document Frequency) calculation [15]. The number of times that the word appears in the document is multiplied by the negative log of the number of documents that contain that word (could be considered as a weighted entropy measurement). The log of the sum of all the TF-IDF value of all the words in textual fields is the result of the metric. For example, if the title field of a record is "Lecture in Java", given that "lecture" and "java" are common words in the repository, will have lower score (lower quality) than a record in which the title is "Introduction to Java objects and classes", not only because "objects" and "classes" are less frequent in the repository, but also because the latter title contains more words.

- **Readability:** This metric tries to measure how accessible the text in metadata is. This metric applies a readability index, for example the Flesch Index [16], to assess how easy is to read the description of the learning object. The 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 is to read a text. For example, a description where only acronyms or complex sentences are used will receive a higher score (lower quality) than a description where normal words and simple sentences are used.

2.1. Evaluation of the Metrics

We designed an experiment to evaluate how the quality metrics presented above correlate with quality assessment by human reviewers. During the experiment, several human reviewers graded the quality of a set of records sampled from the ARIADNE repository [17]. We selected metadata records about objects on Information Technologies objects that were available in English in the repository. From this universe (425 records), we randomly selected 10 with metadata generated manually and 10 with metadata generated by an automated indexer. Following a common practice to reduce the subjectivity in the evaluation of the quality of metadata, we used an evaluation framework. The selected framework was the same 7 quality characteristics proposed by [12] on which the metrics are loosely based. The experiment was carried out online using a web application. After reading the instructions, the user was presented with a list of the 20 selected objects in no specific order. When the user selected an object, a representation of its LOM record 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 give grades in a 7-point scale (From "Extremely low quality" to "Extremely high quality") for each one of the 7 parameters. Only participants that graded all the objects were considered in the experiment. The experiment 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. All of them had a full understanding of the nature and meaning of the examined objects and their metadata, and were trained in the evaluation framework. Parallel to the human evaluation, an implementation of the quality metrics described before was applied to the same set of data that was presented to the reviewers. For the weighted complete-

ness calculation, the weights were obtained from the frequency of use of the fields in the ARIADNE searches [18].

2.2. Data Analysis

Because of the inherent subjectiveness in measuring quality, the first step in the analysis of the data was to estimate the reliability of the human evaluation. In this kind of experiment, the evaluation could be considered reliable if the variability between the grades given by different reviewers to a record 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 [19] over the average quality grade (the sum of the value given to each of the seven quality characteristics, divided by 7). We calculated the 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 result obtained was 0.909, much higher than the 0.7 threshold needed to be considered acceptable. In other words, the ICC suggests that the reviewers provided similar quality scores and that further statistical analysis can be performed.

The next step was to average the value of all the human reviewers for each record and correlate it with the values obtained from the calculation of the quality metrics over the same records. The results are presented in the Table 1.

Table 1. Correlation between the Human Evaluation Score and the Metrics Scores

		Simple-Complet.	Weighted Complet.	Nominal Info. Content	Textual Info. Content	Read-ability
Human Evaluation	Pearson	-.395	-.457	-.182	.842	.257
	Sign.	.085	.043	.443	.000	.274

The Textual Information Content Metric correlates in a high degree (0,842) with the average quality value given by the human reviewers. The significance of that correlation is very high (<0,01), that means that the correlation is real, and that it is not produced by chance. The correlation is even visible if we plot both scores (textual information content and the average score for quality) for the 20 examined objects (Figure 1). The line on top is the Average Quality as calculated from the scores given by the human reviewers. The bottom line is the value that the metric returned. While the lines do not follow the same exact pattern, much of the first line behavior could be explained by the second one.

Digging deeper, a multivariate regression analysis shows that roughly the 70% of the score given by humans could be explained by the Textual Information Content metric. Another 10% of the variation could be explained by the origin of the metadata (automatic or manual). Other tested metrics do not contribute to explain the grade. The results of the regression can be seen in Table 2.

