On the Use of Learning Object Metadata: The GLOBE Experience

Xavier Ochoa[1], Joris Klerkx[2], Bram Vandeputte[2], and Erik Duval[2]

[1] Information Technology Center, Escuela Superior Politecnica del Litoral,
Guayaquil, Ecuador
xavier@cti.espol.edu.ec
[2] Dept. Computerwetenschappen, Katholieke Universiteit Leuven,
Celestijnenlaan 200A, B-3001, Heverlee, Belgium
{Joris.Klerkx,Bram.Vandeputte,Erik.Duval}@cs.kuleuven.be

Abstract. Since IEEE LTSC LOM was published in 2002, it is one of the widest adopted standard for the description of educational resources. The GLOBE (Global Learning Objects Brokered Exchange) alliance enables share and reuse between several Learning Object Repositories worldwide. Being the largest and more diverse collection of Learning Object Metadata, it is an ideal place to perform an analysis of the actual use of the LOM standard in the real world. This paper presents an in-depth analysis of the use and quality of 630.317 metadata instances.

Keywords: Interoperability, Learning Object Metadata.

1 Introduction

The Learning Object Metadata standard for describing learning resources has been published in 2002. It is based on early metadata schemes that were developed by the ARIADNE Foundation [1] and the IMS Global Learning Consortium [2]. Main goal of LOM is "to facilitate search, evaluation, acquisition, and use of learning objects, for instance by learners, instructors or automated software processes" [3] through the description of a common set of metadata elements. LOM proposes around 50 different metadata elements grouped into nine categories: General, Lifecycle, Meta-Metadata, Technical, Educational, Rights, Relation, Annotation and Classification.

As important as the LOM standard is for sharing and reusing learning materials, very little research has been performed on the actual use of the standard in real-world situations. To our knowledge, only three studies have been performed on the use of LOM. Neven and Duval [4] made a first survey of LOM based repositories, but only considered the use of qualitative criteria about the repository infrastructure and did not consider the actual use of metadata. Najjar and Duval [5] performed a quantitative study of the use of metadata for indexing objects in ARIADNE. While the original ARIADNE format was similar to LOM, it was simpler, lacked several key elements (e.g. description and learning resource type) and had different ways to store the information (i.e. the

C. Delgado Kloos et al. (Eds.): EC-TEL 2011, LNCS 6964, pp. 271–284, 2011.

fixed Semantic Section in ARIADNE versus the flexible Classification category in LOM). Another effort to characterize LOM usage was performed by Friesen [6], who collected samples from several repositories worldwide to study their usage of LOM. However, this survey has deep methodological flaws (e.g the metadata was not randomly selected, but hand-picked in each repository) and was based on a very small sample that weakens the validity of the conclusions drawn by the author. After this last work, there has been a large silence in the research community about the actual use of LOM.

This lack of research has not stopped the wide adoption of LOM as the way in which most Learning Object Repositories (LORs) represent their metadata internally, or the way in which they interchange their metadata with search facilities and other repositories. The largest organization that uses LOM as a common medium to enable the sharing of learning materials is GLOBE (Global Learning Objects Brokered Exchange). Jointly, nearly 1,2 million learning objects are shared inside the GLOBE alliance. Being the largest and more diverse collection of Learning Object Metadata, GLOBE is an ideal place to perform an analysis of the actual use of the LOM standard in the real world. This paper presents a large-scale study of the use and quality of more than 50% (630 317) of LOM instances in GLOBE.

We start this paper in Section 2 with information about the collection of data. Section 3 presents basic statistics about the metadata. A study how the various LOM elements and their values have changed over time the presence of new data elements is presented in Section 4. Section 5 performs a deeper study on the (mis-)use of LOM and the actual information that is contained in the metadata. We conclude in Section 6 with lessons learned and future work.

2 Data Collection

To obtain a representative sample of LOM metadata, approximately half of the instances of GLOBE has been obtained by using the OAI-PMH harvester tool developed in ARIADNE. In total, 630 317 LOM metadata instances were harvested from those GLOBE repositories that provide a metadata harvesting service based on the OAI-PMH protocol [7]. All data, code and figures used in this work can be downloaded[1]. The obtained metadata have a high degree of diversity which is the result of different methodologies used to create and aggregate metadata in the different repositories. For example, KOCW, LACLO and OUJ expose a single collection of metadata instances with a common provenance. Other GLOBE repositories expose the result of the aggregation of several metadata collections that have different provenance. This is the case of ARIADNE, where only a 5% of its objects has been produced by its members and the remaining 95% has been collected from other repositories worldwide. LRE, OER and LORNET also aggregate metadata instances from different provenances, but those are altered and enriched after-collection to follow a repository-wide standard. Another example of diversity is the methodology used to create the metadata. While most

