DCPAPERS

Proc. Int'l Conf. on Dublin Core and Metadata Applications 2013

Reusing Textual Resources in Educational Assessment: Adding Text Readability Metrics to Learning Metadata

Muriel Foulonneau
Eric Ras
Public Research Centre Henri Tudor,
Luxembourg
{muriel.foulonneau, eric.ras}@tudor.lu

Elie Abou Zeid
Talar Atéchian
Antonine University,
Lebanon
elie_abouzeid@hotmail.com,
talar.atechian@upa.edu.lb

Abstract

Many digital libraries have identified learners as a core audience. Indeed, many of their resources can be reused in educational contexts. Nevertheless, the search criteria used for retrieving texts as a specific multimedia type are limited. They often do not include properties specific to educational contexts. Assigning LOM metadata to a theatre play or a painting is difficult, since it was not created for a particular learning context. However, it is possible to assign metadata to textual resources based on their characteristics and map these characteristics to an IEEE LOM or DCMI *Audience* metadata element. Text readability metrics for instance can be mapped to educational audiences. In the scope of the iCase project, we are developing an assessment item generation system. We have therefore analyzed metadata models for assessment resources and defined a set of metadata which should be assigned to the multimedia components of assessment items. A major challenge consists in relating multimedia resources to the specific audience metadata. In order to include external resources such as texts, we developed a component available as a Web service to assign metrics related to text readability. In this paper, we present metadata for assessment items and introduce readability metrics.

Keywords: education, assessment, text readability metrics, Web mining.

1. Introduction: Metadata for reusing multimedia resources in education

Many data curators from cultural digital libraries ambition to improve the use of digital resources in educational contexts. This ambition sometimes takes the form of an addition of LOM metadata. While this certainly applies to a lecture, it is unclear that a level of difficulty or an audience can be assigned to the Mona Lisa or a piece of literature. A multimedia resource is indeed often not a learning object on its own but it rather requires taking into consideration the learning context in which it can be useful. Metadata models for learning resources, such as IEEE Learning Object Metadata (LOM) or the Dublin Core Education application profile are usually applied to resources specifically created for learning purposes such as curricula. Other metadata models are applied to any multimedia resources that do not refer to their potential usage in learning contexts, such as MODS[1]. Many resources are stored and curated in digital repositories. They are available through a digital library interface with no specific feature for facilitating their reuse in educational contexts.

Assessment items (test questions) are particular educational resources, where the choice of words, as well as multimedia resources are very important since they can directly impact the outcomes of a test. Items are composed of a stem (i.e. question), potentially answer options, and auxiliary information such as external multimedia resources (e.g., texts, pictures) (Gierl et al., 2008). Methods have been developed to assign psychometric properties (e.g., item difficulty) and verify the quality of assessment items, i.e. their reliability to measure the construct (e.g.,

[1] http://www.loc.gov/standards/mods/

https://doi.org/10.23106/dcmi.952136290

knowledge or skill). These indicators are recorded in IMS-QTI (Question and Test Interoperability) metadata[2]. However, no specific provision is made regarding education or assessment specific metadata which should be attached to auxiliary information.

In the scope of the iCase project, we aim to develop tools for e-assessment. We have supported the delivery of tests for the Programme for International Student Assessment (PISA) and Programme for the International Assessment of Adult Competencies (PIAAC) studies of the OECD[3] on comparative educational levels, as well as school monitoring, adaptive tests in language learning, and finally tests for detecting the formative efficiency of documents to raise awareness of children on health issues. We focus in particular on the generation of assessment items from semantic resources and the inclusion of external resources, typically from the Web or multimedia repositories (e.g., Currier, 2007).

In this paper, we present our work to create an open-source component which assigns various readability metrics to texts from digital repositories. By adding readability metrics we aim to increase the reusability of text resources in the generation process of assessment items. The long-term objective is to improve the item authoring interface of our e-assessment platform, and second to allow automatic item generation (AIG) approaches reuse extensively multimedia resources from digital repositories and the Web within assessment items.

