

Karampiperis P. and Sampson D. (2003). A Schema-Mapping Algorithm for Educational Metadata Interoperability. *In Proc. of the 15th World Conference on Educational Multimedia, Hypermedia and Telecommunications ED-MEDIA 2003*, Honolulu, Hawaii, USA.

A Schema-Mapping Algorithm for Educational Metadata Interoperability

Pythagoras Karampiperis, Kostas Kastradas, Demetrios Sampson
Informatics and Telematics Institute, Centre for Research and Technology Hellas,
42, Arkadias Street, Athens, GR-15234, Greece
and
Department of Technology Education and Digital Systems, University of Piraeus
150, Androutsou Street, Piraeus, GR-18534 Greece
e-mail: {pythk, kkastrad, sampson}@iti.gr

Abstract: The interoperability between systems and the reusability of the stored and managed information is a key issue when designing an educational metadata management system. Converting from one data representation to another is time-consuming and labor-intensive, with few tools available to ease the task. In this paper we present a schema-mapping algorithm that is capable of transforming the metadata description of an educational object created using specific metadata specification to a representation of the same object using either another specification (for example from Dublin Core to IEEE LOM and vice versa), or the same specification in another language (for example mapping between different translations of the IEEE LOM standard). We show that using advanced conversion techniques such as the proposed algorithm can result in lower standardization efforts. Simulation results over a wide range of learning objects demonstrate that the proposed algorithm except from requiring less human intervention, can handle real case mapping problems between not only different metadata models but also between models that use different standards for data representation.

Introduction

The main goal when designing a metadata management system is to achieve interoperability between similar systems, so as to be able to reuse the stored and managed information, both at a lower representation level (physical level) and at the level of description and organization (logical level). The first goal can be achieved using standard interchange technologies such as XML (Extensible Mark up Language). The second goal can be achieved by adopting commonly agreed learning technology specifications. Although today a generally accepted international standard for describing educational material exists, namely the IEEE Learning Object Metadata standard, many metadata management systems are still using other metadata models for describing learning objects (Dublin Core, Ariadne Educational Metadata Recommendation, GEM Element Set) or previous versions of the IEEE LOM standard [S. Sutton, 1999].

Furthermore, the internationalization of each specification defined by the CEN/ISSS Learning Technologies Workshop as the sum of processes whose purpose is to facilitate search, evaluation, reusability, and processing of learning objects within a multicultural and multilingual scenario, lead to the existence of multiple translations of each specification, providing evidence that two systems may not be able to interact, even when they use the same learning object metadata specification.

A possible solution to this problem is to define other specifications describing the interoperability issues between different guidelines or the internationalization issues of a specific standard, but this implies extra effort and extra cost. Another more simple solution is to use schema-mapping algorithms that are capable of transforming the metadata description of an educational object created using a specific standard to a representation of the same object using another specification, or another language.

In this paper, we propose a schema-mapping algorithm that is tested to be effective when applied to a wide range of learning objects; and does not require any information about the source or the destination standard schemas; or any special knowledge from the user of the mapping platform.

In the next section we discuss the relationship between standardization and conversion. We will show that using advanced conversion techniques such as the proposed algorithm we result in lower standardization efforts. The third section presents the design philosophy as well as the details of the proposed schema-mapping algorithm. The fourth section discusses the simulation results of the proposed algorithm when is used for the transformation of metadata repositories storing educational objects, against IBM's Translator Generator; a well-known completely automatic tool for schema-mapping. Finally, we present our conclusions on the use of the proposed schema-mapping algorithm.

Standardization and Conversion

The basic function that underlies systems intercommunication is the exchange of information. The major barrier that prevents system intercommunication, limiting the interoperability between metadata management systems, is the use of different specifications that define the structure of the exchanged information (*standardization diversity*). However, assuming that two systems use the same standardization format, interoperability cannot be ensured if this common format is described in different natural languages (*internationalization problem*).