[1] http://ariadne.cs.kuleuven.be/lomi/index.php/LOM_in_GLOBE

Table 1. LOM Repositories studied

Repository	Instances	Provenance	Creation
ARIADNE Foundation	374 857	Aggregated	Mostly Manual
Learning Resource Exchange (LRE)	169 736	Enriched	Manual
Community on Learning Objects (LACLO)	49 943	Single	Automatic
OER Commons (OER)	25 794	Enriched	Manual
Korean OCW (KOCW)	7 183	Single	Manual
LO Repository Network (LORNET)	1 804	Enriched	Manual
Open University Japan (OUJ)	1 000	Single	Manual

metadata are created manually, some repositories, like ARIADNE, have a non-negligible amount of instances that were produced by semi-automatic metadata generators [8]. An extreme case is LACLO which is exclusively composed of automatically generated metadata. While this diversity makes it harder to analyze the metadata instances as a whole, it is unavoidable in a real-world scenario and it makes the result of our analysis richer and more relevant. Table 1 lists the details from those repositories and the number of instances obtained from them.

3 Basic Statistical Analysis

3.1 LOM Size

The first useful analysis that can be conducted over the collected data is to determine the average size of a LOM instance. The size of the instance is determined by the binding in which it is represented. In this study, we used the XML binding of LOM. The arithmetic mean of the size of the XML files is 4,25 Kb. However, the distribution of the sizes does not follow a normal distribution. The size distribution for each repository, as well as for the aggregated set, presents a right tail (positive skewness). As suggested in literature [9], this distribution can be approximated (but not exactly fitted) by a log-normal distribution, meaning that if the logarithm of the file size is taken, the obtained values follow approximately a normal distribution. An ANOVA analysis of the size of the files in the different repositories shows that there is no major difference in the size of the LOM instances between the GLOBE repositories. The average size and the distribution of values could therefore be used to model the space requirements for LORs and the capacity needed to interchange those instances over the network.

3.2 LOM Data Elements Usage

Our second analysis is a frequency analysis of the data elements in LOM. For this analysis, only top-level fields were counted. For example, in the XML binding of LOM, the field *General.Structure* has two subfields: *General.Structure.source* and *General.Structure.value*. In this case, only the number of appearances of *General.Structure* is counted. The justification for this is that not all repositories have the *source* subfield and, in most cases, the number of appearances of

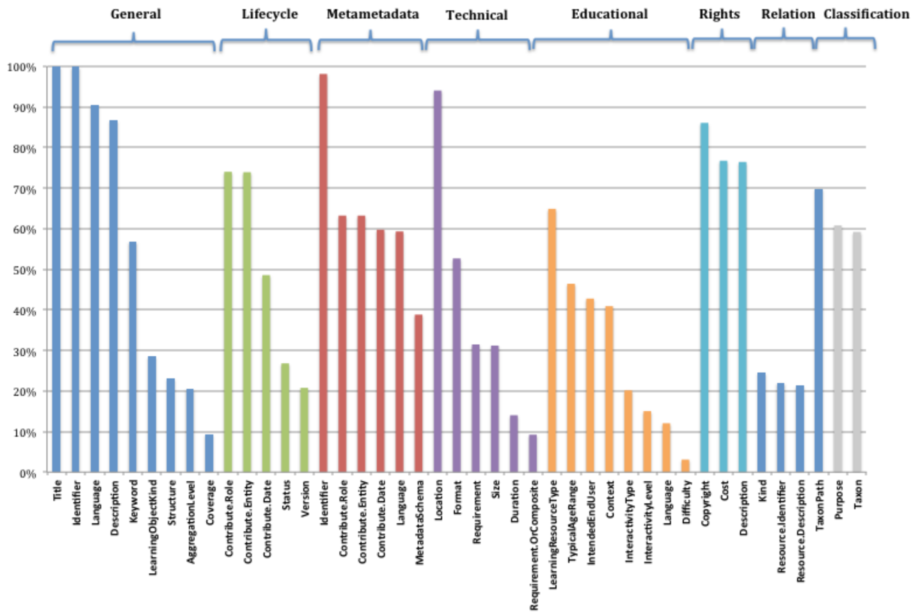
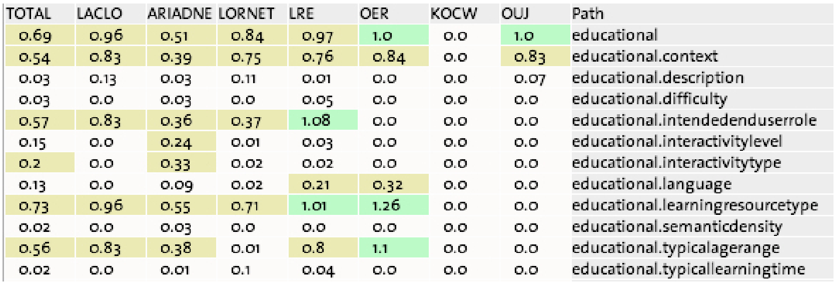


Fig. 1. Percentage of Usage of Different LOM Data Elements in GLOBE

top-level fields is equal to the number of appearances of the *value* subfield. If the data element is filled more than once, it is counted only once. Figure 1 shows the percentage of metadata that have a value for the different data elements of LOM in GLOBE.

