Quality Prediction of Open Educational Resources
A Metadata-based Approach

Mohammadreza Tavakoli[1], Mirette Elias[2], Gábor Kismihók[1], Sören Auer[1]
[1]TIB – Leibniz Information Centre for Science and Technology, Germany {reza.tavakoli, gabor.kismihok, soeren.auer}@tib.eu
[2]University of Bonn and Fraunhofer IAIS, Germany {melias@uni-bonn.de}

***Abstract*—In the recent decade, online learning environments have accumulated millions of Open Educational Resources (*OERs*). However, for learners, finding relevant and high quality OERs is a complicated and time-consuming activity. Furthermore, metadata play a key role in offering high quality services such as recommendation and search. Metadata can also be used for automatic OER quality control as, in the light of the continuously increasing number of OERs, manual quality control is getting more and more difficult.**

In this work, we collected the metadata of 8,887 OERs to perform an exploratory data analysis to observe the effect of quality control on metadata quality. Subsequently, we propose an OER metadata scoring model, and build a metadata-based prediction model to anticipate the quality of OERs. Based on our data and model, we were able to detect high-quality OERs with the *F1 score* of 94.6%.

***Index Terms*—OER, open educational resources, metadata quality, OER quality, big data, data analysis, quality prediction**

I. Introduction

Nowadays, Open Educational Resources (OERs) play a significant role in informal learning. However, the low quality of search services and recommender systems limit the use of OERs [1]. Obviously, the lack of metadata that thoroughly describe OERs has a negative effect on the performance of these services [2], [3].

Furthermore, the vast amount of OERs, which are provided daily by content creators around the world, forces us to put more emphasis on automatic controlling of OERs quality. We depart from a strong assumption that OERs metadata and content quality have tight relationship with each other: the more OERs have high quality metadata, the higher the probability of high quality content is. Currently, manual methods are often used to evaluate both the quality of OER content and metadata [4], which solutions are time consuming and not scalable [3]. Therefore, we expect that a thorough automatic metadata analysis will significantly improve the quality control of OERs. There are some attempts already, which aim at automatizing the quality assessment of metadata (e.g., [3], [5], [6]). However, these focus only on the criteria definitions and metrics to evaluate existing OER metadata (e.g., [7], [8], [9]) without building an intelligent model, or models, to predict the quality of OERs based on metadata.

In this paper, we discuss the details of our exploratory data analysis on the metadata of 8,887 OERs from SkillsCommons[1], in order to provide insights about the quality of metadata in existing OERs, and the effect of quality control on metadata quality. Then, we build a metadata-based scoring and prediction models to anticipate the quality of OERs based on the results of our exploratory analysis.

[1]http://skillscommons.org

II. Related Work

OER metadata is important not only to aid learners in finding relevant content among large amount of OERs, but also to indicate OER quality [10]. In the literature, the quality of OER metadata has been determined in terms of the following dimensions: completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility [7]. Ochoa and Duval have converted these dimensions into a set of calculated metrics, which have been reused by most of the researchers addressing quality of OER metadata [8]. They partially evaluated their metrics (i.e., completeness, accuracy) on a list of 425 OERs from the ARIADNE Learning Object Repository [3].

Most studies about OER metadata quality have mainly focused on the completeness of metadata, by means of the availability of metadata elements, the presence of their values [11], and the evaluation of those values [12]. Pelaez and Alarcon have evaluated the completeness and consistency of OERs [13] by building their calculation on Ochoa and Duval's metrics [8]. They evaluated consistency of metadata elements values with respect to the standardized domain values (e.g. Language should be according to ISO 639-111 language standard). However, most of these approaches are not automatic, and either conceptual [14], [9] or focusing on one, or only a few dimensions [12], [15]. Therefore, there is a need for automatic and intelligent metadata quality assessment in order to improve the discoverability, usability, and reusability of OERs [16].

Based on the state-of-the-art, it is clear that: 1) it is worthwhile and timely to analyze OER metadata to improve OER-based services; and 2) there is a lack of intelligent prediction models, which evaluate the quality of OERs based on their metadata to facilitate the quality control. For these reasons the main research questions and objectives of our current work are:

- Conducting an exploratory data analysis on large amount of OERs' metadata.
- Building a scoring model with a data-driven approach that helps OER repositories and authors to evaluate and improve the quality of their OER metadata.

- Predicting the quality of OERs based on their metadata. This should guide automatic quality control processes and ultimately result in higher OER quality.

III. Data Collection and Research Method

In this section, we explain our steps towards our proposed model. First, we collected and maintained a large dataset of OER metadata. Second, we performed an exploratory data analysis and deduced results. Third, we built a scoring model accordingly, and finally, we proposed a prediction model to anticipate the quality of OERs.

A. Data Collection

We built an OER metadata dataset after retrieving all search results for the terms "Information Technology" and "Health Care" via the *SkillsCommons API*[2] resulting in a metedata pool of 8,887 OERs[3]. Each OER contains the following metadata: url, title, description, educational type, date of availability, date of issuing, subject list, target audience-level, time required to finish, accessibilities, language list, and quality control (a categorical value that shows if a particular OER went through manual quality control or not).