In both cases, there are two possible ways in order to achieve interoperability between educational metadata management systems: either the use of a neutral, standardized format or a conversion between varying formats (either different standards or same standards that are described with different languages) [E. Wustner et al., 2002].

Figure 1 shows overall standardization costs and overall conversion costs depending on the standardization level. Standardization costs contain all the costs that are necessary to implement a standard [T. Buxmann et al., 1999], e.g., software costs, hardware costs and personnel costs. Obviously standardization costs are proportional to the level of standardization.

The graph of overall conversion costs is just reversed, since with high standardization hardly any conversion is necessary, whereas precise conversions between multitudes of specifications and guidelines cause comparatively high costs. The overall conversion costs, as schematized in figure 1, are the sum of:

- *Costs for generating the converter:* These accrue through developing the necessary software and through acquiring a thorough knowledge of the data that has to be converted. The process of acquiring this kind of knowledge is mentioned by CEN/ISSS Learning Technologies Workshop that addresses the problem of internationalization in the case of the IEEE LOM standard. : ‘...internationalization starts with a study on the capability of each data element to support different localizations, both in terms of applicability to diverse languages (multilinguality) and to different cultural contexts..’ [G. Da Bormida et al., 2002].
- *Costs resulting from an insufficient conversion result:* These costs can occur if the conversion instrument is error-prone or information loss could not be avoided. The probability of information loss obviously increases with the heterogeneity of the used metadata schemas. Costs resulting from insufficient results include expenses for manual post editing of the conversion result.

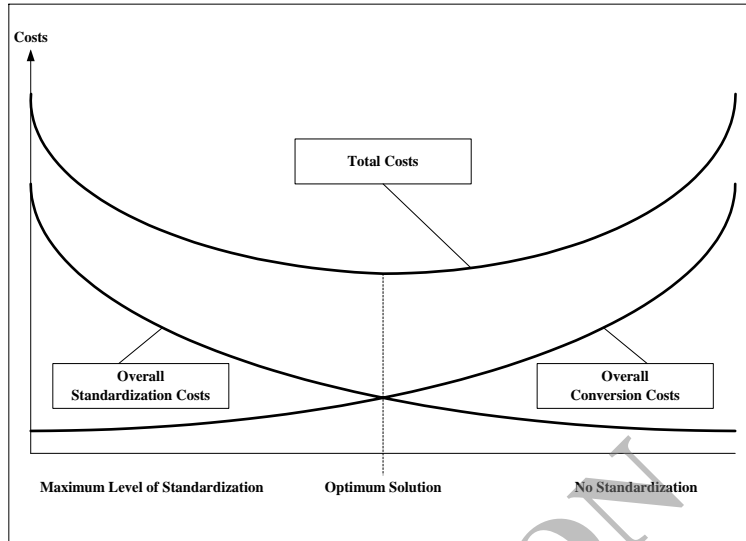


Figure 1: Trade-off between standardization costs and conversion costs

Apparently there is a trade-off between overall costs of standardization and overall conversion costs. The optimum is where the sum of standardization and conversion costs is minimal. We make the assumption that the use of the mapping algorithm implicates a right-shift of the overall conversion cost, as illustrated in figure 2. This right-shifting is due to the fact that the conversion costs have been reduced, since there is less need required for the user to interfere in the conversion process, thus reducing the cost of acquiring knowledge of the data that needs to be converted. On the other hand, this right shifting also implicates further reduce of the needed standardization costs. This does not mean that no standardization of the educational metadata is needed, but fewer efforts are required to support the exchange of information in the context of internationalization.