Main finding of this study is that only a small fraction of the LOM standard is frequently used to describe the learning objects. Only 20 of the 50 data elements are used more than 60% of the time. Also, 16 data elements are used less than 10% of the time. At first sight, this result seems to corroborate the Friesen study [6] conclusion that the added complexity of LOM is not used in the real-world. However, a similar study performed by Wand [10] over almost 1 million metadata instances from the Open Archives Initiative (OAI) that uses the much simpler Dublin Core (DC) [11] metadata schema, found that of 15 DC data elements, five (creator, identifier, title, date, and type) are used 71% of the time and the five least used elements (language, format, relation, contributor and source) are used less than 6% of the time. Contrasted with this last study, it seems that LOM, while being more complex, helps to collect proportionally more information than a simpler and more general schema such as DC. The inequality of usage of the different data elements seems to be something inherent to the creation of metadata. This inequality deserves further research, through a comparative analysis of the use in different metadata standards.

Considering that LOM is specifically designed to describe educational material, it is important to review the use of the Educational section. 4 out of 11 educational data elements (Learning Resource Type, Intended End User Role, Typical Age Range and Context) are used more than 40% of the time, 3 elements



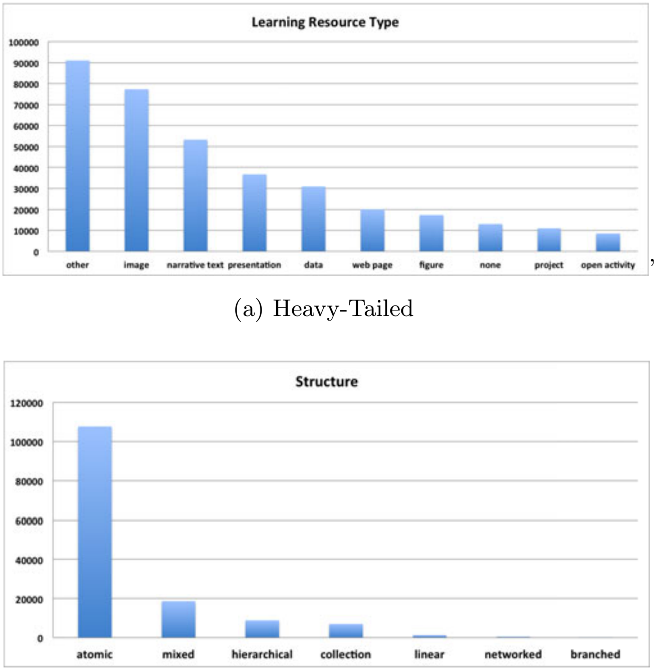
TOTAL	LACLO	ARIADNE	LORNET	LRE	OER	KOCW	OUJ	Path
0.69	0.96	0.51	0.84	0.97	1.0	0.0	1.0	educational
0.54	0.83	0.39	0.75	0.76	0.84	0.0	0.83	educational.context
0.03	0.13	0.03	0.11	0.01	0.0	0.0	0.07	educational.description
0.03	0.0	0.03	0.0	0.05	0.0	0.0	0.0	educational.difficulty
0.57	0.83	0.36	0.37	1.08	0.0	0.0	0.0	educational.intendedenduserrole
0.15	0.0	0.24	0.01	0.03	0.0	0.0	0.0	educational.interactivitylevel
0.2	0.0	0.33	0.02	0.02	0.0	0.0	0.0	educational.interactivitytype
0.13	0.0	0.09	0.02	0.21	0.32	0.0	0.0	educational.language
0.73	0.96	0.55	0.71	1.01	1.26	0.0	0.0	educational.learningresourcetype
0.02	0.0	0.03	0.0	0.0	0.0	0.0	0.0	educational.semanticdensity
0.56	0.83	0.38	0.01	0.8	1.1	0.0	0.0	educational.typicalagerange
0.02	0.0	0.01	0.1	0.04	0.0	0.0	0.0	educational.typicallearningtime

Fig. 2. Heat map table comparing the usage of educational elements

(Language, Interactivity Level and Interactivity Type) are used more than 10% but less than 20%, and the 4 remaining elements (Description, Difficulty, Semantic Density and Typical Learning Time) are used less than 10% of the time. While these values are lower than desirable, they provide proof that LOM is used in the real-world to capture educational information about digital objects. Current research on automatic generation and enrichment of LOM that is based on the context and use of the resource [8] seems to be a way to increase the completeness of the Educational section. To easily find differences in the usage of the educational data elements across the different providers, a heat map view is presented in Figure 2. It presents the relative frequency of the use of a given data element in each studied repository. Values higher than 1 represent that the described element is commonly used more than once in each instance. The first column is the average relative frequency over all the repositories. At this level, it is easy to see the large diversity with respect to the use of the Educational section of LOM that was hidden in the previous general analysis. There are two different repositories, LRE (manually enriched metadata) and LACLO (automatically generated metadata), that use several educational data elements in more than 70% of the cases. Other repositories, such as KOCW and OUJ provide almost no information in this section. ARIADNE, being in itself a collection of several other repositories, mirrors very closely the results of GLOBE as a whole. This suggests, by an admittedly unsafe extrapolation, that the use of the Educational section found in the studied set would be similar to the one existent in the theoretical universe of LOM instances.