2. Metadata on multimedia resources integrated in assessment items

The IMS-QTI metadata model aims to describe assessment items. It is an application profile of the IEEE LOM metadata model. Among the core differences with LOM is a section dedicated to *usage data*. Many metrics can indeed be used to assess item quality, based in particular on calibration. Metrics used in Item Response Theory (Reise et al., 2005) for instance can be recorded as usage data. Because of the large number of metrics which can be assigned to items, the usage data section can include any item statistics, composed of one or multiple values. A statistic can have a name, a value, a standard error and a standard deviation. Glossaries are vocabularies defined using the VDEX model (Vocabulary Definition Exchange)[4] in order to represent common statistics in a harmonized manner[5]. Item statistics are part of Usage data and complement the descriptive Metadata section (IEEE LOM profile) of the IMS-QTI standard (FIG 1).

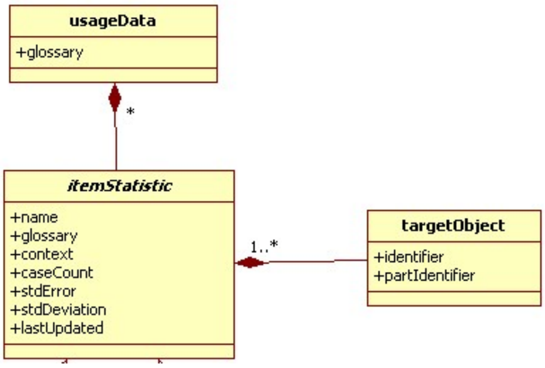

FIG 1 - Modeling of usage data in IMS-QTI 2.1

Metadata models for assessment items (Sarre et al., 2010) do not allow defining metadata for item components such as multimedia resources. Following the framework suggested by Currier (2007), multimedia assets and assessment items should live in distinct repositories. No specific relation should be made between IMS-QTI metadata and assets (in our case textual resources).

[2] http://www.imsglobal.org/question/

[3] http://www.oecd.org/pisa/, http://www.oecd.org/site/piaac/

[4] http://www.imsglobal.org/vdex/index.html

[5] http://www.imsglobal.org/question/qtiv2p1/imsqti_mdudv2p1.html#section10042

https://doi.org/10.23106/dcmi.952136290

Nevertheless, certain features of multimedia resources make them suitable for particular learning contexts.In Foulonneau et al. (2011), we investigated the criteria that need to be taken into consideration when selecting resources to include in assessment items. We highlight the importance of information that is not directly part of LOM, because it is related to multimedia resources in learning objects rather than learning objects themselves. Indeed, the choice of multimedia resources is mainly guided by the intellectual property rights, the type of resource, and the potential bias it can entail.

Bias is a critical issue in the context of e-assessment, since it includes a parameter which is unrelated to the construct (i.e., the knowledge or skill that needs to be assessed) and can impact the outcome of the item. Typically, the presence of faces on a picture can lead to a cultural bias because facial expressions can be understood differently according to cultures (Gruba, 1997). Different levels of vocabulary can lead to a socio-cultural bias or a bias towards minorities (Drasgow et al., 2006). Different formats can entail a bias when a particular population is more familiar with pictures for instance (Van de Vijver, 2004). Regarding texts, difficulty or readability metrics have been proposed in order to define the audience information and therefore the level and age of students for whom a particular text would present no reading difficulty. In order to increase the reusability of textual resources in education, Heilman, Zhao et al. (2008) have therefore proposed annotating texts with readability metrics.

Metadata assigned to multimedia resources can include a qualitative evaluation (e.g., *Difficulty* with a value space from *very easy* to *very difficult*) or an interpretation of statistics (e.g., an *Audience* metadata element with a value K-12). However, numerous readability metrics can be implemented. Like IMS-QTI usage data, adding them as an aggregate in a single metadata value excludes to convey the method used for the creation of the metadata element and any different use which could be made of the metrics. In the context of the iCase project, we investigate the use of readability metrics for creating specific metadata that facilitate the reuse of textual resources in assessment items as suggested by Heilman, Zhao et al. (2008). In the next sections, we present readability metrics that can be used either as a derived metadata property or as a set of distinct metrics.

3. Text readability metrics

Text readability skills are identified as a major component of education in the US. The *Common Core State Standards for Reading* [6] aim to harmonize the level of difficulty of the texts given to learners. They can be used by teachers and parents to guide their choice of reading assignments.

Reading difficulty metrics typically take into consideration the vocabulary used in the texts, whether known or unknown by any particular user, as well as text structure and style. They result in both quantitative and qualitative indicators.

