HTML5, MICRODATA AND SCHEMA.ORG
Towards an Educational Social-semantic Web for the Rest of Us?

Lars Johnsen
Institute of Business Communication and Information Science, University of Southern Denmark
Engstien 1, 6000 Kolding, Denmark

Keywords: Metadata, HTML5, Microdata, Schema.org, The Educational Social-semantic Web.

Abstract: The aim of this short position paper is to argue that the combination of the HTML5 platform, embedded microdata and schema.org vocabularies may pave the way for an educational social-semantic web, a network of structured learning content and supporting compliant software to which "the rest of us" may contribute.

1 THE EDUCATIONAL SEMANTIC WEB

For some years now the vision of the educational semantic web has been a popular topic in some parts of the e-learning community. The educational semantic web may be described as a subset of the semantic web for educational purposes comprising a web of structured data, rather than hyperlinked documents, and services exposing these data to various kinds of software such as search engines, virtual agents, etc. The idea of a semantic web of learning content and supporting software is attractive to many educational technologists because of its potential benefits: improved discoverability of learning resources, re-use of learning objects, more sophisticated content adaptation based on user models and user actions, seamless data integration across disparate platforms and systems, etc.

Until now, work in the field has mainly focused on two areas:

- The application of semantic technologies like RDF, Topic Maps and Linked Data to e-learning resources, tools and systems
- The design, construction and publication of learning metadata models, taxonomies and ontologies such as LOM, SCORM or ALOCOM

Although much useful work has been done and considerable progress made, the vision of the educational semantic web is still to a great extent confined to academic environments. To the average teacher or professor the idea of semantically encoding his or her teaching materials and exposing them via a web service to compliant software is still somewhat alien. There are obviously many reasons for this, but one key factor is, I believe, the conceptual and technical barriers inherent in current semantic web models, technologies and tools. This in turn means that a critical mass of semantically enriched learning content on the web, let alone an educational semantic web, is nowhere near its realization.

The aim of this short position paper is to argue that the introduction of HTML5, microdata, a standard for marking up structured data in HTML5 documents, and the launch of schema.org, a set of general purpose vocabularies for semantic annotation of web content, may well be just the thing needed to kick-start developments towards an educational semantic web for "the rest of us" – a kind of social-semantic web to which the ordinary teacher or professor can contribute.

2 HTML5, MICRODATA AND SCHEMA.ORG

HTML5 will no doubt be the presentation format of most web-based learning resources in the years to come. This latest version of HTML has improved functionality to embed multimedia objects like video and audio files and natively supports various types of interactivity, for example drag-and-drop actions.

Johnsen L..
HTML5, MICRODATA AND SCHEMA.ORG - Towards an Educational Social-semantic Web for the Rest of Us?.
DOI: 10.5220/0003895901010104
In *Proceedings of the 4th International Conference on Computer Supported Education* (CSEDU-2012), pages 101-104
ISBN: 978-989-8565-06-8

Likewise, it is likely to play a major role in m-learning as it can be employed for the creation of mobile web apps through the application of Javascript libraries like jQuery Mobile. As the format becomes accepted as *the* standard for user-oriented web publishing, it will be underpinned by an increasing number of development tools ranging from simple text editors, many accessible online, open source plug-ins facilitating design activities such as canvas drawing to full-fledged development environments aimed at widget construction or the like.

To make possible simple forms of semantic encoding within HTML5 documents, a syntax called microdata has been developed and published alongside HTML5. Called HTML5's "best-kept secret" by one blogger (Gilbertson, 2010), microdata have been relatively unknown beyond web developer circles until recently when Google, Microsoft and Yahoo jointly announced schema.org, a set of standardized categories and properties for identifying and describing general purpose semantics in web pages using microdata (events, persons, organization, places, etc.). The aim of schema.org is primarily to enhance the performance and presentation capabilities of the three major search engines, thereby hopefully improving user experiences and satisfaction (and ultimately, no doubt, increasing revenues).

Both the microdata standard and schema.org have been met with a certain amount of criticism. For example, it can be argued that microdata lack the expressiveness and flexibility of RDFa, an already existing standard for encoding meaning in web documents and does not therefore constitute any real added value in the context of semantic mark-up. And as for schema.org, the entire project may in a way be seen as a step away from open standards with their insistence on implementation in open forums and permanent availability. For instance, the three companies in question can alter, delete parts of, or altogether remove the documentation at schema.org at any time if they choose to do so.