3.3 LOM Vocabulary Usage

Vocabulary elements are data elements that can only be filled with a limited set of values established by the metadata standard. The main goal of these vocabulary elements is to provide a higher level of semantic interoperability [3]. In this analysis, the distribution of permitted values for the vocabulary elements has been obtained by parsing the XML instances in a similar way than in our previous studies. Based on the distribution of their values, it was easy to identify at least 3 main groups of vocabulary elements. The Heavy-Tailed group is



(a) Heavy-Tailed

(b) Light-Tailed

Fig. 3. Different Vocabulary Elements Groups based on their Distribution

characterized by two or three values that are heavily used (more than 25% of time), followed by several values with a lower, but still significant usage between 5% and 25%. An example of this distribution can be seen in Figure 3a. The Language (of the resource and the metadata), Format, Resource Type and Context elements are part of the Heavy-Tailed group. The Light-Tailed group is characterized by a dominant value (usage of more than 70% of the time), followed by few (2 or 3) values that are used more rarely (from 2% to 10% of the time). An example of this distribution can be seen in Figure 3b. Structure, Intended User Role, Interactive Type and Status seem to be part of the Light-Tailed group. The third group, called Symmetrical group, contains all vocabulary elements that are based on a scale. In this group, the central value of the scale is by far the most common value (60% to 70%). Difficulty, Interactivity Level, Semantic Density and Aggregation Level are part of this group. Other vocabulary fields, such as Cost, Copyright, Classification Purpose, Contributor Role, etc. lie between these groups.

One possible explanation for these field distributions involves the objectivity of the information and the use of default values. The elements in the Heavy-Tailed group, contains information that could be easily obtained from the object or the context where it is used. For example, it is easy for the indexer to determine the

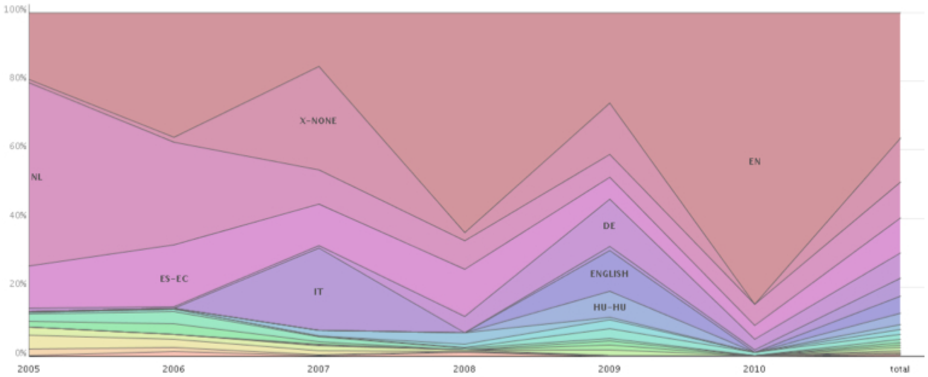


Fig. 4. Change over time of the values for general.language

language of the element, specifically because it is most probably in a language that he knows. On the other hand, data elements in the Light-Tailed group, seem to be more difficult to determine objectively (such as the Structure of the object) or there is safe default value for most objects (for example, it is safe to assume that the Intended End User is the learner). The existence of a safe default (usually the middle value) seems to be also a possible explanation for the value distribution in the elements in the Symmetrical group. Although there is anecdotal evidence to support this interpretation, further quantitative and qualitative studies of the indexation process are needed in order to provide a strong explanation.

4 LOM Evolution and Expansion

The previous section showed basic statistics about the GLOBE metadata. This section presents the use of LOM data elements through time to find existing trends in metadata indexation. To investigate this, only those metadata instances with value for $lifecycle.contribution$ date have been considered. "Many Eyes", a set of data visualization tools provided by IBM, has been used to create stack graphs for the values of the data elements which is meant for visualizing the total change over time of a group of quantities. Each ribbon of color in the stack graph represents a data item changing over time. Figure 4 shows the change over time of the values for *general.language*. This figure shows a number of trends:

In 2005, more than 50% of investigated metadata instances were in dutch. After 2005, a decrease is noticed by 40%. This raises the question why so many instances were indexed in dutch around 2005. One explanation could be the uptake of sharing educational resources by several institutions such as ARIADNE and Klascement in Flanders and the Netherlands around that time. Following this reasoning, organizations from Italy started to share their resources in 2007. English is the most used language for educational resources which is obvious.