Since the 1920's, more than 200 formulae have been proposed (Dubay, 2004). Statistical approaches use the length of sentences, the length of words and their scarcity (Fry, 2002). Fry Readability, New Date-Chall, Gunning-Fog and Flesh Kincaid Reading Ease are among the most famous examples of those metrics. The Flesh-Kincaid Reading Ease and Grade Level metrics which use the average number of syllables per word and the average number of words per sentence are still the most commonly used (Heilman, Collins-Thomson et al., 2008). Metrics such as Lexile Scale and Mean Log Word Frequency use corpora. The frequency of words is typically calculated from a corpus. Despite their efficiency, the bias of these metrics against technical texts which contain rare words for instance has been criticized (Sheehan et al., 2012).

Other types of metrics have therefore been developed which analyze the structure and linguistic features of the text. Coherence metrics represent the level of organization of concepts and relations in a text. It is typically calculated from the semantic similarity between neighbor

[6]. http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf

https://doi.org/10.23106/dcmi.952136290

sections of a text in order to detect any change of topic. Latent Semantic Analysis has been used in the Intelligent Essay Assessor (Folz et al., 1999) in order to obtain a coherence metric. The importance of coherence in readers' understanding of a text is however debated (Foltz et al., 1998; Dubay, 2004).

McNamara et al. (2010) have demonstrated the importance of cohesion metrics for measuring text readability. Text cohesion represents the relations between the various components of a text. This can typically use the semantic relations between terms, without referring necessarily to their position in the text.

A number of systems have been developed in order to support both teachers and parents in the selection of texts. Existing systems range from manually categorized Web pages (NetTrekker[7]) to a set of syntactic and semantic metrics (Huff, 2008). Coh-Metrix[8] developed by the University of Memphis (McNamara et al., 2010) assesses text coherence and cohesion and provides a mapping between text characteristics and the expected level of users. Read-X (Miltsakaki et al., 2007) classifies texts from the Web along a readability scale based on a corpus analysis. Toreador predicts the difficulty of the vocabulary contained in a text according to its frequency in a domain or according to a particular educational level. REAP[9] from Carnegie-Mellon recommends texts according to the vocabulary known by a user (Brown et al., 2004). Most of these systems are adapted to the analysis of Web pages in order to select Websites for education. Classic statistical analyses of texts still work well on relatively homogeneous corpora, whereas other metrics can improve their performance (Nelson et al. 2012). Nevertheless, their efficiency depends on the educational context in which they are used (e.g., language learning or biology) and the type of text (e.g., news vs. literature). SourceRater[10] developed by ETS (Educational Testing Service, developer for instance of the TOEFL for English language certification) also takes into consideration the genre of the text Sheehan et al. (2010).

The most recent evolutions therefore combine readability metrics and tend to refine them according to the educational context and the genre of texts and investigate the use of personalized metrics (Fry, 2002; Sheehan et al., 2010).

4. An annotation component to generate metadata on text readability

In order to author e-assessment items, external multimedia resources, such as a text may be added. In this case, the use of a digital library can help identify relevant resources. However, it is necessary to analyse textual resources so as to add specific metadata related to the readability. Whereas *Flesh Kincaid Grade Level* for instance propose a direct mapping between text statistics and a US grade level to fill for instance *dcmi:Audience* metadata, it fails to take into consideration the specific features of different types of texts and the importance of the user profile (e.g., on the extent of the known vocabulary). We are therefore developing a component which implements the following metrics: a) *Flesh Kincaid Reading Ease* (readability metric) and b) various indicators which can be combined in order to define a tailored readability metric, i.e., the similarity of words used in neighbour sentences (coherence), references (third person and demonstrative pronouns), conjunctions, and identifications (grammatical cohesion), exact repetitions (lexical cohesion), words frequency, and known words (customizable readability) (Abou Zeid et al., 2012).

As opposed to classical approaches, we use semantic technologies and Web mining. The similarity of words was determined based on their semantic network in the WordNet lexical dataset[11]. Word frequency was calculated through a Web mining approach targeted to specific

[7] http://www.nettrekker.com
[8] http://cohmetrix.memphis.edu
[9] http://reap.cs.cmu.edu/
[10] http://www.ets.org/research/topics/as_nlp/educational_applications
[11] http://wordnet.princeton.edu/

https://doi.org/10.23106/dcmi.952136290

Websites. Known words are encoded in an extended version of the *Personal Information Model Ontology* (Sauermann et al., 2007).

