

## E-Learning and Microformats: A Learning Object Harvesting Model and a Sample Application

Ahmet Soylu<sup>1</sup>, Selahattin Kuru<sup>2</sup>, Fridolin Wild<sup>3</sup>, Felix Mödritscher<sup>3</sup>,

<sup>1</sup> K. U. Leuven, Interdisciplinary Research on Technology, Education and Communication (iTec), Campus Kortrijk, Etienne Sabbelaan 53, 8500 Kortrijk, Belgium  
ahmet.soylu@kuleuven-kortrijk.be

<sup>2</sup> Işık University, Department of Computer Science and Engineering, Kumbaba Mevkii Şile, 34980, Istanbul, Turkey  
kuru@isikun.edu.tr

<sup>3</sup> Vienna University of Economics and Business Administration, Department of Information Systems, Augasse 2-6, 1090 Wien, Austria  
{fridolin.wild,felix.moedritscher}@wu-wien.ac.at

**Abstract.** In order to support interoperability of learning tools and reusability of resources, this paper introduces a framework for harvesting learning objects from web-based content. Therefore, commonly-known web technologies are examined with respect to their suitability for harvesting embedded meta-data. Then, a lightweight application profile and a microformat for learning objects are proposed based on well-known learning object metadata standards. Additionally, we describe a web service which utilizes XSL transformation (GRDDL) to extract learning objects from different web pages, and provide a SQI target as a retrieval facility using a more complex query language called SPARQL. Finally, we outline the applicability of our framework on the basis of a search client employing the new SQI service for searching and retrieving learning objects.

**Keywords:** E-Learning, Learning Object, Metadata, Semantic Web, Interoperability, SQI, RDF, SPARQL, GRDDL, Microformats.

### 1 Motivation

Learners are not bound neither to individual learning environments as closed box of pure information nor to classical in-class learning environments anymore. Instead they are facing with several tools (web 2.0) within their learning environment, which enable them to collaborate, to reach an endless amount of information on the web, to remix and share it, and to create social networks. The e-learning market has already been overpopulated with tools and platforms to support different types of learning communities with learning management, content management, and communication facilities [1]. Depending on case, these tools are being used individually by learners, or by means of mash-ups, or as heterogeneous systems which involve several tools

and might be centered around one particular learning system. All these aspects end up in several technological challenges. Our work specially targets the following ones;

- Re-use: Learning content can be easily split up, re-used, and arranged again in order to create different contexts.
- Interoperability: Multiple learning systems or tools can communicate with each other to share and mix their learning resources.
- Accessibility: Learning content can be accessed by several learning tools and devices, besides appropriateness of content can be ensured.
- Durability: Users are not bound to any specific system, tool or device, interoperability and reusability is ensured.

We consider this work to be particularly important for mash-ups because the shift towards learner centric approaches accompanied by an ever-increasing degree of interaction and rising volumina of content provided by learners through the use of diverse social and web 2.0 tools emphasizes the importance of collective and personalized use of these distributed applications. Therefore, our work is considered as an approach to address the re-usability, interoperability, accessibility, and durability challenges of mash-ups by making existence, retrieval, sharing, and harvesting of learning objects independent of their applications, where these challenges were neglected due to the limiting features provided (RSS, import/export facilities, etc.).

## 2 Semantic Web

Computers are not able to automatically apply actions like ‘derive’, ‘associate’, or ‘link’ onto information available through the web – in the way humans do. Instead, they in general deal only with presentation and routine processing, as the web does most often does not provide the mark-up necessary for automatically investigating and understand the meaning of the data provided. For example, collections of documents on the web are linked to each other without allowing inspecting the motivation why these resources have been linked. The semantic web offers a variety of new opportunities to overcome these shortcomings in today’s hypertexts. Amongst others, the semantic web aims at enriching the content of the world wide web pages with meaningful structure to create an environment in which software agents can roam from page to page to carry out more intelligent tasks for their users [2]. Thereby, the semantic web integrates (earlier) research from knowledge representation, however, in the same decentralized manner the web is known for.