Note however, that besides “en” also “english” is added as a value. Looking back to the data, “eng” was found for a minority of resources. Developers of applications that make use of this data set, should thus be careful to consider this interoperability issue. Another interesting trend is the use of the category relationship. There are several possible values for a relationship type provided by the standard such as “has worked on”, “is required by”, “requires”, “is version of”, etc. However, there are only two types in use that both take up about 50% of the instances: “is part of” and “is referenced by”. One could at least expect that indexers would index the relationship in both directions so that at least the types “has part” and “references” are used as much as the other types.

Application profiles (APs) are based on existing standards and specifications [12] but can add elements if needed for the purpose of the provider. One of our goals was to list elements used by the GLOBE providers that are not in the LOM standard but were added by extending the standard in one or more application profiles. The following 3 elements were found to be the most significant additions in the GLOBE metadata set:

– *general.learningobjectkind* is added in the MACE Application Profile to distinguish between a real world object (i.e. a building) or a media object (digital resource) that e.g. describes that real world object. As this is a mandatory field in the AP, 34.532 metadata instances were found in the investigated set that made use of this field.
– *technical.geolocation* is also added in the MACE AP to be able to add geo-coordinates in the case of a real world object such as a building. This field enabled the creation of typical mobile applications where users walk through a city and get extra information about the buildings in their neighborhood. However, a mismatch was found between the number of geo-locations (7 369) and the number of real world objects (11 622).
– *general.subtitle* is added by the AP of KOCW. It has a value in 73% of their metadata instances. In LOM, only 1 title is allowed with one or more translations of that title which is obviously not enough in the case of KOCW.

5 Metadata Quality Metrics

5.1 LOM Compliancy

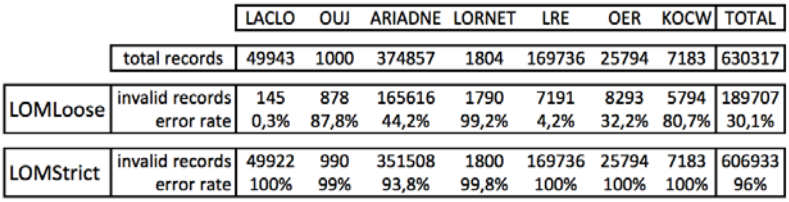
As the amount of metadata instances grows, so does the need for automatic validation against agreed rules of conformance, based on an application profile for example. An open source tool, the ARIADNE validation service, has been developed for this purpose. On the GLOBE metadata set, the conformance against the IEEE LOM loose and LOM strict bindings have been checked. LOM loose mostly focuses on validating the structure of the elements in the LOM namespace, checking for patterns of datatypes like date stamps, and the peculiarities of the vCard structure. The vCard standard is used to represent information about an individual, as defined by the IETF proposed standard RFC 2426[2].

[2] http://www.ietf.org/rfc/rfc2426.txt

LOM Strict has the same set of rules as LOM loose, with additional constraints on the used vocabularies. Furthermore, LOM strict does not allow to extend the structure with new elements. Figure 5 shows a summary of the validation error rate in total and split down per provider. The overall error rate against LOM Loose is 30,1%, which means that 30% of the records can cause issues when processing these with tools that assume LOM compliancy. Looking at the error rate per provider, it can be seen that OUJ, LORNET and KOCW have an error rate of over 80%. Also striking is the error rate of the LOM strict validation. Only 4% of all records are fully LOM Strict compliant. These numbers are explored in more detail further down this section. The validation errors can be split up in the following categories:

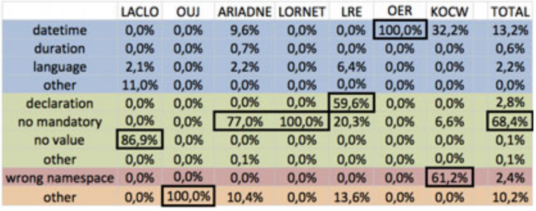
- Pattern matching errors : Certain datatypes have very specific syntax constraints. These are checked with pattern matching rules. All kinds of pattern matching errors are put in this category. An example mistake in this category is a wrongly denoted timezone in a date stamp. For example, "GMT-5" should be written as "-05:00".
- vCard errors : This category contains all errors concerning the vCard standard. This datatype is used in the lifecycle.contribute.entity and metaMetadata.contribute.entity elements.
- LOM extension errors : When extending LOM in an application profile, elements can be added using a separate namespace and different vocabularies can be used. However they can not be used inside the same LOM namespace, otherwise they are seen as a validation error. Furthermore, as said earlier, LOM strict imposes a fixed set of allowed values for all elements with a vocabulary datatype. Thus, all other values are considered invalid when validating against LOM strict.
- Various errors : This contains all other types of errors. Mostly these are structural errors, e.g. putting a child element in a parent element must only contain plain text.