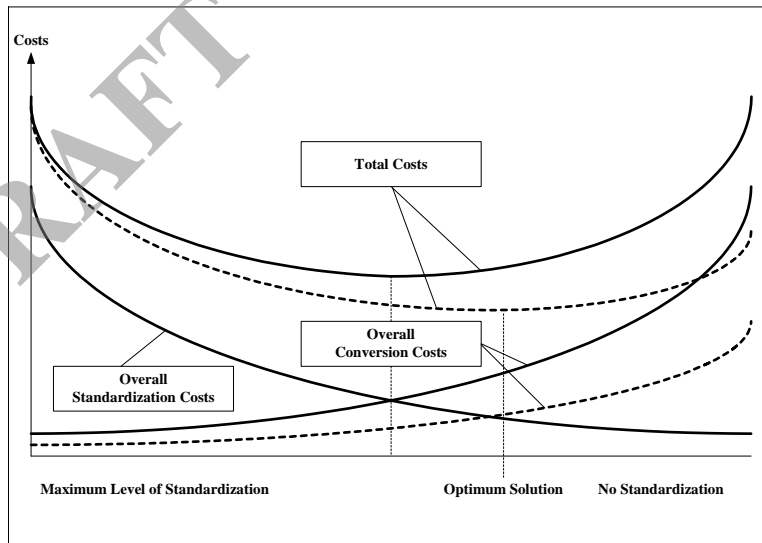


Figure 2: Trade-off between standardization costs and conversion costs when a schema-mapping algorithm is applied

The Mapping Algorithm

For solving the mapping problem between two different schemas representing the same real-world entity, we can use two approaches:

- Attribute-Driven, The mapping is based on the names of the attributes and not on the values that they hold.

- Data-Driven, The mapping is based on the similarity of the data values that the attributes hold.

The Data-Driven methods have better performance since the corresponding map can be the result of comparing more than one example. This property does not exist in Attribute-Driven methods, which produce the mapping only by comparing the name of the attributes between the two given schemas. The two categories of methods have comparable performance when only one input example is used by a Data-Driven method.

The proposed algorithm is Data-Driven and consists of three nested parts. The main part of the algorithm produces the mapping and requires a measurement of the similarity between two attributes – second part, which in turn requires a measurement of the similarity between two tokens – third part.

So, suppose that two different schemas (schema A and schema B), that describe the same real-world entity, are given. The algorithm has the form shown below:

- Step 1: Select the schema that contains the largest number of elements. Let this schema, be schema A.
- Step 2: Select the first element of schema A
- Step 3: Select the first element of schema B
- Step 4: Compute the similarity between the two selected elements. If the similarity is greater or equal than the similarity threshold parameter – defined by the user – accept the mapping between the two elements, else not.
- Step 5: Are there any other non selected elements in schema B? If yes, select the next element and repeat Step 4. If not, move to next step.
- Step 6: Are there any other non selected elements in schema A? If yes, select the next element and repeat Step 3. If not, end of algorithm.

For the measurement of the similarity between two elements (let them be elements e_1 and e_2), we have:

- Step 1: Select the element with the smallest number of token. Let this be element e_1 .
- Step 2: Select the first token of element e_1 .
- Step 3: Select the first token of element e_2 .
- Step 4: Compute the similarity between the two selected tokens.
- Step 5: Are there any other non selected tokens in element e_2 ? If yes, select the next token; if no, set as similarity of the selected token of element e_1 , the maximum value of similarity between the token of e_1 and the tokens of element e_2 and proceed to next step.
- Step 6: Are there any other non selected tokens in element e_1 ? If yes, select the next token and repeat Step 3. If not, end the algorithm with total similarity, between two elements e_1 and e_2 , equal to the average of the similarity of the tokens that compose element e_1 .

For the measurement of the similarity between two tokens (let them be token t_1 and t_2), we have:

- Step 1: Select the token with the smallest number of characters. Let this token be t_1 .
- Step 2: Check if token t_2 , starts with the same characters as token t_1 . If yes, then finish the algorithm with similarity between the two tokens equal to the number of characters of the remaining token to the number of characters of the starting token t_1 . If no, then remove the last character from token t_1 and repeat Step 2. If token t_1 has only one character then instead of repeating Step 2, finish the algorithm with similarity equal to zero.