XML and RDF technologies constitute a basis for knowledge representation for expressing web-based semantics. An important point is that XML just describes the structure of data and does not say anything about its meaning, while RDF is used to formalize semantics. However, RDF is not enough to describe complex relationships between objects such as cardinality constraints, unions, disjoint classes etc. RDFS gives a limited capability to define only simple relationships. At this point ontologies come up as a solution, they are used to define complex relationships and sets of inference rules between data objects by providing more vocabulary. There are also many efforts for an ontology-based semantic web, important ones to mention here are

the DAML and OWL ontology languages. Both of these languages are based on XML, RDF, and RDFS. XML, RDF, and OWL can be seen as the layers where each layer requires different skills and targets different needs [3].

XML, RDF, OWL, and all of these technologies aim at expressing semantic in a machine understandable way, but not to be understood by humans. It would be a big burden to duplicate efforts by creating a piece of data in the form of both RDF, for machines, and simple XHTML, for humans. Additionally, this would also be problematic in terms of data consistency and synchronization. To avoid this double work burden, eRDF, RDFa, and microformats have been proposed, allowing for embedded semantic mark-up within web pages and being understandable for both machines and humans. RDFa is a W3C specification based on expressing structure via attributes of languages of XHTML and HTML [4]. eRDF, i.e. embedded RDF, facilitates the embedding of crucial parts of the RDF model, but does not attempt to extend this to the full RDF support [5]. eRDF is inspired by some basic principles of microformats. A microformat is a set of XHTML tags that is used to embed information into web pages which are understandable both by machines and humans while considering the human as first priority [6].

RDFa and eRDF are based on the framework provided by RDF whereas microformats offer both syntax and a restricted vocabulary that does not rely on RDF or any other framework. Therefore, microformats are domain specific. A microformat only becomes meaningful, when its syntax and vocabulary are defined by a community. This also implies that the extracting procedure is same for every eRDF and RDFa statement expressed in XHTML pages whereas it is different for every microformat. eRDF and RDFa are based on RDF. They therefore enable users to mix and use different name spaces. Microformat, however, use a flat name space that is already predefined and cannot be extended or mixed; new metadata elements cannot be added. It is obvious that RDFa and eRDF provide more flexibility and functionality than microformats, however there are hardly any real life applications of RDFa or eRDF whereas there are many real life examples of microformats. Microformats do not aim at becoming a panacea for expressing taxonomies, ontologies, or other abstractions which are not extensible or open ended [7].

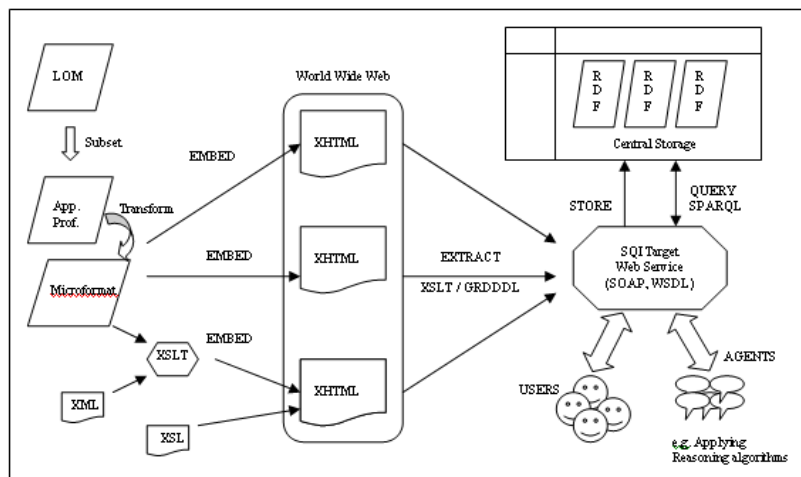
While Microformats can encode explicit information to aid machine readability, they do not address implicit knowledge representation, ontological analysis, or logical inference [8]. The simplicity of this approach may be the reason of its success. Its constraints, however, are the reason why it is called 'lower case' semantic web. A living community being the key, it is not astonishing that most microformats express well-established and accepted meta-data such normalized by Dublin Core, iCal, vCard, FOAF, or ATOM. For instance, using the hAtom microformat, a website can immediately provide a feed by intelligently marking up items. There is no longer a need to publish the same content in other formats like RSS since applications can extract raw ATOM data from the (X)HTML of the website, again without converting to and from RDF [9].

Microformats are well accepted by now. Within the known limitations, they provide already simple and straightforward capabilities to manipulate meaningful structures added to human-understandable pages.