A more in depth look at the validation errors against LOM loose is shown in Figure 6. The most prominent error is a vCard error, where the vCard is missing 1 of 3 mandatory elements (i.e. N, FN or VERSION). It represents 68% of all LOM loose validation errors. vCard errors in general are the most common



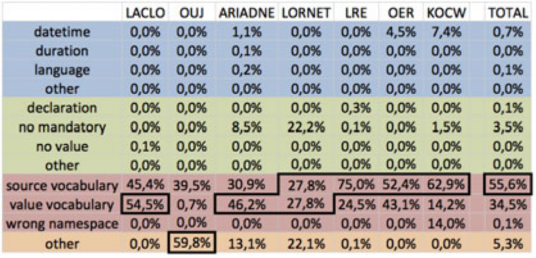
		LACLO	OUJ	ARIADNE	LORNET	LRE	OER	KOCW	TOTAL
	total records	49943	1000	374857	1804	169736	25794	7183	630317
LOMLoose	invalid records	145	878	165616	1790	7191	8293	5794	189707
	error rate	0,3%	87,8%	44,2%	99,2%	4,2%	32,2%	80,7%	30,1%
LOMStrict	invalid records	49922	990	351508	1800	169736	25794	7183	606933
	error rate	100%	99%	93,8%	99,8%	100%	100%	100%	96%

Fig. 5. A summary of error rates in GLOBE against LOM loose and LOM strict



	LACLO	OUJ	ARIADNE	LORNET	LRE	OER	KOCW	TOTAL
datetime	0,0%	0,0%	9,6%	0,0%	0,0%	100,0%	32,2%	13,2%
duration	0,0%	0,0%	0,7%	0,0%	0,0%	0,0%	0,0%	0,6%
language	2,1%	0,0%	2,2%	0,0%	6,4%	0,0%	0,0%	2,2%
other	11,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
declaration	0,0%	0,0%	0,0%	0,0%	59,6%	0,0%	0,0%	2,8%
no mandatory	0,0%	0,0%	77,0%	100,0%	20,3%	0,0%	6,6%	68,4%
no value	86,9%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,1%
other	0,0%	0,0%	0,1%	0,0%	0,0%	0,0%	0,0%	0,1%
wrong namespace	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	61,2%	2,4%
other	0,0%	100,0%	10,4%	0,0%	13,6%	0,0%	0,0%	10,2%

Fig. 6. Validation results against LOM loose. Blue represents pattern matching errors, green, vCard errors, red, LOM extension errors and orange, other errors.



	LACLO	OUJ	ARIADNE	LORNET	LRE	OER	KOCW	TOTAL
datetime	0,0%	0,0%	1,1%	0,0%	0,0%	4,5%	7,4%	0,7%
duration	0,0%	0,0%	0,1%	0,0%	0,0%	0,0%	0,0%	0,0%
language	0,0%	0,0%	0,2%	0,0%	0,0%	0,0%	0,0%	0,1%
other	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
declaration	0,0%	0,0%	0,0%	0,0%	0,3%	0,0%	0,0%	0,1%
no mandatory	0,0%	0,0%	8,5%	22,2%	0,1%	0,0%	1,5%	3,5%
no value	0,1%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
other	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
source vocabulary	45,4%	39,5%	30,9%	27,8%	75,0%	52,4%	62,9%	55,6%
value vocabulary	54,5%	0,7%	46,2%	27,8%	24,5%	43,1%	14,2%	34,5%
wrong namespace	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	14,0%	0,1%
other	0,0%	59,8%	13,1%	22,1%	0,1%	0,0%	0,0%	5,3%

Fig. 7. Validation results against LOM strict. Blue represents pattern matching errors, green, vCard errors, red, LOM extension errors and orange, other errors.

errors, spread across almost all providers. In 4 out of 7 providers, vCard errors represent at least 60% of all errors. Only 2 providers have no vCard errors at all. Further analysis has shown that 79% of the invalid records contain at least one vCard error. Solving only this vCard error would turn at least 68% of the invalid records into valid records. This indicates that vCards have been wrongly used across all providers. When analyzing the error type of the providers with an error rate of over 80%, it was found that 100% of the errors of LORNET are about a missing mandatory element in a vCard. Looking at OUJ's errors reveals that all of these are structural errors, by putting a "string" or a "entry" element inside an invalid parent element. Finally, validation errors of KOCW are largely due to extending LOM in the wrong way, by adding a new element without using a new namespace for this element. When examining the validation errors against LOM strict, as shown in Figure 7, a completely different picture can be seen. Here, 90% (55,6% source vocabulary and 34,5% value vocabulary errors) of the validation errors are caused by the "LOM extension" category. Hence it is also no surprise that for 6 of the 7 providers, this same category contains more than 50% of the errors. By looking at a number of randomly hand-picked errors, it becomes clear that the larger the amount of invalid records for one error, the bigger the chance that the error is more syntactical in nature. A possible explanation could be that these errors are produced by (semi-)automatically generated metadata or by programming bugs in client tools. To conclude this section, we can safely say that the vCard standard is very error prone, and not very portable. One of the main reasons is its syntax. It relies on new line characters, it is heading space sensitive, etc. There are also many DateTime formatting issues. One cause is an inconsistency in the XML binding of LOM and the ISO 8601:2000 standard it claims to adhere to. When a TZD (TimeZone Designator) is present, the digits representing a decimal fraction of a second become mandatory in LOM, whereas this should actually not be the case. Finally, a major problem with metadata validation is that providers are usually not aware of these errors, and often they do not use validators to check their own compliancy. People are also not aware of how to properly extend LOM, and they do not understand properly how xml namespaces work.