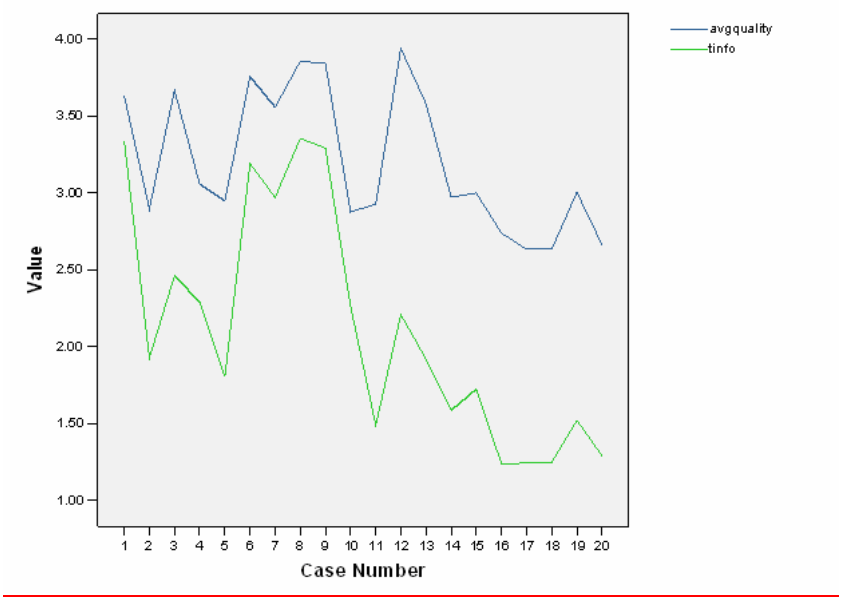

Fig. 1. Average Quality Score and the Textual Information Content Metric values

Table 2. Multivariate Regression Result

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
Textual lnformation	.842	.710	.694	.25406
Textual Information + Origin of Metadata	.905	.819	.798	.20623

We can conclude from the experiment that the Textual Information Content can be used as a good predictor of the human evaluation of the metadata record. This suggests that human reviewers mainly focus in the free text fields when they are examining the record. This result could have implications in the way that the metadata is presented to final users. For example: only textual fields should be shown by default and non-textual fields should only be presented by user's request. The other metrics, while could be useful to assess the fitness of the record for other applications, do not seem to be useful to predict the quality score of the record given by human reviewers.

3. Applications

Once we have identified good metrics to predict the value of some kind of quality of a metadata record we can use it in several applications:

- **Automatic Record Improvement**: We can compute the metric of all the records present in the repository and process those with a low score by an automatic extractor of metadata, such as [3]. Then we merge the newly generated record with the existing one and evaluate if the score improved. This could be used in legacy repositories, especially when the quality of the original metadata is low.
- **Quality Visualization**: The metrics values can be used to create visualizations of the repository in order to gain a better understanding of the distribution of the quality problems. For example, a treemap visualization [20] 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 on Figure 2. The treemap represent the structure of the ARIADNE repository. The global repository contains several Local repositories (BLKL, CS_L, UGAL, etc). The different authors publish metadata in the local repositories. The boxes represent the set of learning objects metadata records published by a given author. The color of the boxes represents the average of the Textual Information Content metric score of that set of records. The color scale goes from red (low quality) to yellow (medium quality), to green (high quality). This visualization helps us to easily spot authors that provide good textual descriptions to their objects.