It is clear that the performance of the algorithm depends on the selection of the similarity threshold parameter, which represents the minimum similarity of two paths that should be mapped. A greater value will give less but more secure mappings, a smaller value will produce more mappings but with greater probability of making a mistake. So the best practice when selecting a specific value for the similarity threshold parameter is to select the minimum possible value in order to have maximum number of connections, without affecting the correctness of the mappings.

Case Study: Transforming Metadata Repositories for Educational Objects

The main problem of most educational metadata management systems [D. Sampson et al., 2002] is the existence of a large number of learning resources that are already described in a non-standard manner and therefore describing them in globally accepted way is a very time-consuming and expensive process. Although a generally accepted standard for describing educational material (IEEE Learning Object Metadata) exists, many educational metadata management systems are using other guidelines; or previous versions of the IEEE standard; or even different translations of the IEEE LOM. This fact makes the interoperability between those systems almost impossible. Therefore, in order to increase the interoperability between educational metadata management systems, efficient tools and procedures that will enable the transformation between different metadata schemes should be used.

In traditional data integration tools such as CLIO [L. Popa et al., 2002], a semi-automatic schema mapping tool developed at the IBM Almaden Research Center, the users of the system are required to understand both their source's data and the target representation, in order to manually create a mapping between them. Other tools have been presented, trying to completely automate the mapping process such as IBM's Translator Generator, but these tools are expecting from the user to describe with completely the same way the content of the different schemas.

The proposed algorithm except from being fully automatic, handles in a more precise way real case mapping problems such as the mapping between different educational metadata description schemes that uses not only different description schemas but also different standards for representing the content information [D. Sampson and P. Karampiperis, 2003].

In this section we are going to present the simulation results of the mapping algorithm previously described when the algorithm is used to map one educational metadata schema to another or same educational schemas that use different languages to describe them, against IBM's Translator Generator.

Setting the Simulation Platform

In our simulation experiment we try to map different representations of the same real-world entity using the Dublin Core, Ariadne, Gem, IEEE LOM (English) and IEEE LOM (Greek) educational metadata models. In order to examine the efficiency of the proposed algorithm, we have designed several datasets for each one of the five previously mentioned educational metadata schemes. For the designing of the testing datasets we used the ISO 639 and ISO3166-1 standards as the language format scheme and the ISO 8601 standard as the date format scheme, according to all metadata schemes of our example.

An example of those datasets, which shows that although the different schemes are describing the same learning objects, they can use also different structures for the representation of the same content, is shown in Table 1.

Metadata Scheme	Element	Value
Dublin Core	Title	Graphic Java Mastering the JFC
	Creator	David Man Geary
	Subject	Programming Languages
	Publisher	Prentice Hall Publications
	Date	1999-11-25
	Language	en-US
IEEE LOM	Title	Graphic Java Mastering the JFC
	Language	en-GB
	Keyword	Programming Languages
	Role	Author
	Family Name	Geary
	Given	David
	Additional Name	Man
	Role	Publisher
	Organization Name	Prentice Hall Pub.
	Date	1999-11-25
	Age Range	25-45
Intenced End User Role	Learner	
ARIADNE	Title	Graphic Java Mastering the JFC
	Family Name	Geary
	Given Name	David
	Additional Name	Man
	Prefix	Mr.
	Publication Date	1999-11-25
	Language	en-GB
	End User Type	Learner
GEM	Institution	
	Age	25-45
	Role	Author
	Creator	Man Geary, David
	Date	1999-11-25
	Language	en-GB
	Role	Publisher
	Publisher	
	Keyword	Programming Languages
Title	Graphic Java Mastering the JFC	

Table 1: An example of a Testing Dataset.