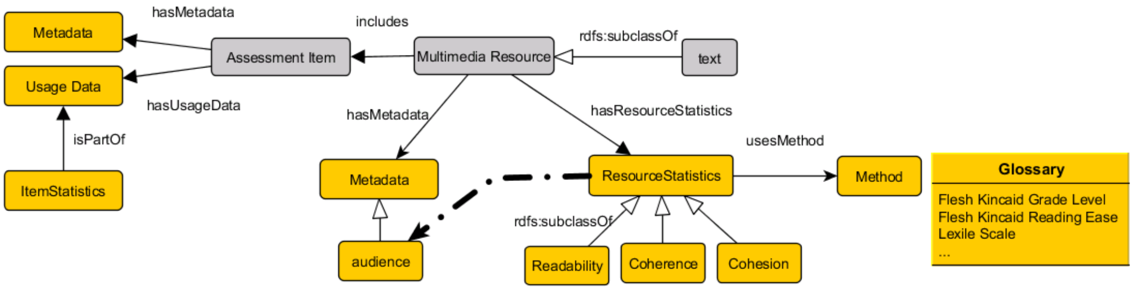

FIG 2 - Resource statistics bound to textual resources

The objective of the component is to record statistics about the resource so as to infer metadata such as *dcmi:Audience* (bold arrow), which for convenience was represented as a class rather than a property on Figure 2.

The metrics can be obtained by loading a text in PDF, Word or HTML format. A Web interface is available as well as a SOAP Web service interface. The various indicators can then be combined to create readability metrics, either on the fly, based on a particular user profile or as a background metadata creation process for a set of documents.

5. Conclusion and future work

In the context of our work on e-assessment in various domains, we have to take into consideration the many aspects involved in the selection of multimedia (often external) resources in assessment items. This aims to avoid any bias and to reliably assess a particular skill without the candidate to be penalized by an unrelated difficulty such as an unknown vocabulary term. Our system defines a set of classical metrics using statistical and uses innovative semantic-based and Web mining technologies to determine the audience of a textual resource.

Although this can be applied to all types of educational resources, our project aims to complement the test item authoring environment of the open source platform used for item creation and delivery. Given the cost of authoring assessment items, a number of projects have been designed in order to generate assessment items (Gierl et al., 2013). However, generated assessment items currently do not include multimedia resources. In In order to select automatically multimedia resources, a number of key information, related to the IPR, the format and the content should have been available Foulonneau (2011). Our work aims to enable the analysis of multimedia resources in order to include it in assessment items, either through the item authoring interface or through the generation of items.

However, the assignment of a single readability metric in an *Audience* metadata (IEEE LOM or DCMI) is insufficient. Indeed, we realized that a single readability metric is not sufficient, since the readability of texts is more and more assessed according to user/learner profiles. Therefore, metadata related to the genre of the text as well as linguistic metrics, such as cohesion, coherence, and statistical metrics of readability can help selecting resources. A personalized approach to readability metadata may then be implemented.

This raises an issue regarding the way in which metadata values are created. Indeed, in the case of IMS-QTI, it is possible to add general information on the quality of the assessment item. However, the detail of the quality metrics can optionally be provided in a dedicated *usage* data section with the specification of vocabularies to describe commonly accepted metrics. Similar issues have been raised to aggregate paradata on learning resources in the scope of the learning registry (Jesukiewicz et al., 2011). Readability metrics can be automatically generated. They do not therefore represent a significant additional metadata creation costs. They are an example of metadata created for increasing the reusability of any type of resource in educational contexts.

https://doi.org/10.23106/dcmi.952136290

Our future work will be dedicated to the improvement of metrics, their adaptation to other languages, the creation of metadata for texts, and their integration into a metadata structure, similar to the usage data section of the IMS-QTI metadata model.

References

Abou Zeid, Elie, Muriel Foulonneau, M., and Talar Atéchian. (2012). Réutiliser des textes dans un contexte éducatif. Document numérique, 15(3), 119-142.

Brown, Jonathan and Maxine Eskenazi. Retrieval of authentic documents for reader-specific lexical practice. Proceedings of InSTIL/ICALL Symposium 2004. Venice, Italy, 2004.

Currier, Sarah. (2007). Assessment item banks and repositories. JISC-CETIS paper. http://wiki.cetis.ac.uk/Assessment_Item_Banks_and_Repositories

Drasgow, Fritz and Krista Mattern. (2006). New Tests and New Items: Opportunities and Issues. In Dave Bartram, Ronald K. Hambleton (eds.) Computer-Based Testing and the Internet Issues and Advances. John Wiley & Sons Ltd., Chichester, UK.