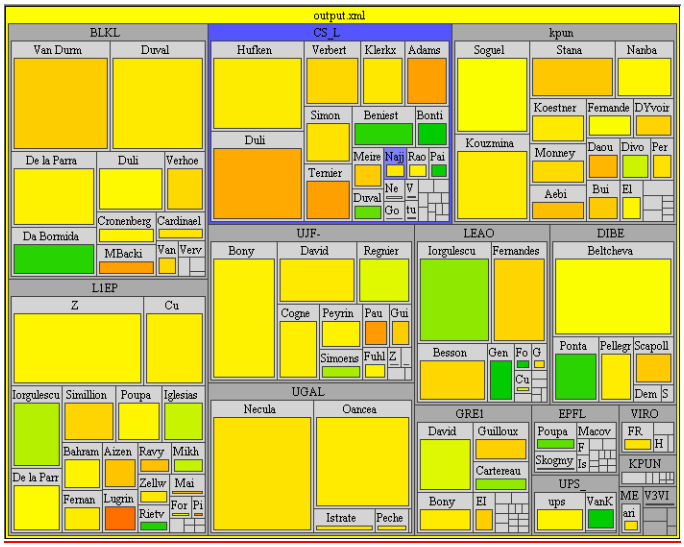

Fig. 2. Visualization of the ARIADNE Repository

- **Repository Interoperability**: One of the problems of the federation of different learning object repositories is the different quality standards for the metadata in the different communities of practice [21]. To avoid degrading the experience of a user in a given repository, the quality metrics could be integrated into a SQI Query

Language [4] to restrict the search of metadata records that have a similar quality level as the destination repository.

4. Conclusions

We have proposed 5 of quality metrics. Based in a small experiment, one of the metrics proposed, the Textual Information Content, seems to be useful to predict the quality score assigned by human reviewers to learning object metadata. While a bigger experiment with more records and different repositories is necessary in order to corroborate this result, it suggests that simple metrics could be use to gain insight in something as subjective and complex as the perception of metadata quality. If the correlation persists in further evaluations, it can be concluded that instead of showing the full metadata fields to the user, as several learning object applications do, it will be enough to present them the textual fields of title and description, much like in modern web searches provide only the title and excerpt of web pages.

While the quality metrics has been developed for learning object metadata, the calculations are "standard-agnostic". The same metrics could be easily extended to most kind of metadata records. For example, they could be applied to evaluate the quality of a digital library using, for example, Dublin Core or MARC metadata standards.

A lot more research and experimentation in quality metrics is needed, but it is clear that if we want to make the transition from small and isolated learning object repositories, to a fully integrated learning object ecosystem where millions of objects are created, indexed and changed daily, we need to have some kind of automatic evaluation of quality to avoid the system to break apart. This work is a first step that contributes to enabling such automated quality assessment.

5. Further Work

The quality value, assessed by the reviewers and predicted by the Textual Information Content metric, can be considered as "synthetic" quality. No real application present the learning object metadata record to the user in order to obtain a quality evaluation. The "real" quality of the metadata can be defined as the fitness of the metadata to fulfill a given purpose [9]. According to [1], the main purposes of the learning object metadata are to facilitate the search, evaluation, acquisition, and use of the learning object. Under these premises, there should be 4 different dimensions of quality to be measured: Retrieval Quality, Evaluation Quality, Accessibility Quality and Re-use Quality. As further work to be done in order to obtain the desired automatic evaluator of learning object metadata quality, we present some possible ways to calculate quality sources for the different quality dimensions. Those scores should be correlated with new quality metrics to find, among those metrics, useful predictors that can be used to develop the automatic evaluator of quality.

- **Retrieval Quality**: The logs of the searching tools for learning objects can be can analyzed in order to calculate the number of times that a given object appear in the result list when a relevant search has been performed. High quality metadata

should allow the object to appear in high percentage of relevant queries. The position of the object inside the result list could also be taken in account for the retrieval quality score.

- **Evaluation Quality**: The tools that present the metadata record to the user for evaluation can include a small survey asking if the information shown has been useful in order to decide if the object is relevant for him/her. The percentage of times that the metadata information has been found useful could be considered evaluation quality score.
- **Accessibility Quality**: Learning object tools could log the times that a learning object could not be accessed or retrieved because of errors in the metadata information. For example when there are errors in the identifier, URI, or cost of the learning object. The number of errors detected should decrease the accessibility quality score.
- **Re-use Quality**: The re-use of the learning object is not only determined by the characteristics of the metadata, but also for the characteristics of the learning object itself. Nonetheless, objects that are re-used more often, most probably, have a suitable metadata that allow users to integrate it in their learning systems. This, the re-use quality score could be determined by the number of times that the object has been repurposed in different learning contexts.

6. Related Work

While, to the authors' knowledge, there are no currently other initiatives to build an automatic quality assessment system for learning object metadata, in the Semantic Web area, [22] propose an automated system to assess the quality of the automatic extraction of semantic information from web pages. While the core idea, to have an automated way to ensure quality of an automated system, is similar to ours, it focus mainly in the correction of spelling errors and concept disambiguation through the use of ontologies and Semantic Web tools.