Simulation Results

As already explained, the efficiency of the proposed mapping algorithm depends on the selection of the similarity threshold parameter. In order to test the robustness of the algorithm for several different values of the similarity threshold parameter, we used three different simulation scenarios. The first one uses the value of 0.5 for the similarity threshold parameter and the other two the value of 0.6 and 0.7 respectively. In order to evaluate the total efficiency of the proposed algorithm and to compare the mapping results with those of IBM's Translator Generator we have designed three different evaluation criteria, which are defined by:

$$\text{Confidence} = \frac{\sum \text{Similarity of Content}}{\forall \text{mapping} \text{ Total number of mappings produced}}$$

$$\text{Success} = \frac{\text{Number of correct mappings}}{\text{Total number of mappings produced}}$$

$$\text{Mistakes} = \frac{\text{Number of wrong mappings}}{\text{Total number of mappings produced}}$$

All of them are used to evaluate a total criterion that is the mean value of confidence, success and mistakes for all of the different educational metadata description schemes. It is obvious that the efficiency of the mapping algorithm depends also on the similarity between the entity values of the different schemas, on which the algorithm is applied. We have split the testing datasets in three categories according to the measure of the similarity between entity values, which is defined by:

$$\text{DataSetSimilarity} = \frac{\sum_{\forall \text{entity of the DataSet}} \text{Similarity}}{\text{Total number of entities in DataSet}}$$

So we have identified datasets with low (less than 40%), medium; and high (more than 70%) similarity. In Figures 3a, b and c the results of the proposed mapping algorithm for the three categories of the datasets are shown respectively; for the three simulation scenarios of the similarity threshold parameter.

In all out tests, selecting a threshold with value over than 0.7 gave the same indicator values with those of threshold with value equal to 0.7. Since our main objective when selecting a specific value for the similarity threshold parameter is to select the minimum possible value in order to have maximum number of connections and maximum total indicator (min-max criterion), the best similarity threshold value seems to be that of 0.7.

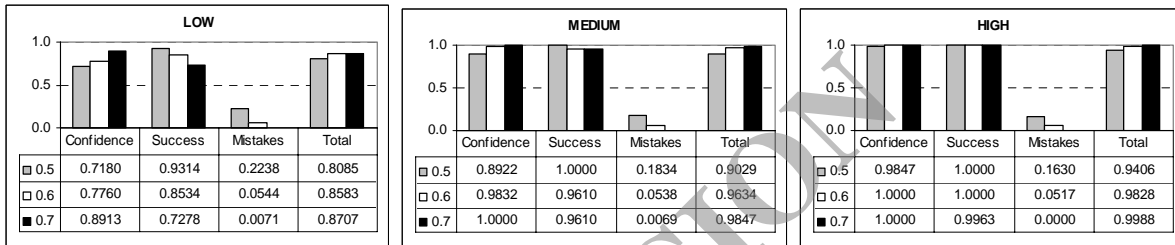


Figure 3a, b and c: Results of the proposed algorithm applied to datasets with low, medium and high similarity between entity values respectively

In Figures 4 a, b and c the evaluation criteria as well as the total evaluation results for both the proposed algorithm and the IBM's Translator Generator are shown. From the simulation results it is clear that the proposed schema-mapping algorithm succeeds in producing more correct mappings than those produced by a general-purpose schema-mapping platform such as the IBM's Translator Generator, and thus it has a better total evaluation indicator.

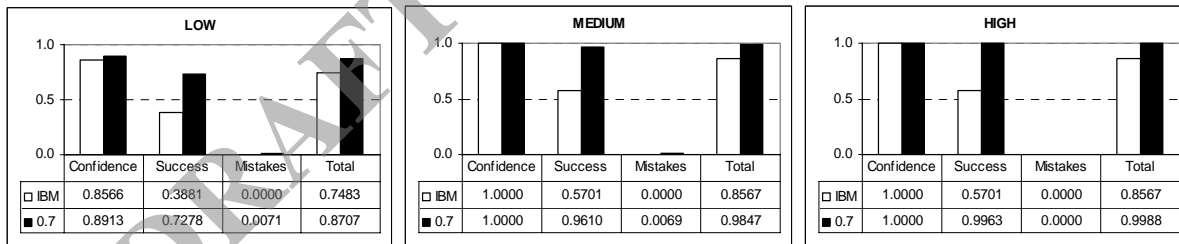


Figure 4a, b and c: Comparison results of the proposed algorithm and the IBM's Translator Generator applied to datasets with low, medium and high similarity between entity values respectively

