DCMF: DC & Microformats, a Good Marriage

Eva Méndez
University Carlos III of Madrid, Spain
emendez@bib.uc3m.es

Leandro M. López
Freelance, Argentina
inkel.ar@gmail.com

Arnau Siches
esbudellat.net, Spain
asiches@gmail.com

Alejandro G. Bravo
Webposible, Spain
alejandrogbravo@yahoo.es

Abstract

This report introduces the Dublin Core Microformats (DCMF) project, a new way to use the DC element set within X/HTML. The DC microformats encode explicit semantic expressions in an X/HTML webpage, by using a specific list of terms for values of the attributes "rev" and "rel" for <a> and <link> elements, and "class" and "id" of other elements. Microformats can be easily processed by user agents and software, enabling a high level of interoperability. These characteristics are crucial for the growing number of social applications allowing users to participate in the Web 2.0 environment as information creators and consumers. This report reviews the origins of microformats; illustrates the coding of DC microformats using the Dublin Core Metadata Gen tool, and a Firefox extension for extraction and visualization; and discusses the benefits of creating Web services utilizing DC microformats.

Keywords: microformats; Dublin Core; DCMES; Web 2.0; metadata; X/HTML; RDF; embedded Web semantics; social applications; bibliographic data repositories

1. Introduction

During the Web 1.0 years ("Altavista Age" that you probably remember), the usual method for including semantic information within documents was using the (X)HTML header <meta> elements, as well as <title>, <address>, <link>, <del>, <ins> elements and "title" and "cite" attributes. This continues in the present, but the abuse ("black SEO") and misuse (inconsistencies) of <meta> elements forces search engines to ignore this information. With the introduction and growing popularity of XML, and the first Recommendation status of W3C's RDF in February 1999 (W3C, 1999b), the potential and versatility of metadata has increased tremendously, supporting more precise and interoperable information gathering and retrieval. The Semantic Web aims to transform the current Web into a machine-readable Web, while maintaining its ability to be directly and easily read by people. However, metadata in webpages is not person-oriented, but search engine-oriented. This metainformation is only available through visualizing source code or using metadata visualization tools (Firefox Dublin Core Viewer extension (2005), or, historically, using special user agents like Metabrowser, which allowed the user to browse both the information and the metainformation within a webpage.

2. Microformats

Microformats originated from a grassroots movement lead by Tantek Çelik to make recognizable data items (such as events, contact details or geographical locations) capable of automated processing by software agents, as well as directly readable by human beings (Knowledge@Wharton, 2005). The official website of microformats.org says that they are *"designed for humans first and machines second, microformats are a set of simple, open data formats built upon existing and widely adopted standards".*

A microformat is a Web-based data formatting approach seeking to re-use existing content as metadata, using only X/HTML classes and other attributes for encoding. Microformats are simple

https://doi.org/10.23106/dcmi.952109254

conventions for embedding semantic markup in human-readable documents. They make use of implicit and explicit X/HTML characteristics in order to add simple semantic information via:

- relationship links using "rel" and "rev" attributes on <a> and <link> elements. Besides the default defined types of relationship in the HTML specification, they can also be extended using profiles.
- "class" and "id" attributes of most X/HTML elements. In this case, in addition to its support for display (as in CSS), these attributes may be used for other different functionalities.

Web developers frequently make use of meaningless values for class names and identifiers. However, source code comprehension can be enhanced and *extra information* added for instance using "header", "menu" and "footer" for page layout definition. In December 2005 Google did an analysis of a sample of slightly over a billion documents, extracting information about popular class names, elements, attributes, and related metadata. One of the goals of that project (Google Web Authoring Statistics) was to know if any logic or semantics were used in class names. The conclusion was that there is no uniformity in naming classes. As a consequence, it is hard to parse documents in order to extract semantic information, except when microformats are used.

The main goal of microformats is to solve problems created by inconsistent labeling, for instance, defining events, people, relationships, etc. through the creation of simple elements and element sets. Some of the microformat element sets are associated with widely adopted standards or schema, such as hCard (based on the vCard standard for business cards) and hCalendar (based on iCalendar for events); some others have a newer origin, like "rel-tag" microformats, used to simplify blog indexing through Technorati. There are also other globally used microformats such as "vote-links" for electronic voting, "hReview" for media reviews, "hResume" for resumes, and "XFN" for social networks, etc.

One of the most obvious and important benefits of using microformats —besides easy encoding and quick distribution— is the ability to easily parse web documents to look for microformats and extract them. There are a number of Web services that exploit this semantic information such as: Technorati[28] to find Weblog posts, Upcoming.org[29] to extract hCalendar definitions of events, and Yahoo! Tech[30] publishing of products reviews etc. Yahoo! has also implemented a search engine for Creative Commons licensed documents[31], and Yahoo! Search parses almost every defined microformat.