References

1. IEEE (2002). IEEE Standard for Learning Object Metadata. http://ltsc.ieee.org/doc/wg12/
2. Singh, A., Boley, H., & Bhavsar, V. C. (2004). A learning object metadata generator applied to computer science terminology. In Presented at the Learning Objects Summit, March 29-30, 2004.
3. Cardinels, K., Meire, M., & Duval, E. (2005). Automating metadata generation: the simple indexing interface. In Proceedings of the 14th WWW conference (pp. 548-556). New York, NY, USA: ACM Press.
4. Simon, B., Massart, D., van Assche, F., Ternier, S., Duval, E., Brantner, S. et al. (2006). A Simple Query Interface for Interoperable Learning Repositories. In B. Simon, D. Olmedilla, & N. Saito (Eds.), (pp. 11-18).

5. GLOBE (2006). Global Learning Objects Brokered Exchange. http://taste.merlot.org/initiatives/globe.htm
6. Verbert, K., Jovanovic, J., Gaševic, D., & Duval, E. (2005). Repurposing Learning Object Components. In. OTM 2005 Workshop on Ontologies, Semantics and E-Learning
7. Currier, S., Barton, J., O'Beirne, R., & Ryan, B. (2004). Quality assurance for digital learning object repositories: issues for the metadata creation process. ALT-J, Research in Learning Technology, 12, 5-20.
8. Barton, J., Currier, S., & Hey, J. (2003). Building Quality Assurance into Metadata Creation: an Analysis based on the Learning Objects and e-Prints Communities of Practice. In Proceedings 2003 Dublin Core Conference: (pp. 39-48). Seatle, Washington.
9. Guy, M., Powell, A., & Day, M. (2004). Improving the quality of metadata in Eprint archives. Ariadne.
10. Weibel, S. (2005). Border Crossings: Reflections on a Decade of Metadata Consensus Building. D-Lib Magazine, 11.
11. Ochoa, X. & Duval, E. (2006). Quality Metrics for Learning Object Metadata. In Proceedings ED-Media 2006. (pp. 1004 -1011)
12. Bruce, T. & Hillman, D. (2004). The continuum of metadata quality: defining, expressing, exploiting. In D.Hillman & L. Westbrooks (Eds.), Metadata in Practice (Chicago: American Library Association.
13. Duval, E. (2005). LearnRank: the real quality measure for learning materials. In A.McCluskey (Ed.), Policy and Innovation in Education - Quality Criteria (pp. 26-29). European Schoolnet.
14. Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In Proceedings of International Joint Conferences on Artificial Intelligence (pp. 448-453).
15. Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. Information Processing and Management, 39, 45-65.
16. Foltz, P., Kintsch, W., & Landauer, T. (1998). The measurement of textual coherence with Latent Semantic Analysis. Discourse Processes, 25, 285-307.
17. Ariadne Foundation (2005). Ariadne Foundation. http://www.ariadne-eu.org
18. Najjar, J., Ternier, S., & Duval, E. (2004). User Behavior in Learning Objects Repositories: An Empirical Analysis. In Proceedings of ED-MEDIA 2004
19. Shrout, P. & Fleiss, J. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. Psychological Bulletin, 2, 420-428.
20. Bederson, B., Shneiderman, B., & Wattenberg, M. (2002). Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies. ACM Transactions on Graphics (TOG), 21, 833-854.
21. Shreeves, S., Knutson, E., Stvilia, B., Palmer, C., Twidale, M., & Cole, T. (2005). Is "Quality" Metadata "Shareable" Metadata? The Implications of Local Metadata Practices for Federated Collections. In Proceedings of the Twelfth National Conference of the Association of College and Research Libraries Chicago, IL: Association of College and Research Libraries.
22. Lei, Y., Sabou, M., Lopez, V., Zhu, J., Uren, V., & Motta, E. (2006). An Infrastructure for Acquiring High Quality Semantic Metadata. In Proceedings of the 3rd European Semantic Web Conference (ESWC 2006)