In order to make a more extensive examination of the efficiency of the proposed algorithm, we created full datasets of LOM and Dublin Core metadata schemes representing the same learning objects. The result of the mapping algorithm using any similarity threshold parameter above 0.6 was according to Annex B (Mapping to Unqualified Dublin Core) of IEEE 1484.12.1-2002 (Draft Standard for Learning Object Metadata) standard, in which the mapping between LOM and Dublin Core metadata schemes is defined. This fact proves that the automatic, without the user interference, mapping between two different schemas is not impossible and also that this process can be as effective as the manual mapping of a very experienced user.

Conclusions

Converting from one data representation to another is time-consuming and labor-intensive, with few tools available to ease the task. In traditional data integration tools the users of the system are required to understand both their source's data and the target representation, in order to manually create a mapping between them. Other tools have been presented, trying to completely automate the mapping process, but these

tools are expecting from the user to describe with completely the same way the content of the different schemas. The proposed algorithm except from being fully automatic, handles in a more precise way real case mapping problems such as the mapping between different educational metadata description schemes that use not only different description schemas but also different standards for representing the content information, or the mapping between metadata descriptions that represent the same real-world entity in different languages.

References

IEEE Draft Standard for Learning Object Metadata, IEEE P1484.12.1/d6.4, 2002.

Dublin Core Metadata for Resource Discovery. Internet RFC 2413. URL address: <http://www.ietf.org/rfc/rfc2413.txt>, accessed December 2002.

ARIADNE project. URL address: <http://ariadne.unil.ch>, accessed December 2002.

S. Sutton, "Conceptual Design and Deployment of a metadata framework for educational resources on the Internet". *Journal of the American Society for Information Science* 50(13), pages 1182-1192, 1999.

E. Wustner, T. Hotzel & P. Buxmann, "Converting Business Documents: A Classification of Problems and Solutions using XML/XSLT", *Proceedings of the 4th IEEE International Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems (WECWIS) 2002*.

T. Buxmann, F. Weitzel, F. Westarp & W. Konig, "The Standardization Problem in Networks – A General Framework". In Jakobs K., (eds.): *Standards and Standardization: A Global Perspective*, Idea Publishing Group, 1999.

G. Da Bormida, E.Ovcin & L.Anido-Rifon, "Internationalization of the IEEE Learning Object Metadata", *CEN/ISSS Information Society Standardization System, Learning Technologies Workshop*, URL address: <http://www.cwnorm.be/iss/Workshop/lt/>, accessed December 2002.

D. Sampson, V. Papaioannou & P. Karadimitriou, "EM2: An environment for editing and management of educational metadata, Special Issue on Innovations in Learning Technologies", *Educational Technology and Society Journal of International Forum of Educational Technology and Society and IEEE Computer Society Learning Technology Task Force*, 5(4), October 2002.

L. Popa, M.A. Hernandez, Y. Velegrakis, R.J. Miller, F. Naumann & H. Howard, "Mapping XML and Relational Schemas with Clio", *Proceedings of the 18th IEEE International Conference on Data Engineering (ICDE) 2002*.

IBM's Translator Generator, IBM AlphaWorks URL Address: <http://www.alphaworks.ibm.com/tech/>, accessed December 2002.

D. Sampson, P. Karampiperis, "Reusable Learning Resources: Developing a Metadata Management System supporting Interoperable Learning Object Repositories", in Rory McGreal (Editor), *Online Education Using Learning Objects*, Morgan Kauffmann, 2003.