5.2 Information Content

One of the main purposes of any metadata schema is to carry information about the object that is described. One method that has been extensively used to determine the informativeness of a metadata instance is to measure its completeness; the presence or absence of a value for the different data elements of the metadata schema. However, the mere existence of a value in 1 of the data elements does not necessarily mean that this value is carrying information. For example, if all metadata instances in the repository have the same title, that data is useless to find and select a particular object within that repository. On the other hand, if values for a data element are well distributed among the metadata instances, the information carried by that element is maximized and that field could be easily used to filter and select elements from the repository. Shannon was the first to propose variety or disorder as a measurement of information content [13]. He proposed "entropy" as a measurement of information contained in a message. This subsection will apply different measurements to measure entropy of values contained in various types of data elements of LOM in order to determine the amount of information they carry.

Vocabulary Data Elements. When a variable can take only a determined set of values, the information content of that variable can be determined by its entropy. The entropy is equal to the sum of the probability of all given values, multiplied by the negative of the logarithm of the value probability. If the logarithm used is base-2, entropy is measured in bits. This concept is applied to determine the entropy of the vocabulary data elements [14]. For each of the studied repositories, entropy values have been calculated for the 14 vocabulary data elements presented in Table 2. If the repository does not have any value for that element it is marked with an X. This entropy value depends on the variety of the distribution of values in the data elements. For example, in LACLO, all instances have "es-EC" as the value for "Language". The calculated entropy for this case is 0 bits, as there is no variety in the data. On the other hand, LRE has materials from various european countries which makes its content highly multi-lingual. As expected, entropy value for "Language" in LRE is the highest in GLOBE (3,43 bits). The entropy value also depends on the number of choices for each element. For example, the maximum entropy for the element Cost, that could only take two values (Yes or No) is 1 bit, while the maximum entropy for the element Difficulty that could take five values is 2,32 bits. Because of this, the value of entropy is only comparable between elements with the same number of possible values.

It can be easily deduced from Table 2 that LRE has the most diverse information in their vocabulary elements. One explanation for this observation is the process of metadata enrichment that has been performed over most metadata [15] in that repository. On the other hand, LACLO instances carry very little information because all the values for an element are the same. The lack of sophistication of the automatic algorithm used to generate the metadata of this repository (for example adding the Aggregation Level atomic to all its instances) is responsible for these results. However, LACLO has the highest entropy value for "Format", which indicates that the automatic metadata generator seems to be better equipped than manual indexers to obtain the MIME type of digital objects. The main conclusion that can be obtained from this analysis is that quality control processes, such as the enrichment conducted in the LRE repository, helps to improve the amount of information contained in the metadata instance.

Free Text Data Elements. When data elements can be filled with free text, traditional entropy measurement is not possible as the amount of values that the text could take is infinite. However, the entropy formula could be adapted to count the information carried by each word in the free text elements, divided by the number of words used. In this way, a value of entropy can be calculated for each instances based on the text that they contain. The Relative Entropy metric provides such adaptation [16]. Figure 8 presents the distribution of the Relative Entropy metric for each repository. In general, all repositories have a main peak between 6 and 12 bits per word. Repositories that are the aggregation of other repositories present one or more lower peaks at different values (e.g. at 6 and 17 for LRE). While in this study, the enriched LRE again contains more information per word in their text fields, the automatically generated instances from LACLO, that took their free text from the text available in the context where the object was published, seems to contain a similar amount of information than other repositories with manually generated metadata such as LORNET, OUJ and OER.

Table 2. Entropy value for the Vocabulary Data Elements in each Repository

	ARIADNE	LRE	LACLO	OER	KOCW	LORNET	OUJ
Language	2,22	3,43	0	0,52	1,38	1,10	0,70
Structure	1,65	0,60	X	X	X	1,39	X
Aggregation Level	1,09	1,85	0	X	X	1,30	0,21
Status	0,99	0	X	X	X	1,09	X
Format	2,72	X	3,04	2,10	0,07	2,70	2,27
Interactivity Type	0,71	1,55	X	X	X	1,02	X
Learning Resource Type	2,87	3,04	1,76	3,61	0	1,05	X
Interactivity Level	1,18	2,02	X	X	X	1,49	X
Semantic Density	1,51	X	X	X	X	0	X
Intended End-User Role	1,56	1,91	X	X	0	2,42	X
Context	2,62	2,23	X	1,43	0	2,55	0,14
Difficulty	1,24	1,71	X	X	X	0	X
Cost	0,12	0,01	0	0	0	0,69	0
Copyright	0,82	0,55	0	0,30	0	0,94	X

