Automatic Generation of SCORM Compliant Metadata for Portable Document Format Files

Antonio Maratea, Alfredo Petrosino, Mario Manzo

Abstract: *The Shareable Content Object Reference Model (SCORM) is a widely adopted collection of specifications for web-based e-learning to which most Learning Management Systems adhere. While it allows reusability of content, it requires extensive, slow and expensive metadata annotation, and this fact prevents many content producers from properly creating and using Learning Objects. We propose an automatic metadata generation procedure that allows to label specific Learning Objects (scientific papers) with general metadata compliant to the SCORM. As some metadata are intrinsically unrelated to structure while others are strictly connected to structure, two different techniques were developed: one based on vocabularies and the other based on structural features. Results show that, in the provided context and for the "general" metadata category, the accuracy of annotations is comparable to that of a human expert.*

Key words: *Automatic Metadata Generation, SCORM, LOM, Portable Document Format, Vector Space Model.*

INTRODUCTION

While the production of educational content is every day easier and technological supported learning is already a widespread reality, standards are needed to overcome the risk of a new Babel in distance learning and to lower the production cost and time for courses. The three pillars on which the most recent and popular standards are built are: *reusability of content*, *interoperability of platforms* and *didactic neutrality*. For all of them, a detailed and diffuse metadata annotation of learning content is essential, especially concerning reusability.

The Shareable Content Object Reference Model [4] (SCORM) is a collection of specifications for web-based e-learning widely adopted worldwide. The SCORM standards are governed and published by the ADL[1] following an initiative of the United States Department of Defense. It is a reference model, not a standard by itself, that means it is a collection of standards from various specialized organizations (i.e. IMS[2] and AICC[3]). Primary goal of SCORM is to provide a reference model that allows interoperability, easy access and reuse of web-based learning units for Industry, Government and University. The main advantage of SCORM is the reusability of Learning Objects (LO), at the cost of a thorough annotation that often can be guaranteed only by a scarce, slow and expensive human expert.

Automatic Metadata Generation (AMG) [3] began with the introduction of digital documents since the years '50 and concerns indexing, abstracting, and classifying automatically documents [5]. Data sources are usually represented by documents in various formats and their context information, while metadata to be filled may concern any aspect of document usage, content or context. The process used by AMG algorithms for the metadata retrieval is known as Metadata Extraction (ME). In this paper we propose an AMG procedure that allows to label specific Learning Objects (scientific papers) with metadata compliant to the SCORM. The purpose is to ease and to speed up the annotation process, while supporting the extensive annotation and use of scientific as didactic content. The paper is organized as follows: first related work is briefly discussed;

CompSysTech'12, June 22-23, 2012, Ruse, Bulgaria.

then the used algorithms are presented in detail, followed by results of their application to a limited set of papers; finally conclusions are drawn and future work outlined.

RELATED WORK

The goal of good AMG algorithms is to find a balance between quality and complexity of obtained metadata. In [4] it is reported a review of existing and potential metadata automation approaches. In fact, many organisations are also looking at automated metadata systems to reap automatic metadata generation benefits. In [16] the use of Hidden Markov Models is explored, learning a model structure from labeled and unlabeled data. Subsequently it is shown how a manually built model containing multiple states related to extraction fields outperforms a model with one state per field, and, strategies for learning the model structure automatically from data are discussed. In [9] the capabilities of two Dublin Core[4] automatic metadata generation applications are tested. Samples of 29 resources, from National Institute of Environmental Health Sciences (NIEHS), were submitted to the systems. The experimental phase shows that extraction processing algorithms can contribute to useful automatic metadata generation and that harvesting metadata from META tags created by humans can have a positive impact on automatic metadata generation. In [11] a framework for automatic metadata extraction starting from electronic documents, both text and images, is described. It can be useful to ease metadata creation process in case of papers. The system is composed of three blocks: a text conversion module for converting electronic document into standard text file format, a task-oriented parser module for automatically extracting metadata from text using a pre-defined grammar, and data verification module for identifying and correcting the errors in extracted metadata. In [7] a system based on a novel spatial/visual knowledge principle for extracting metadata from scientific papers stored as PostScript files is described. The system embeds general knowledge about the graphical layout of a scientific paper to guide the metadata extraction process and can effectively assist the automatic index creation for digital libraries. In [13] Natural Language Processing and Machine Learning techniques are used to automatically generate metadata for learning resources. Evaluation phase is performed on metadata assigned to learning resources by either automatic tagging or manual assignment. Results show minimal difference between automatically generated metadata and manually assigned metadata. In [10] a technique for automatically generating Qualified Dublin Core metadata is proposed. The descriptions cover ten of the fifteen standard Dublin Core metadata elements and semantic precision is increased by element refinement and encoding scheme qualifiers. The produced descriptions could most obviously be used by tools for resource discovery but also by local data management applications. Finally, in [6] a tool of Automatic Metadata Generation is proposed to assist in the metadata creation process, enabling LOs to share common contextual metadata while receiving additional high quality LO specific metadata without the need for manual metadata creation efforts.