### 3 A Learning Object Harvesting Model for Web Content

The proposed model for harvesting learning objects bases assumes that learning objects (also called ‘learning resources’) are digital and non-digital artifacts (images, texts, etc.) that can be used for educational purposes. Learning objects can be part of the courseware ranging in size and complexity from a single graphic to an entire course [10]. Enabling learning objects to be re-usable and sharable by different tools, systems, and applications is the core challenge for e-learning interoperability.

Hereby, meta-data plays a key role: there are several bodies which have published standards and specifications to describe metadata information. In the context of this work, LOM (Learning Object Metadata) standardized by IEEE is of prominent interest, though other standards are available and compatible with the approach chosen. Most of these standards consist of many elements with varying degrees of relevance (depending mainly on the application and the usage context). For instance, LOM currently offers more than 71 elements. From this standardized set offered by LOM, an application profile has been derived mixing in elements from DC and additional custom elements. The application profile forms the basis for a microformat to embed learning resources into web pages that can later-on be harvested by crawlers. Embedding microformat structures into XHTML can be realized by hand with simple XHTML mark-up or it can be automated with the help of XSLT transformations over the XML bindings of LOM.



**Fig. 1.** Harvesting model for extracting learning object meta-data from web-based content.

After proposing a light-weight microformat for learning objects, a web service (based on SOAP and WSDL) has been created which harvests learning objects (bound to microformat) in XHTML pages and transforms them into RDF format via XSLT or XSLT and GRDDL [11] combined. If the XHTML file references an XSL file providing transformation information on how to translate microformat bindings of

learning objects into RDF then GRDDL is used. Otherwise a predefined XSL transformation is applied to perform the necessary transformation into RDF.

Additionally, an SQI target is provided which uses above mentioned harvesting mechanism and allows querying for learning objects meta-data harvested into the central storage facility. SQI is an interoperability infrastructure that enables heterogeneous systems to communicate for the purpose of learning object retrieval by using a common query language (in our case SPARQL) [12]. Fig. 1 provides an overview on the workflow of the proposed harvesting and retrieval model. As a proof of concept, a search application was implemented which uses the SQI target to query learning resource meta-data.

#### **4 Implementation Work and Experiences**

Simplicity and a minimalist approach have been considered as the guiding mantra of the prototypical implementation described below, especially while selecting data types and meta-data elements. An e-Learning application profile guideline, which was submitted to 2006 CEN/ISSS WS/LT workshop [13], has been followed to guide the construction of the application profile (roughly consists of identifying the following steps: requirement analysis, selection of data elements, multiplicity requirements, data elements from multiple name spaces, local data elements, obligation of data elements, value spaces, relationship and dependencies, and data type profiling). The primary purpose of the proposed application profile is to facilitate re-usability and interoperability of learning objects both by individuals and automated software applications such as agents. The minimalist approach defines the borders of this application profile by the vocabulary of our microformat. Therefore, it is important to address common needs instead of providing a comprehensive profile, since schemas of microformats are not subject to change often. Elements of the application profile have been derived from IEEE LOM 1482-12-1 2002 and Dublin Core, enriched by two custom elements. The application profile's reference model and XML bindings are based on SCORM 2004 3rd Edition.

The selection follows and extends the proposals for application profiles done by Friesen and Campbell [14, 15] and later work of Godby [16] (which again is based on Friesen and Campbell's spreadsheets). A total of 35 application profile has been investigated and lead to the conclusion that there is a close match between Dublin Core elements and the frequently used IEEE LOM elements. Therefore we followed the idea to map the 15 Dublin Core elements into the IEEE LOM elements if available, which means we included only LOM elements having a respective Dublin Core match, and if there is no such match, the Dublin Core element is taken as is. A total of 18 elements have been proposed, seven of them adopted from the Dublin Core name space, nine of them adopted from the IEEE LOM name space, and two additional ones that have been derived from IEEE LOM through a custom name space, this is done just in order to simplify those two aggregate elements. Tab. 1 provides an overview on the selected elements. Further details such as data types, multiplicity requirements etc. have been omitted due to space limitations.