Nonetheless, the combination of microdata and schema.org vocabularies have the potential, in my view, to become the tool of the trade for many learning content creators willing to add semantic metadata to their web-based materials but reluctant to delve into the finer details of specialized learning metadata models often couched in slightly arcane XML dialects. Here are some reasons why I think this is so:

3 WHY USE MICRODATA?

Firstly, microdata (based on schema.org vocabularies) are simple and relatively easily learned. Microdata encode so-called items, entities or objects, categorize them in one or more classes and assign property values to them. The content of an HTML5 element (section, paragraph, heading, etc.) may thus be marked up to indicate that it deals with, say, an item of the type "person" which has the name property of "Shakespeare":

```
<p itemscope
itemtype="http://schema.org/Person">

<span
itemprop="name">Shakespeare</span> was
born in …

</p>
```

Because microdata are embedded directly, but unobtrusively, in HTML5 elements, they are also accessible to local programming scripts and may in this way be used for content adaptation or user interaction purposes. A simple example would be the visual foregrounding or extraction of all items of a certain semantic type or the assignment of behaviours to items with certain semantic properties. But this is not all. Since semantic encoding is done in a standardized way, useful scripts may be developed, shared and employed on a global scale, and, equally significantly, *across disparate subject matters, disciplines and subjects.* For instance, learning content developers embedding microdata about places and locations in their materials might be be able to download, or point to, scripts, such as jQuery files, mapping these microdata to Google Maps while authors textually describing subject matter concepts and concept relations might be able to utilize available plug-ins to visualize these as concept maps.

Arguably, this is a somewhat novel way of thinking about learning content metadata: Here metadata is not conceived of as ancillary information, detached from the content itself, but as an integral part of its actual *learning design*, possibly initially hidden from the user but ready to be "activated" for specific communicative or didactic purposes, such as catering for different learning styles among users. In semiotic terms, microdata may thus be characterized as embedded resources or vehicles for making meaningful changes in learning material. They may aid in *transforming* learning content, i.e. making changes in the same representational mode, say text, or

contribute to the *transduction* of learning content, i.e. the transfer of material across representational modes, say from text to visual (see Bezemer & Kress, 2008). Metaphorically speaking, microdata may act as "semiotic enzymes" in learning designs.

Secondly, microdata allow authors to mark up digital and non-digital entities alike. This means that a teacher or professor may not only attach metadata to learning objects such as videos, graphics and images providing information about their production history, copyright, intended audience or context of use, etc. but may also specify in some detail what these learning objects are really about. In other words, learning content developers can link learning *objects* to learning *topics* and more generally *documents* to *domains*:

```
<div itemscope
itemtype="http://schema.org/ImageObj
ect">
<img
src="upload.wikimedia.org/wikipedia/
commons/a/a2/Shakespeare.jpg"
itemprop="contentURL" />
<p itemprop="about" itemscope
itemtype="http://schema.org/Person">
<span
itemprop="name">Shakespeare</span>
</p>
</div>
```

In this simple example, it is indicated that an item of the type "image object" located at a specific web address is about an item of the "person" type having the property value of "Shakespeare". But of course much richer semantic descriptions can be attached to the central topic of the image (or any other learning object) if need be. One obvious thing to do is to assign taxonomic classes to domain topics ("Shakespeare is a playwright") and/or to relate domain topics in conceptual structures using relevant associative relations ("Shakespeare wrote Hamlet").

Thirdly, microdata can refer to any vocabulary, taxonomy or ontology class or property with a unique URL on the web and even mix different ones in the same document. So, for instance, one could point to the ALOCOM vocabulary to label the didactic functions of the individual document components of an HTML5 file (glossary, exercise, learning objective and so on) while referring to schema.org for the identification of the domain categories, properties and topics of the subject matter.

This functionality makes it possible, at least in theory, to mark up a variety of meaning types, their representations, and the way these interact. Authors may encode *ideational* semantics, i.e. what the text is about ("the subject matter"), *textual* semantics, i.e. the communicative functions of individual document parts ("introduction", "abstract", "summary", etc.) and *interpersonal* semantics, i.e. meanings relating to the relationship between the writer and his or her audience ("claim", "argument", "evidence" and so on). To make explicit the interaction of ideational, textual and interpersonal meaning in learning materials, a relational genre model along the lines of Martin & Rose (2008) might be applied. In their approach, so-called educational micro-genres – recurrent goal-oriented configurations of meaning in larger texts (aka macro-genres) – can be defined and categorized, in part, according to the way they communicate about "entities" and "events" respectively. While stories and histories are examples of event-oriented micro-genre families, reports of various types classify, describe and explain objects and phenomena, real or imagined. By encoding and exposing micro-genres, their subject matter, and their modes (text, image, video, etc.), we may, eventually, not only be able to search for embedded learning content about a particular topic but also specify its representational characteristics, its didactic function and design and the learning or teaching activities it is intended to support.