---

[1] Advanced Distributed Learning Initiative, http://www.adlnet.org
[2] IMS Global Learning Consortium, http://www.imsglobal.org/
[3] Aviation Industry CBT Committee, http://www.aicc.org/
[4] Dublin Core Metadata Initiative, http://dublincore.org/

MATERIALS AND METHOD

The IEEE LOM (Learning Object Metadata) standard groups all its meta-data in nine top level categories: *general, life cycle, meta-metadata, technical, educational, rights, relation, annotation* and *classification*. Each category may have subcategories of different depths, depending on the complexity of the learning object being annotated. Accurate filling of all these fields is time consuming and may be difficult even for an expert, due to the high level of detail and skill required.

In this paper we considered the automatic generation of metadata only for the *general* category, namely: *identifier, title, language, description, keyword, coverage, structure* and *aggregation level*. As the learning objects considered are single scientific articles, the *aggregation level* and *structure* attributes are not relevant, while *identifier* may always be generated artificially with a sequence. For the remaining attributes, used techniques are based on natural language processing and vocabularies. The preliminary steps are to transform the pdf file in plain text and to choose the vocabularies. For each field to be filled a slightly different strategy was implemented, from two categories: with structural information and without structural information.

Medatata Extraction with Structural Information (MESI)

Structure should always reflect relevance: words assume a different meaning depending on the place or the format they are given. Completely ignoring the format of a document implies loosing some precious information, and flattening its content makes relevance recognition of some words (section headers, keywords, footnotes) very hard.

To effectively recognize relevance of words in various sections of a paper, given a pdf file, first it has to be transformed in a way that accounts for its structure. To this purpose, a Layout-Aware PDF Text extraction tool (LA-PDFText) [15] was used as preprocessing step. It produces an output XML file divided into blocks corresponding to sections of the source file. Each block wraps the plain text contained in the corresponding section and adds to it some structural features, as character height, font, number of lines etc. Pages and blocks are organized hierarchically and numbered.

After preprocessing, a rule-based parsing strategy was used on the file produced by LA-PDFText, and all block were analysed to identify relevant information.

Specifically, the metadata for which this approach was followed are:

- **Title:** by definition the "name" given to the learning object. Title was found selecting the block of text with maximum font size.
- **Description:** by definition a textual description of the content of the Learning Object. The block identified as the "abstract" with an occurrence search of the word "abstract" was used.

Medatata Extraction without Structural Information (MENSI)

Large part of the information available on the web is unstructured, mostly in the form of free text. This fact required tools for natural language processing and algorithms able to cope with free text, especially in a Information Retrieval (IR) perspective. In this paper natural language processing was done through the classical Vector Space Model, that models terms by their frequency in a document and allows the use of classical quantitative analysis algorithms.

To effectively recognize metadata attributes not related to the paper organization, given a pdf file, first it has to be transformed in plain text and all its formatting structure removed. To this purpose, any pdf to txt conversion tool may be used. Then a vocabulary must be chosen - and which one depends on the application: it may be a vocabulary of scientific terms for a given subject or a true natural language dictionary. The document terms are then stemmed and a score matrix of similarity of each term of the document to any term of the vocabulary is filled.

Terms / Docs	T_1	T_2	$\cdots$	T_n
D_1	$s_{1,1}$	$s_{1,2}$	$\cdots$	$s_{1,n}$
D_2	$s_{2,1}$	$s_{2,2}$	$\cdots$	$s_{2,n}$
$\vdots$	$\cdots$	$\cdots$	$\ddots$	$\vdots$
D_m	$s_{m,1}$	$s_{m.2}$	$\cdots$	$s_{m,n}$

Where s_{ij} is the frequency of term T_j in document D_i. The *m*-dimensional vector of the averages by row gives the mean score of similarity of the *m* documents with the whole vocabulary. The highest values of s_{ij} within the matrix represent the most represented terms of the vocabulary in corresponding documents.