**Tab. 1.** Elements of the proposed microformat (LO: LOM, DC: Dublin Core, CO: Custom)

| Set | Selected Elements  |
|-----|--|
| LOM | Title, Language, Description, Keyword, Coverage, Type, End User, Context, Format |
| DC  | Date, Creator, Contributor, Publisher, Rights, Relation, Source                  |
| CO  | Identifier, Classification   |

The proposal of microformat was quite straight forward in this case, because the base applications profile itself is already simple and flat as there is no aggregate element used in application profile at all. Most of the elements has been directly associated with a ‘class’, i.e. class design pattern, attribute. The ‘creator’, ‘contributor’ and ‘publisher’ elements are used together with ‘vcard’ identifier. This is because the domains associated with these elements are based on ‘vcard’. The abbr-design pattern has been used for the ‘date’ and ‘language’ elements. The only elemental microformat, i.e. the ‘rel-tag’ microformat, is utilized for the ‘keyword’ element. Fig. 2 depicts the extraction of the proposed microformat.

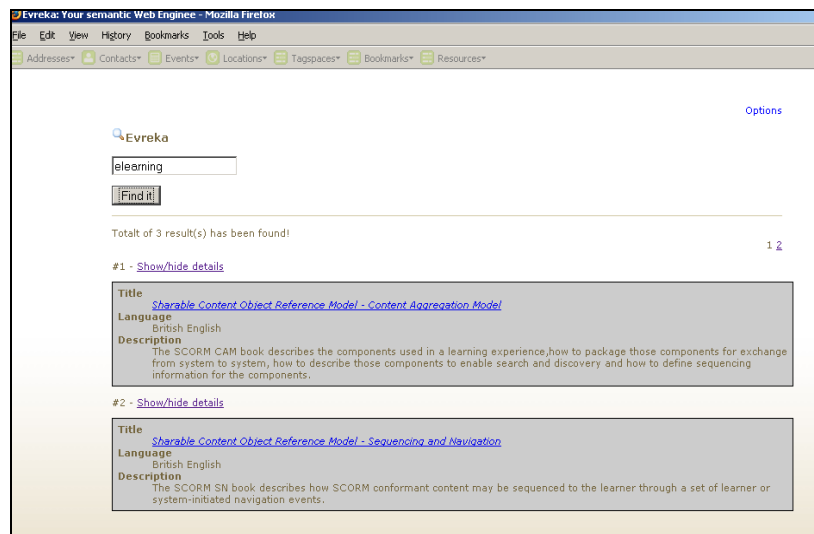
|                        |  |
|------------------------|--|
| <b>Title</b>           | Sharable Content Object Reference Model - Sequencing and Navigation  |
| <b>Language</b>        | British English  |
| <b>Description</b>     | The SCORM SN book describes how SCORM conformant content may be sequenced to the learner through a set of learner or system-initiated navigation events. |
| <b>Creators</b>        | Daren  |
| <b>Contributors</b>    | Alice  |
| <b>Publisher</b>       | Advanced Distributed Learning (ADL)  |
| <b>Identifier</b>      | <a href="http://course.isikun.edu.tr/LO2">http://course.isikun.edu.tr/LO2</a>  |
| <b>Classification:</b> | ADL Scorm Cocepts  |
| <b>End User:</b>       | Learner  |
| <b>Context:</b>        | School   |
| <b>Coverage</b>        | eLearning  |
| <b>Keywords</b>        | <a href="#">eLearning</a> , <a href="#">Learning Object</a>  |
| <b>Type</b>            | narrative text   |
| <b>Date</b>            | 30 July 2007   |
| <b>Format</b>          | txt/html   |
| <b>Rights</b>          | <a href="http://www.isikun.edu.tr/termsandconditions.htm">http://www.isikun.edu.tr/termsandconditions.htm</a>  |
| <b>Relation</b>        | Part of Scorm 3rd Edition, Content Aggregation Model   |
| <b>Source</b>          | Scorm 3rd Edition  |

**Fig. 2.** Extracted meta-.data of an exemplary learning object encoded with our microformat.

An SQI target has been built and enhanced with GRDDL and SPARQL by using the RAP RDF API [17]. Several sample microformat-enriched web-pages have been generated which are available to the SQI target. Then, a simple search client which is based on AJAX, JavaScript, CSS, and PHP has been set up to employ this SQI target, after its harvesting the sample pages via GRDDL, to query microformat-encoded semantics via SPARQL. These web-pages has been generated for test purposes, however any individual or authoring tool supporting our proposed microformat is capable of producing such microformat embedded pages by means of individual authoring or by means of automated transformation of current content as far as a valid

match found between their previous (or base) metadata descriptions and our microformat (this is usually the case if DC or LOM has been employed as base).