B. Exploratory Analysis of OER Metadata

As a point of departure, we used our dataset to explore the availability of metadata values, which are related to the category quality control ("with control" or "without control"). Our analysis showed a clear increase in OER metadata quality (in terms of *availability* of metadata) in the quality controlled OERs, which can be interpreted as a result of OER quality control. However, our analysis also indicated that the proportion of manual OER quality control in our dataset has been decreasing over the last years (from more than 60% in 2016 to less than 40% in 2019). We believe that the growing number of OERs is among the main reasons for this change. To conclude the results of our exploratory analysis:

1) Quality controlled OERs can be used to define benchmarks for quality of metadata fields
2) There is a need to define a method that facilitates the automatic assessment of OER metadata quality, and consequently the quality control of OERs.

C. OER Metadata Scoring Model

As the first step when building our scoring model, we defined the importance of each metadata field based on those OERs, which went through quality control.

For this purpose, we set the importance rate of each metadata field according to its availability rate among quality controlled OERs (between 0 and 1). For instance, all quality controlled OERs have a *title* and therefore, we set the importance rate of *title* to 1, and for *Time Required*, we set it to 0.58 since 58% of the controlled OERs have *Time Required* metadata. Moreover, we normalised the calculated importance rates as normalized importance rate.

Afterwards, for each field, we created a rating function in order to rate metadata values. We fit a normal distribution on values (lengths) of the following metadata fields: *title, description, and subjects*, as they have distributions similar to normal and used the reverse of *Z-score* concept (as $\frac{1}{\lceil |x-\bar{x}|/s \rceil}$ where $\bar{x}$ and s is the mean and standard deviation respectively of the field in the dataset) to rate the metadata values based on the properties of the quality controlled OERs. Thus, the closer an OER *title/description/subjects* length is to the mean of distributions, the higher is the rate. It should be mentioned that when a value is equal to the mean, the rate will be 1 and when it is empty the rate will be 0. Moreover, we used a boolean function for the four fields: *level, length, language, and accessibility* which assigns 1 when they have a value and assigns 0 otherwise. Table I illustrates the metadata fields, importance rates, normalized importance rates, and the rating functions.

TABLE I: OER metadata fields and importances

Type	Importance Rate [0-1]	Normalized Importance Rate [0-1]	Rating Function [0-1]
Title	1	0.17	$\frac{1}{\lceil \lvert x-5.5 \rvert / 2.5 \rceil}$
Description	1	0.17	$\frac{1}{\lceil \lvert x-54.5 \rvert / 40 \rceil}$
Subjects	0.86	0.145	$\frac{1}{\lceil \lvert x-4.5 \rvert / 3.5 \rceil}$
Level	0.98	0.165	If available: 1; else: 0
Language	0.92	0.155	If available: 1; else: 0
Time Required	0.58	0.098	If available: 1; else: 0
Accessibilities	0.59	0.099	If available: 1; else: 0

Finally, we defined the following two scoring models in order to cover the availability and adherence of the defined benchmarks:

Availability Model. We calculate the availability score of an OER o as Equation (1) where $norm\_import\_rate(k)$ is *Normalized Importance Rate* of metadata field k. This score shows how complete that metadata is in a weighted summation, in which the normalized important rates are the weights. Therefore, the more an OER contains important fields, the higher the availability score is. For instance, an OER with metadata about *title*, *description* and *level* (metadata fields with the highest importance rates), achieves a higher availability score than another one which has metadata for *subjects*, *time required*, and *accessibilities*.

$$avail\_score(o) = \sum_{k=available fields} norm\_import\_rate(k) \quad (1)$$

Normal Model. We calculate the normal score of an OER o as Equation (2), where $norm\_import\_rate(k)$ is the *Normalized Importance Rate* of metadata field k, and *rating(o, k)* is the assigned rating to OER o based on the rating function of k. This score shows how close metadata to the defined benchmark is (based on metadata of the OERs with quality control). With this scoring model, an OER which has the

[2]http://support.skillscommons.org/home/discover-reuse/skillscommons-apis/

[3]Our dataset can be downloaded from: https://github.com/rezatavakoli/ICALT2020_metadata

most similar metadata properties with the metadata of quality controlled OERs, achieves the highest normal score.

$$norm\_score(o) = \sum_{k=fields} norm\_import\_rate(k) * rating(o,k) \quad (2)$$

D. Predicting the quality of OERs based on their metadata

We used 80% of our data as a training set and trained a machine learning model to predict the quality of OERs based on their metadata and our scoring model. Therefore, we got the OERs "with control" as higher quality class (containing 4,651 OERs), and set the remaining as lower quality class (containing 4,236 OERs). As a classifier, a Random Forest model was trained to make a binary decision (i.e., high-quality or low-quality) based on the fields: *Importance score, Availability Score, Level Metadata Availability, Description Length, Title Length, and Subjects Length*.