The similarity in Vector Space Model is obtained by using associative coefficients based on the inner product of the document vector and query vector, where word overlap indicates similarity. The most popular similarity measure is the cosine coefficient, but also Jaccard and Dice coefficients may be used.

Specifically, the metadata for which this approach was followed are:

- **Language:** by definition the primary human language (or languages) used within learning object to communicate to the intended user. In this case each document is compared with K vocabularies and K matrices were filled. The chosen vocabularies were actual natural language dictionaries publicly available[5] and the maximum mean score of similarity among vocabularies was used to recognize the language.
- **Keyword:** by definition a keyword or phrase describing the topic of learning object. The chosen vocabulary was a publicly available e-learning related vocabulary[6] (any other subject would be suitable as well). For each document, the four terms of the vocabulary with the highest score are considered keyword of that document.
- **Coverage:** by definition the time, culture, geography or region to which this learning object applies. The chosen vocabulary was built with a set of world cities. The city with highest score in each document is considered the coverage of that document. If it is available, a public vocabulary like the Thesaurus of Geographic Names[7] can be used to obtain a more detailed geographical annotation.

EXPERIMENTS, RESULTS AND DISCUSSION

We used 14 papers concerning e-learning written in 7 different languages (English, French, German, Spanish, Italian, Portuguese and Polish) to test the method; of these only 7 [4,14,2,1,17,8,12] are reported here (one for each language), but results for the

[5] http://www.winedt.org/Dict/
[6] http://www3.imperial.ac.uk/ict/services/teachingandresearchservices/elearning/aboutelearning/elearningglossary#x
[7] http://www.getty.edu/research/tools/vocabularies/tgn/

remaining were absolutely similar and coherent to what is shown. The papers were chosen considering availability, language and subject.

Test results are reported in table 1, while top-k scores for Language and Keyword are reported in fig. 1.

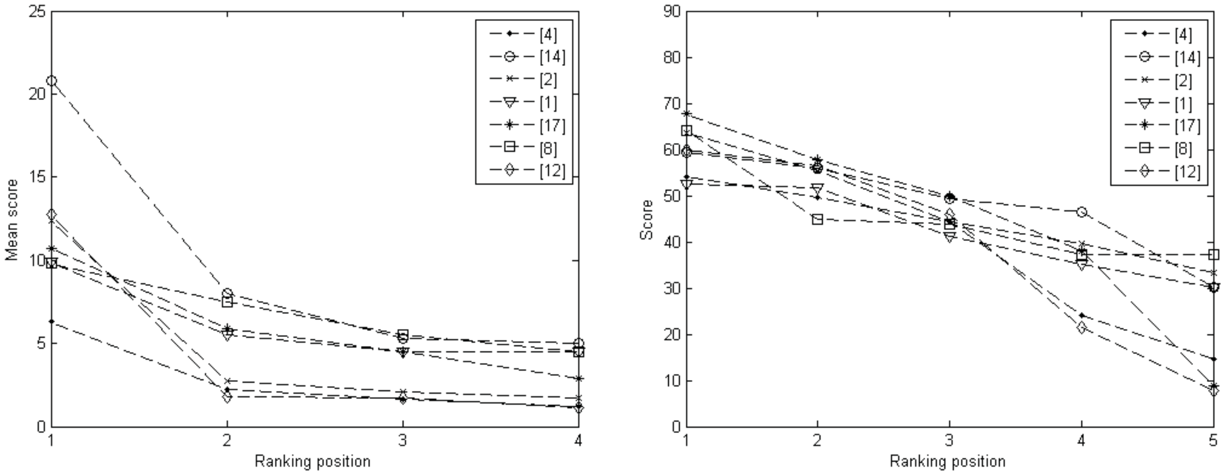

Figure 1 – Top-k scores for language (left) and keyword (right)