DuBay, William H.. (2004). The Principles of Readability. Impact Information, Costa Mesa, California.

Foltz, Peter W., Walter Kintsch, and Thomas K. Landauer. (1998). The Measurement of Textual Coherence with Latent Semantic Analysis. Discourse Processes*,* 25(2-3), 285–307.

Foltz, Peter W., Darrel Laham, and Thomas K. Landauer. (1999). The Intelligent Essay Assessor: Applications to Educational Technology. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1(2).

Foulonneau, Muriel. (2012). Generating educational assessment items from linked open data: the case of DBpedia. In The Semantic Web: ESWC 2011 Workshops, 16-27. Springer Berlin Heidelberg.

Foulonneau, Foulonneau, Eric Ras, and Thibaud Latour. Reusing multimedia resources in assessment items – practices and impact. Poster abstract In Whitelock, D., Warburton, W., Wills, G., and Gilbert, L. (Eds.), CAA 2011 International Conference, University of Southampton.

Fry, Edward. (2002). Readability versus Leveling. The Reading Teacher*,* 56(3), nov. 2002, International Reading Association, 286-291.

Gierl, Mark J., Jiawen Zhou, and Cecilia Alves. (2008). Developing a Taxonomy of Item Model Types to Promote Assessment Engineering. Journal of Technology, Learning, and Assessment, 7(2).

Gierl, Mark J., and Thomas M. Haladyna. (2013). Automatic Item Generation: Theory and Practice. Routledge, New York.

Gruba, Paul. (1997). The role of video media in listening assessment. System, 25(3), 335-345.

Heilman, Michael, Le Zhao, Juan Pino, and Maxine Eskenazi. (2008). Retrieval of Reading Materials for Vocabulary and Reading Practice. Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications, 80–88, Columbus, Ohio, USA, June 2008. Association for Computational Linguistics.

Heilman, Michael, Kevyn Collins-Thompson, and Maxine Eskenazi. (2008). An Analysis of Statistical Models and Features for Reading Difficulty Prediction. Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications, 71–79, Columbus, Ohio, USA, June 2008. Association for Computational Linguistics.

Huff, Leslie. (2008). Review of *netTrekker d.i.* Language Learning & Technology, 12(2), June 2008, 17-25.

Jesukiewicz, P., & Rehak, D. R. (2011). The Learning Registry: Sharing Federal Learning Resources. In *The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC)* (Vol. 2011, No. 1). National Training Systems Association.

McNamara, Danielle S., Max M. Louwerse, Philip M. McCarthy, and Arthur C. Graesser. (2010). Coh-Metrix: Capturing linguistic features of cohesion. Discourse Processes, 47(4), 2010, 292–330.

Miltsakaki, Eleni, Audrey Troutt. (2007). Read-X: Automatic Evaluation of Reading Difficulty of Web Text. World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2007, 1, 7280-7286.

Nelson, Jessica, Charles Perfetti, David Liben, and Meredith Liben (2011). Measures of Text Difficulty: Testing their Predictive Value for Grade Levels and Student Performance. Technical report, Student Achievement Partners.

Reise, Steven P., Andrew T. Ainsworth, and Mark G. Haviland. (2005). Item Response Theory Fundamentals, Applications, and Promise in Psychological Research. In Current Directions in Psychological Science, April 2005 14(2), 95-101

Sarre, Sandrine and Muriel Foulonneau. (2010). Reusability in e-assessment: Towards a multifaceted approach for managing metadata of e-assessment resources. In Internet and Web Applications and Services (ICIW), 2010 Fifth International Conference on (pp. 420-425). IEEE.

https://doi.org/10.23106/dcmi.952136290

Sauermann, Leo, Ludger van Elst, and Andreas Dengel. PIMO – a framework for representing personal information models. I-SEMANTICS '07, Graz, Austria, 2007, 270-277.

Sheehan, Kathleen M., Irene Kostin, Yoko Futagi, and Michael Flor. (2010). Generating automated text complexity classifications that are aligned with targeted text complexity standards. Research report ETS RR-10-28, 2010, Educational Testing Service.

Van de Vijver, Fons, Norbert K. Tanzer. (2004). Bias and equivalence in cross-cultural assessment: an overview. Revue européenne de psychologie appliquée. 54,(2), 119–135.

https://doi.org/10.23106/dcmi.952136290