For instance the metadata for our sample learning objects are created with respect to the XML LOM bindings. Thereafter, a simple XSL transformation embeds this semantics into the web-pages according to our microformat automatically, and our SQI target harvests this semantic information using GRDDL and transforms it into RDF. Then, the query results are sent to search client in the form of a LOM XML binding, and the search client visualizes the extracted XML via XSL. In Fig. 3 the search results for the query term ‘elearning’ are displayed.



**Fig. 3.** Results for term ‘elearning’ visualized with our search client employing the SQI target.

It is important to highlight that the search client serves as a proof of concept, particularly for SPARQL, i.e. the SQI target only employs SPARQL’s basic facilities. Once the learning objects are represented with RDF or even empowered ontologically, query languages for RDF like SPARQL will allow the application of complex relationship queries and operations over learning objects in the sense of retrieval, analysis, content construction etc.

Comparing our microformat-based to an ontology-based approach, it has to be mentioned that microformats have clear shortcomings against RDF. Particularly, microformats lack strengths like ‘open-ended design’, ‘the ontological expression power’, and ‘extensibility’. However, a switch from Microformats to RDFa is in principle possible, though at the time of writing not promising due to the widespread acceptance and due to the ease-of-use of microformats. The drawbacks of microformats constitute the main drawbacks of the model and the sample application.

## 5 Conclusion and the Future Work

Within this paper, we have introduced a model and a prototypical sample application for learning object harvesting with the help of a new microformat in order to address mainly interoperability and reusability of learning content. Apart from the drawbacks which originate from the simplicity of microformats, this model needs to be complemented by strong reasoning and data mining algorithms, and, additionally, by techniques not just for harvesting objects but also for analyzing learning patterns and learning activities described on the web. Efficient agents for crawling need to be added as well. Those agents are not expected to serve just traditional desktop computers, but should rather serve any device which has access to internet such as PDAs or cell phones. In this way, learning could be embedded into real life.

## References

1. Kieslinger, B., Wild, F., Arsun, O.I.: iCamp – The Educational Web for Higher Education. In: Proc. of the EC-TEL, pp. 640--645 (2006)
2. Berners-Lee, T., Hendler, J., Lassila O.: The Semantic Web. Scientific American (2001)
3. Wilson, M., Matthews, B.: The Semantic Web: Prospects and Challenges. In: Proc. Of the International Conference on Databases and Information Systems, pp. 26--29 (2006)
4. Adida, B., Birbeck, M.: RDFa Primer. W3C (2008). Available at <http://www.w3.org/TR/xhtml-rdfa-primer/#id85078> (Aug. 2008)
5. Talis: Rdf in HTML: Embedded RDF (2006). Available at <http://research.talis.com/2005/erdf/wiki/Main/RdfInHtml> (Aug. 2008)
6. Allsopp, J.: Microformats: Empowering Your Markup for Web 2.0. FriendsofED, Berkeley (2007)
7. Simpson, J.: Microformats vs. RDF: How Microformats relate to the Semantic Web (2007). Available at <http://www.semanticfocus.com/blog/entry/title/microformats-vs-rdf-how-microformats-relate-to-the-semantic-web> (Aug 2008)
8. Khare, R., Çelik, T.: Microformats: A pragmatic path to the Semantic Web. In: Proc. of the International World Wide Web Conference, pp. 865--866 (2006)
9. Graf, A.: RDFa vs. Microformats. DERI Technical Report (2007)
10. Collier, G.: e-Learning Application Infrastructure. *Whitepaper*. Sun Microsystems (2002)
11. Connolly D.: Gleaning Resource Descriptions from Dialects of Languages. W3C (2004). Available at <http://www.w3.org/2004/01/rdxh/spec> (Aug. 2008)
12. Prud'hommeaux E., Seaborne E.: SPARQL Query Language for RDF (2008). Available at <http://www.w3.org/TR/rdf-sparql-query/> (Jan. 2008)
13. CWA: Guidelines and support for building application profiles in e-learning (CEN Workshop Agreement 15555:2006E), European Committee for Standardization (2006)
14. Friesen, N.: Survey of LOM Implementations, CanCore (2003)
15. Campbell, M.L.: UK LOM Core Update. CETIS Technical Report (2003)
16. Godby, C. J.: What do Application Profiles Reveal about the Learning Object Metadata Standard? ARIADNE, October 2004 (41)
17. Melnik, S.: RAP: RDF API for PHP. SourceForge (2008). Available at <http://sourceforge.net/projects/rdfapi-php> (Aug. 2008)