Ref.	Title	Language	Description	Keywords	Coverage
[4]	**Height** 15 **Block** 1 **Page** 1 The Sharable Content Object Reference Model (SCORM) - A Critical Review.	**Language** / **Mean Score** English 6.30 Dutch 2.24 French 1.68 Italian 1.17	Learning technology standards are…	**Keyword** / **Score** LMS 54.07 SCORM 49.69 Metadata 44.26 e-Learning 24.20 Reusability 14.61	**City** Kassel **Score** 7.71
[14]	**Height** 20 **Block** 1 **Page** 1 Une architecture orientée services pour la création et le cheminement d'objets pédagogiques de type questionnaire	**Language** / **Mean Score** French 20.82 Polish 8.02 English 5.30 Dutch 4.98	Dans cet article nous proposons…	**Keyword** / **Score** XML 59.44 IMS 55.96 SCORM 49.30 e-Learning 46.46 Distance 30.05	**City** Marseille **Score** 32.98
[2]	**Height** 15 **Block** 3 **Page** 1 E-Learning Standards aus didaktischer Perspektive	**Language** / **Mean Score** Dutch 12.37 English 2.76 Italian 2.05 French 1.73	Einleitung Da der Herstellermarkt…	**Keyword** / **Score** LCMS 63.47 IMS 55.31 e-Learning 44.24 SCORM 39.63 LMS 33.30	**City** Münster **Score** 8.29
[1]	**Height** 17 **Block** 1 **Page** 1 Empaquetamiento y Visualización de Objetos de Aprendizaje SCORM en LMSs de Código Abierto	**Language** / **Mean Score** Spanish 9.88 Portuguese 5.55 Dutch 4.53 English 4.48	En el trabajo se presenta una..	**Keyword** / **Score** SCORM 52.39 LMS 51.59 IMS 41.26 Metadata 35.15 e-Learning 30.13	**City** Valdivia **Score** 8.13
[17]	**Height** 17 **Block** 1 **Page** 1 MOODLE PER L'APPRENDIMENTO LINGUISTICO: elementi critici per una integrazione di sistema	**Language** / **Mean Score** Italian 10.73 French 5.90 Dutch 4.52 English 2.92	L'obbiettivo di questo contributo...	**Keyword** / **Score** Tutor 67.66 SCORM 57.66 e-Learning 49.91 LMS 38.06 Authoring 8.71	**City** Bologna **Score** 38.06

[8]	**Height** 15 **Block** 1 **Page** 1 Web Semântica e e-Learning juntos por uma boa causa	**Language** / **Mean Score** Portuguese 9.80 French 7.52 Spanish 5.53 Dutch 4.53	A sociedade em que vivemos...	**Keyword** / **Score** XML 64.19 e-Learning 44.84 SCORM 43.84 Metadata 37.30 IMS 37.30	**City** Porto **Score** 23.60
[12]	**Height** 13 **Block** 1 **Page** 1 Rozwiązania e-edukacji w zarządzaniu kapitałem ludzkim	**Language** / **Mean Score** Polish 12.74 French 1.81 English 1.74 Dutch 1.10	W artykule przedstawiono…	**Keyword** / **Score** XML 59.76 SCORM 56.51 e-Learning 45.93 LMS 21.46 Education 7.84	**City** Poznan **Score** 18.78

Table 1 – Results

As can be seen from results, in all tested cases the proposed techniques correctly classify language, with a good margin (Figure 1), correctly recognize the title and the abstract (Table 1), correctly classify the place and correctly isolate a good set of keywords. In paper [8] the fifth keyword has the same score of the fourth and should be added to the set.

Method effectiveness clearly depends on the choice of vocabularies and relies on a well structured set of pdf documents like scientific papers, nonetheless it allows a completely blind annotation of content that eases content reusability, creation of Learning Objects and retrieval of information.

CONCLUSIONS AND FUTURE WORK

The use of AMG algorithms for automatic extraction of general metadata from papers has shown to be viable and effective. Results show that in the provided context (scientific papers on the same broad subject) the accuracy of obtained annotations is comparable to that of a human expert. As some metadata are intrinsically unrelated to structure while others are strictly connected to structure, two different techniques were developed, one requiring vocabularies and the other requiring structural features. There are no substantial differences in the accuracy obtained with both. Future work is in trying to handle more complex metadata and less structured documents.

REFERENCES

[1] L.A. Alvarez G., D.P Espinoza P. and S.G. Bucaraya A. *Empaquetamiento y Visualización de Objetos de Aprendizaje SCORM en LMSs de Código Abierto*. Primera Conferencia Latinoamericana de Objetos de Aprendizaje, pp. 1-10, 2006.

[2] P. Baumgartner, H. Häfele and K. Maier-Häfele. *E-Learning Standards aus didaktischer Perspektive*. In: Campus 2002: Die virtuelle Hochschule in der Konsolidierungsphase. G. Bachmann, O. Haefeli und M. Kindt. Münster, Waxmann. 18: pp. 277-286, 2002.