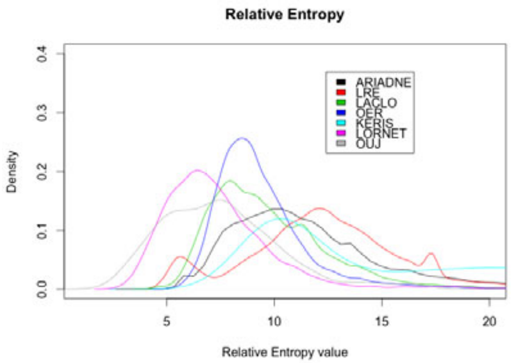


Fig. 8. Distribution of the Relative Entropy (RE) metric

6 Conclusions and Future Work

The main conclusions of this work can be summarized in the following points:

- Out of 50 LOM elements, 20 are used consistently in GLOBE. This is in itself more metadata than traditionally collected metadata in simpler schemas such as DC. Results of this analysis can be used to improve search facilities, focusing on data that is actually contained in the metadata.
- 4 out of 11 educational elements are used in average. However, their actual use is repository/community dependent. The LOM Working group and repository managers should analyze the reasons behind different levels of adoption.
- 3 main extended data elements have been found. These can be a starting point for a LOM Working group discussion about the merit of inclusion of new elements in the next version of the standard. A more exhaustive analysis of Application Profiles should be performed to get a broader understanding of the need for additional elements.
- Solving the vCard related errors would turn at least 68% of the LOM loose invalid records into valid records. Also, in GLOBE it does not make sense to adhere to the LOM strict binding, as only 4% of all records in GLOBE are fully compliant with LOM strict.
- The information content of a record is strongly related with both the quality management process implemented by the repositories and the inherent capabilities of different types of metadata generation.

This work also rises new questions that require further research in order to be answered:

- What is the mechanism behind the usage level of different metadata elements in a given community?
- The distribution of values in the fields follows patterns. What are the forces that shape those distributions?

It is the perception of the authors of this work that this kind of studies should be made an integral part of the development of LOM (or any other metadata schema) solution. Only by analyzing actual use of different schemas in real life settings, decisions can be taken to improve or adapt those schemas so that teachers and learners can really benefit from them.

References

1. Duval, E., Forte, E., Cardinaels, K., Verhoeven, B., Van Durm, R., Hendrikx, K., Forte, M., Ebel, N., Macowicz, M., Warkentyne, K., et al.: The Ariadne knowledge pool system. Communications of the ACM 44(5), 72–78 (2001)
2. IMS-GLC: IMS Learning Resource Metadata specification v.1.2 (2001), `http://www.imsglobal.org/metadata/`
3. IEEE: IEEE 1484.12.1 Standard: Learning Object Metadata (2002), `http://ltsc.ieee.org/wg12/par1484-12-1.html`
4. Neven, F., Duval, E.: Reusable learning objects: a survey of LOM-based repositories. In: Proceedings of the tenth ACM international conference on Multimedia, MULTIMEDIA 2002, pp. 291–294. ACM, New York (2002)
5. Najjar, J., Ternier, S., Duval, E.: The actual use of metadata in ARIADNE: an empirical analysis. In: Duval, E. (ed.) Proceedings of the 3rd Annual ARIADNE Conference, pp. 1–6. ARIADNE Foundation (2003)
6. Friesen, N.: The international learning object metadata survey. The International Review of Research in Open and Distance Learning 5(3) (2004)
7. Van de Sompel, H., Nelson, M., Lagoze, C., Warner, S.: Resource Harvesting within the OAI-PMH Framework. D-Lib Magazine 10(12), 1082–9873 (2004)
8. Meire, M., Ochoa, X., Duval, E.: Samgi: Automatic metadata generation v2.0. In: Seale, C.M.J. (ed.) Proceedings of the ED-MEDIA 2007 World Conference on Educational Multimedia, Hypermedia and Telecommunications, pp. 1195–1204. AACE, Chesapeake (2007)
9. Downey, A.B.: The structural cause of file size distributions. In: Proceedings of IEEE Ninth International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, pp. 361–370 (2001)
10. Ward, J.: Unqualified Dublin Core usage in OAI-PMH data providers. OCLC Systems & Services 20(1), 40–47 (2004)
11. Weibel, S.: The Dublin Core: a simple content description model for electronic resources. Bulletin of the American Society for Information Science and Technology 24(1), 9–11 (1997)
12. Heery, R., Patel, M.: Application profiles: mixing and matching metadata schemas. Ariadne 25, 27–31 (2000)
13. Shannon, C.E.: A mathematical theory of communication. ACM SIGMOBILE Mobile Computing and Communications Review 5(1), 3–55 (2001)
14. Ochoa, X., Duval, E.: Automatic evaluation of metadata quality in digital libraries. International Journal of Digital Libraries 10(2), 67–91 (2009)
15. Massart, D.: Towards a pan-European learning resource exchange infrastructure. In: Next Generation Information Technologies and Systems, pp. 121–132 (2009)
16. Stvilia, B.: Measuring information quality. PhD thesis, University of Illinois (2006)