3. Dublin Core Microformats (DCMF)

We started the Dublin Core MicroFormats (DCMF) project in 2005, taking advantage of Dublin Core's versatility, general purpose applicability, its formal standardization and the wide promotion by the Dublin Core Metadata Initiative (DCMI). Dublin Core is a metadata schema which is syntactic-independent so it is suitable for encoding semantics within a microformats structure. So, DCMF allow us to extend the indisputable advantages of DCMI —simplicity, flexibility, diffusion and appropriateness— to any domain. All of the microformats have been created with a concrete goal, and the general goal of DC Microformats is to describe web resources (as any resources can have a title, keywords, description, author, etc.). But DC microformats are also particularly appropriate to encode bibliographic descriptions of resources, such as magazines, books, articles, in any media, including paper or digital.

[28] Technorati: http://technorati.com/
[29] Upcoming: http://upcoming.yahoo.com
[30] Yahoo! Tech: http://tech.yahoo.com
31 Yahoo Search: Creative Commons Search: http://search.yahoo.com/cc

https://doi.org/10.23106/dcmi.952109254

3.1. Example: DCMF Encoding

Let's see an example of how we can describe Tim Berners-Lee's book using semantic information encoded as DCMF. The following code will represent this information in an X/HTML webpage:

```
<dl class="dublincore">
<dt>Title:</dt>
<dd class="title">Weaving the Web</dd>
<dt>ISBN:</dt>
<dd class="identifier">0062515861</dd>
<dt>Author:</dt>
<dd><a href="http://www.w3.org/People/Berners-Lee" class="creator">Tim
Berners-Lee</dd></dl>
```

According to the example, to use DC microformats, we need:

1. An X/HTML element (in this example <dl>, a definition list) with the class or identifier "dublincore", which acts as container of a DC microformat and identifies it.
2. A string which represents the semantic expressed by the microformat (in the example, "Title", "ISBN" and "Author").
3. An X/HTML element with the "class" or "id" attributes, whose value is the appropriate DC element to indicate the semantic information to machines (in the example "title", "identifier" and "creator"); and also the value of the element/property (in the example "Weaving the Web", "0062515861" and "Tim Berners-Lee").

If we declare the information expressed in the microformat (Web 2.0 approach) in RDF nomenclature (Semantic Web approach), we should speak about: resource, property and value, where:

- resource, is the value of the element with "identifier" class or id, if it exists;
- property, is the value of the class expressed for both; for humans ("Title", "ISBN" and "Author") and for machines ("title", "identifier" and "creator")
- and value, is the content of X/HTML elements with the class or identifiers of the last item ("Weaving the Web", "0062515861" and "Tim Berners-Lee").

3.2. How to Create DCMF: Dublin Core Metadata Gen

There are many tools to extract and/or generate metadata with DCMI elements, but none allows us to create microformats, except Dublin Core Metadata Gen, which was incorporated into the DCMF project. Dublin Core Metadata Gen is an application developed in PHP that generates three kinds of DC metadata: RDF, X/HTML using <meta> elements and also, per the project presented here, DCMF. In Dublin Core Metadata Gen, you can enter the data into a template and get: DC in RDF, DC in X/HTML using <meta> elements, and DC in microformats.

3.3. How to See DCMF: Dublin Core Microformats Viewer

The Dublin Core Microformats Viewer is an add-on for Firefox and Flock browsers. This user agent's extension detects DC microformats when is then included in the X/HTML code of the webpage. Like Dublin Core Viewer Extension add-on (and inspired on it), DC Microformats Viewer installs a little icon in the status bar, letting the users open a pop-up window containing a table with the Dublin Core microformats present in the current page. This tool is only a simple extension with simple functionality, but it also shows the ease of extracting metainformation from DC microformats, and the potential of this approach.

4. Microformats, <meta> Elements and/or RDF

Microformats are another way of expressing metadata in general, and DC in particular, embedded in web resources. If we compare microformats encoding with the use of <meta>

https://doi.org/10.23106/dcmi.952109254

elements, and/or with RDF syntax, microformats have some advantages and distinctive characteristics:

- Easy to create. Microformats make participation in Web 2.0 social collaboration easier for content creators. Any web content creator can write microformats easily. The only required knowledge is basic X/HTML and X/HTML authoring tools.
- Easy to recognize and use. The information (of an event, a business card, bibliographic record, etc.) can be read by people using their user agents. Users also can extend their browsers' functionalities (mainly by add-ons and widgets, such as Operator for Firefox), to combine pieces of information on websites with applications (e.g. Flickr+Google Maps; Upcoming+ Google Calendar; Yahoo! Local+your address book, etc.).