[3] K. Bird, and the Jorum Team. Automated Metadata - A review of existing and potential metadata automation within Jorum and an overview of other automation systems. 31st March 2006, Version 1.0, Final, Signed off by JISC and Intrallect July 2006.

[4] O. Bohl, J. Schellhase, R. Sengler and U. Winand. *The Sharable Content Object Reference Model (SCORM) - A Critical Review*. Proceedings of the International Conference on Computers in Education (ICCE'02), pp. 950–951, vol. 2, 2002.

[5] S. Brin and L. Page. *The anatomy of a large-scale hypertextual Web search engine*. Computer Networks and ISDN Systems, vol. 30, pp. 1-7, 1998.

[6] L.F.H. Edvardsen, I.T. Sølvberg, T. Aalberg and H. Trætteberg. *Using Automatic Metadata Generation to reduce the knowledge and time requirements for making SCORM Learning Objects*. Proceedings of the 3rd IEEE International Conference on Digital Ecosystems and Technologies, pp. 392-397, 2009.

[7] G. Giuffrida, E. C. Shek, and J. Yang. *Knowledge-Based Metadata Extraction from PostScript Files*. Proceedings of the Fifth ACM Conference on Digital Libraries, pp. 77-84, 2000.

[8] V. Gonçalves and E. Carrapatoso. *Web Semântica e e-Learning juntos por uma boa causa*. 8th International Symposium on Computers In Education, 1: pp. 1-10, 2010.

[9] J. Greenberg. *Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications*. Journal of Internet Cataloging, 6(4): pp. 59-82, 2004.

[10] C. Jenkins, and D. Inman. *Server-side Automatic Metadata Generation using Qualified Dublin Core and RDF*. Proceedings of International Conference on Digital Libraries: Research and Practice, pp. 262-271, 2000.

[11] A. Kawtrakul and C. Yingsaeree. *A Unified Framework for Automatic Metadata Extraction from Electronic Document*. Proceedings of IADLC2005, pp. 71-77, 2005.

[12] R. Klaus, M. Dyks. *Rozwiązania e-edukacji w zarządzaniu kapitałem ludzkim*. Komputerowo Zintegrowane Zarządzanie R. Knosala, Oficyna Wydawnicza PTZP,vol. 1, pp. 671-675, 2010.

[13] E.D. Liddy, E. Allen, S. Harwell, S. Corieri, O. Yilmazel, N.E. Ozgencil, A. Diekema, N.J. McCracken, J. Silverstein, and S.A. Sutton. *Automatic metadata generation and evaluation*. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.401-402, 2002.

[14] I. Madjarov, A. Betari, B. Shishedjiev, Z. Bakkoury. *Une architecture orientée services pour la création et le cheminement d'objets pédagogiques de type questionnaire*. Premier Congrès International Technologies Numériques de l'Information et de Communication Educatives – Expériences et Perspectives (TNICE-EP'2007), Marrakech, Maroc, 2-4 mai 2007.

[15] C. Ramakrishnan, A. Patnia, E. Hovy and G.A. Burns. *Layout-Aware Text Extraction from Full-text PDF of Scientific Articles*. Source Code for Biology and Medicine, Vol. 7, No. 1, 2012.

[16] K. Seymore, A. McCallum and R. Rosenfeld. *Learning hidden Markov model structure for information extraction*. Proceedings of Workshop on Machine Learning for Information Extraction, pp. 37-42, 1999.

[17] A. Valdivieso, V. Preti. *MOODLE PER L'APPRENDIMENTO LINGUISTICO: elementi critici per una integrazione di sistema.* Conferenza nazionale italiana Moodlemoot, 2010.

ABOUT THE AUTHOR

Antonio Maratea, University of Naples "Parthenope", Department of Applied Science, Centro Direzionale di Napoli, Isola C4, 80143 Napoli, Italy, e-mail: antonio.maratea@uniparthenope.it

Alfredo Petrosino, University of Naples "Parthenope", Department of Applied Science, Centro Direzionale di Napoli, Isola C4, 80143 Napoli, Italy, e-mail: alfredo.petrosino@uniparthenope.it

Mario Manzo, University of Naples "Parthenope", Department of Applied Science, Centro Direzionale di Napoli, Isola C4, 80143 Napoli, Italy, e-mail: mario.manzo@uniparthenope.it