4 WHY USE SCHEMA.ORG?

The immediate attraction of schema.org for learning content developers, besides of course the obvious fact that it is supported by the major search engines, is no doubt that it emerges, at least at first glance, as a *one-stop shop* for descriptive tools. The average teacher or professor is unlikely to want to spend a lot of time trawling the web for learning content metadata schemes as well as vocabularies for detailing what the metadata is actually about. But if he or she only has to look in one place, the task seems manageable. However, the categories for labelling and describing things and events currently provided by schema.org are, needless to say, very general in nature and as such do not meet the requirements of most disciplines within the field of education: There is neither a vocabulary for specifying properties of chemical substances nor one for encoding the military rank of historical persons. There are ways of alleviating such problems, though.

In addition to "importing" properties from other vocabularies, schema.org classes can be specialized. For example, it might not be sufficient for a historian to categorize George Armstrong Custer as an instance of http://schema.org/Person in his or her textbook on American history. In this case, the author is free to expand the type specification into, say, http://schema.org/Person/Officer or even http://schema.org/Person/Officer/MajorGeneral. These new categories are unknown to search engines but may be of use to other processing software such as local presentation scripts.

Equally importantly, work is under way to support educational metadata at schema.org. The Learning Resource Metadata Initiative (http://www.lrmi.net/), co-led by the Association of Educational Publishers and Creative Commons, have drafted a specification comprising a limited number of properties describing web resources specifically designed for learning, teaching and education. Some of these properties focus on the intended use of these resources – assignment, exercise or group work, say - while others are centred round the task of aligning learning content with competencies (one piece of content may require a certain competency while another may teach a specific one). Once in place, these properties should be attached to existing schema.org vocabularies denoting general web resources like pages, videos, images and so on.

Some final remarks: by applying HTML5, microdata and pointers to shared terminologies at schema.org to our web-based learning materials, we positively respond to the call for the "unbundling" of learning resources on the web, i.e. the release of "micro pieces of knowledge into the open Net" (Breck, 2008). By marking up "didactic data", so to speak, in our web-based materials, we can better unlock content and make it findable and accessible for reuse and repurposing in new contexts and across multiple platforms. And if microdata and vocabularies like those published at schema.org are widely adopted in the field of education, an increasing amount of structured data sets will eventually be exposed on the web and consumed by an increasing number of e-learning systems and tools in real educational settings. This does not, of course, happen overnight. In the shorter term, however, we can hope to see better search results when we look for learning resources on the web using our favourite search engine. And this is in itself not such a bad start. Also, it goes without saying that the combination of HTML5, microdata and schema.org is no panacea and there are challenges on the educational semantic web which are clearly better addressed using core semantic technologies like RDF, Linked Data, Topic Maps, OWL, etc. Issues pertaining to subject identity, inference and data integration may be a case in point. But even in "traditional" semantic web circles the impact of microdata and schema.org is being felt. Only very recently, for example, RDFa Lite, a less complex version of RDFa was released. RDFa Lite is a lightweight syntax for embedding RDF data, so-called triples, in web pages. It is expected that RDFa Lite will be supported by schema.org in due course giving content developers the choice between microdata or RDF. Such developments surely help lower the barriers of the educational semantic web – especially for the rest of us.

ACKNOWLEDGEMENTS

I am greatly indebted to four anonymous reviewers for their valuable comments.

REFERENCES

Allsopp, J., 2009. *Developing with Web Standards.* Pearson Education.

Bezemer, J., Kress, G., 2008. Writing in Multimodal Texts: A Semiotic Account of Designs for Learning. *Written Communication*. Vol. 25, No. 2.

Breck, J., 2008. Unbundling Online Educational Resources. In: Bruck, P.A. & Lindner, M. (eds.) 2008. *Microlearning and Capacity Building: Proceedings of the 4th International Microlearning 2008 Conference. Innsbruck University Press.*

Gilbertson, S., 2010. Microdata: HTML5's Best-Kept Secret. http://www.webmonkey.com/2010/09/microdata-html5s-best-kept-secret/ (visited 2011-11-11).

Martin, J.R., Rose, D., 2008. *Genre Relations. Mapping Culture,* Equinox Publishing Ltd

Pilgrim, M., 2010. *HTML5: Up and Running,* O'Reilly Media, Inc.

Robertson, R.J., 2011. Technical standards in education, Part 5: Take advantage of metadata. http://www.ibm.com/developerworks/industry/library/ind-edustand5/index.html?cmp=dw (visited 2011-11-11).