There are also disadvantages. Probably microformats are less known than the <meta> element, because microformats belong to the emerging domain of the Web 2.0. Also microformats are more limited than RDF; for example, they can not formally define complex relationships and microformats' scope are narrower that the descriptive potential of RDF. Despite all those limitations, microformats are a way to work with DC metadata in the context of Web 2.0, allowing authors to generate semantic information easily comprehensible to both people and machines. Web services can also be developed to support DC microformats, as for any other existing microformats. Examples might include article repositories, books, magazines, etc that allow people to add and find bibliographic records easily.

Microformats, have been also called as "lower-case Semantic Web" but they are a very important inflection point within the Semantic Web. Standards like GRRDL, a recent W3C Recommendation, demonstrate that mechanisms from *Gleaning Resource Descriptions from Dialects of Languages,* are needed to extract Semantic Web Information from X/HTML microformats (W3C, 2007).

5. Conclusions and Future Work

In a post on his blog, Stu Weibel (2006) wrote: *The flexibility that microformats afford is an essential feature of the hyper-innovation that characterizes Web 2.0*, but he wondered if Dublin Core fits in the microformats' philosophy. In this report we answer "YES": Dublin Core fits perfectly in the microformats' philosophy, just as it does in the context of the "classic" Semantic Web. Adopting microformats as a new way to express semantic information with DC allows us to expand the use of DC to new domains that, otherwise, would not use it. In addition, the nature of DC as a general purpose metadata model implicitly suggests its use in microformats for describing resources, especially the bibliographic types of resources previously mentioned.

Microformats avoid the problems of updating and synchronizing the information in many sources (like resumes on employment-related websites) or formats (information visible for people in Web pages, or <meta> elements and RDF for search engines). But microformats especially are intended to allow people to participate in and take advantage of the Semantic Web in the specific situations already mentioned.

The DCMF project intends to combine the simplicity and flexibility of Dublin Core with the possibilities that microformats offer. DCMF is an attempt to make semantic information easy and practical. Furthermore, the ease of parsing web documents with microformats lets us use this semantic information for Web services useful to people.

Future work on DC microformats will be the evolution and improvement of those tools described here (Dublin Core Metadata Gen and Viewer and Dublin Core microformats), and the development of Web services for querying the information within DC microformats.

https://doi.org/10.23106/dcmi.952109254

References

Bravo, Alejandro, and Arnau Siches. (2005). *Dublin Core Metadata Gen: Generator of metadata using Dublin Core.* Retrieved from http://www.webposible.com/utilidades/dublincore-metadata-gen.

Conolly, Dan. (2007). *Gleaning Resource Descriptions from Dialects of Languages (GRDDL) W3C Recommendation 11 September 2007.* Retrieved, April 1, 2008, from http://www.w3.org/TR/2007/REC-grddl-20070911/.

DCMF. (2008). *Microformatos Dublin Core* (Translated to Spanish). Retrieved from http://webposible.com/microformatos-dublincore/.

DCMI. (2008). *Dublin Core Metadata Initiative.* Retrieved from http://dublincore.org.

DCMI. (2008). *Dublin Core Metadata Initiative - Tools and Software.* Retrieved from http://dublincore.org/tools.

Google. (2006). *Web Authoring Statistics.* Retrieved from http://code.google.com/webstats/.

Kaply, Michael. (2008, May 21). Operator 0.9.3 for Firefox. Posted to https://addons.mozilla.org/es-ES/firefox/addon/4106.

Knowledge@Wharton. (2005, July 27). *What's the Next Big Thing on the Web? It May Be a Small, Simple Thing - Microformats.* Retrieved, April 1, 2008, from http://knowledge.wharton.upenn.edu/index.cfm?fa=printArticle&ID=1247.

Kumar, Amit. (2008, March 13). The Yahoo! Search Open Ecosystem. Message posted to http://www.ysearchblog.com/archives/000527.html.

Lauke, Patrick H. (2005). *Firefox Dublin Core Viewer Extension.* Retrieved from http://www.splintered.co.uk/experiments/73/.

López, Leandro M. (2008, March 19). Visor de Microformatos Dublin Core: Extensión para Firefox. Message posted to http://wses.wordpress.com/2008/03/19/visor-de-microformatos-dublin-core.

Metabrowser. (n.d.). Retrieved from http://metabrowser.spirit.net.au. (Metabrowser is not longer available).

Ora, Lassila, and Ralph R. Swick. (1999b). *Resource Description Framework (RDF) Model and Syntax Specification W3C Recommendation 22 February 1999*. Retrieved, April 1 2008, from http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

Raggett, Dave, Arnaud Le Hors, and Ian Jacobs (Eds.). (1999a). *HTML 4.01 Specification W3C Recommendation 24 December 1999.* Retrieved, April 1 2008, from: http://www.w3.org/TR/html401.

Weibel, Stuart. (2006, April 12). Ockham's Bathroom Scale, Lego™ blocks, and Microformats. Message posted to http://weibel-lines.typepad.com/weibelines/2006/04/ockhams_bathroo.html.

https://doi.org/10.23106/dcmi.952109254