IV. Validation

We built a test set using the remaining 20% of data. The classifier achieved an accuracy of **94.6%**, where 95% of F1-score for "with control" class, and 94% of F1-score for "without control" class[4]. Moreover, we extracted the importance value of each feature for the classification task. Table II represents the features of our model and their importance score [0-1]. The importance values reveal the effect of each feature in our prediction model. The model assigns the highest value to the *Availability Score* and *Normal Score* features, which are the indicators we proposed. Thus, we can infer that these two indicators can illustrate the quality of OER metadata.

TABLE II: OER quality prediction model features

Feature	Importance score [0-1]
Availability Score	0.32
Normal Score	0.25
Level Metadata Availability	0.23
Description Length	0.10
Title Length	0.05
Subjects Length	0.05

V. Conclusion and Future Work

In this study, we collected and analysed the metadata of a large OER dataset to provide deeper insights into OER metadata quality, and proposed a scoring and a prediction model to evaluate the quality of OER metadata and, as a consequence, OER content quality. The model proposed in this short paper not only helps OER providers (e.g. repositories and authors) to revisit and think about the importance of their metadata quality, but also facilitate the quality control of OERs in general. These are essential in the light of the rapidly growing number of OERs and OER providers these days. Applying our model on our Skillscommons dataset indicated that it can detect OERs with quality control with the accuracy of **94.6%**.

[4]The implementation steps and results in Python can be downloaded from: https://github.com/rezatavakoli/ICALT2020_metadata

We consider this study as one of the first important steps to propose intelligent models to improve OER metadata quality and OER content. As future work, we plan to further improve and validate our models by collecting more data from other OER repositories and consider more metadata features (e.g. text-based analysis of title and description). Additionally, we plan to validate our approach in other contexts, for instance by applying our scoring and prediction model to open educational videos on Youtube.

References

[1] J. Chicaiza, N. Piedra, J. Lopez-Vargas, and E. Tovar-Caro, "Recommendation of open educational resources. an approach based on linked open data," in *Global Engineering Education Conference*. IEEE, 2017, pp. 1316–1321.

[2] P. Király and M. Büchler, "Measuring completeness as metadata quality metric in europeana," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 2711–2720.

[3] X. Ochoa and E. Duval, "Automatic evaluation of metadata quality in digital repositories," *International journal on digital libraries*, vol. 10, no. 2-3, pp. 67–91, 2009.

[4] A. Tani, L. Candela, and D. Castelli, "Dealing with metadata quality: The legacy of digital library efforts," *Information Processing and Management*, vol. 49, no. 6, pp. 1194–1205, 2013.

[5] T. Trippel, D. Broeder, M. Durco, and O. Ohren, "Towards automatic quality assessment of component metadata," *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pp. 3851–3856, 2014.

[6] D. M. Nichols, C.-H. Chan, D. Bainbridge, D. McKay, and M. B. Twidale, "A lightweight metadata quality tool," in *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, 2008, pp. 385–388.

[7] T. R. Bruce and D. I. Hillmann, "The continuum of metadata quality: defining, expressing, exploiting," in *Metadata in Practice*. ALA editions, 2004.

[8] X. Ochoa and E. Duval, "Quality Metrics for Learning Object Metadata," *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, no. 2004, 2006.

[9] A. Romero-Pelaez, V. Segarra-Faggioni, N. Piedra, and E. Tovar, "A proposal of quality assessment of oer based on emergent technology," in *2019 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 2019, pp. 1114–1119.

[10] S. Ushakova, "Usability of metadata standards for open educational resources," Oct 2015. [Online]. Available: https://hclemuseum.wordpress.com/2015/10/02/usability-of-metadata...

[11] M. A. Sicilia, E. Garcia, C. Pagés, J.-J. Martínez, and J. M. Gutierrez, "Complete metadata records in learning object repositories: some evidence and requirements," *International Journal of Learning Technology*, vol. 1, no. 4, pp. 411–424, 2005.

[12] M. Margaritopoulos, T. Margaritopoulos, I. Mavridis, and A. Manitsaris, "Quantifying and measuring metadata completeness," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 4, pp. 724–737, 2012.

[13] A. R. Pelaez and P. P. Alarcon, "Metadata quality assessment metrics into ocw repositories," in *Proceedings of the 2017 9th International Conference on Education Technology and Computers*. ACM, 2017, pp. 253–257.

[14] T. Margaritopoulos, M. Margaritopoulos, I. Mavridis, and A. Manitsaris, "A conceptual framework for metadata quality assessment." in *Dublin Core Conference*, 2008, pp. 104–113.

[15] A. Romero-Pelaez, V. Segarra-Faggioni, and P. P. Alarcon, "Exploring the provenance and accuracy as metadata quality metrics in assessment resources of ocw repositories," in *Proceedings of the 10th International Conference on Education Technology and Computers*. ACM, 2018, pp. 292–296.

[16] D. Gavrilis, D.-N. Makri, L. Papachristopoulos, S. Angelis, K. Kravvaritis, C. Papatheodorou, and P. Constantopoulos, "Measuring quality in metadata repositories," in *International Conference on Theory and Practice of Digital Libraries*. Springer, 2015, pp. 56–67.