SEMANTIC WEB ADOPTION: ONLINE TOOLS FOR WEB EVALUATION AND METADATA EXTRACTION

RAFAEL PEDRAZA-JIMÉNEZ, LLUÍS CODINA, CRISTÓFOL ROVIRA

Department de Periodisme i de Comunicació Audiovisual, Area de Coneixement de Biblioteconomia i Documentació, Universitat Pompeu Fabra, La Rambla, 30-32, Barcelona, 08002, Spain

This work briefly analyses the difficulties to adopt the Semantic Web, and in particular proposes systems to know the present level of migration to the different technologies that make up the Semantic Web. It focuses on the presentation and description of two tools, DigiDocSpider and DigiDocMetaEdit, designed with the aim of verifying, evaluating, and promoting its implementation.

1. Introduction

In 2001 Berners-Lee and his colleagues made known to the public at large the project of the Semantic Web [1]. This project augured deep changes that would affect, and, in fact are already affecting, the fields of creation, edition and publication of web pages and sites.

This dissertation briefly analyzes the present state of adoption of the Semantic Web, and two tools designed by the authors are introduced. The first one, the spider DigiDocSpider, springs from the necessity of having indicators that allow us to know and evaluate the actual level of implementation of the Semantic Web.

The second one, the editor DigiDocMetaEdit, was created with the aim of providing Web users with a tool that allows the automatic generation of metadata according to the standards of the W3C, and thus, contributing to the adoption of the Semantic Web.

2. The Semantic Web Today

In December 2007, Scientific American journal published an article entitled "The Semantic Web in Action" [2] in order to show, through the presentation of several cases studies, the potential that Semantic Web technologies have reached.

However, it is important to underline that the existence of this technology does not mean the existence of the Semantic Web. There are very few actual implementations that support the technologies of the Semantic Web, specially when it involves an unambiguous interaction between an agent and a user, or if it requires the capacity of a machine to make a decision or to adapt itself to a context.

Nevertheless, it can be affirmed that an approach of the users of the present Web to the Semantic Web is taking place. Initiatives like FOAF, which allows the description and identification by means of RDF of its users; the emergence of folksonomies, which introduce users in the tasks of description and labelling; or the creation and use of syndicated content channels with RSS or ATOM (also RDF technologies) are good examples of this approach.

3. Indicators for measuring the adoption of Semantic Web

In this way, the tangible outcomes of the Semantic Web scarcely occur and when they do, only in limited contexts of the present Web. In our opinion, the beginning of the migration to the Semantic Web requires, at least, these three conditions:

1. *Source code quality*: the source code of web pages must be consistent and without errors.
2. *Use of metadata*: enough quantity and quality of the metadata used to describe web pages.
3. *Use of RDF:* encoding of metadata using RDF.

These three elements seem to be the most reliable indicators of the evolution to the Semantic Web, according to the W3C's own road map.

First of all, it seems evident that the Semantic Web will not be possible with the actual level of errors on source code of web pages. These errors have been systematically compensated by the high tolerance of browsers to incorrect code.

The problem is that this tolerance can not be transferred to the software that will work in the Semantic Web for two reasons: on the one hand, because all the infrastructure rests on XML, which, in turn, demands a systematic, rigorous and consistent encoding; on the other hand, because the software agents that will work on the Semantic Web will run a kind of processing much more complex, where any error will make insecure its work (or will just stop it).

The quality and the quantity of meta-information is other of the indicators studied. The more and best metadata a page contains, the better to achieve the necessary level of exigency to mark semantically a document, and be able this way to obtain the expected future outcomes of the Semantic Web. Finally, the

encoding on RDF is the adequate format to encode metadata for the Semantic Web, and, therefore, its use is a clear indicator of migration.

4. Online tools for web evaluation and metadata extraction

The research group DigiDoc (http://www.digidocweb.net) has developed two tools in order to favour not only the evaluation of the present Web but also the adoption of the Semantic Web. The first one, the spider DigiDocSpider, originates from the necessity to have indicators that allow us knowing and evaluating the actual level of implementation of the Semantic Web.

The second tool, the editor DigiDocMetaEdit, was created with the aim of providing Web users with a tool that makes possible the automatic generation of metadata according to W3C standards, therefore, contributing to the process of adoption of the Semantic Web. In the following paragraphs, both applications are described in detail.

4.1. *DigiDocSpider*

DigiDocSpider is an open source software with GPL license for the automatic analysis of web pages. It is a spider that enables the analysis of pages of a Web site on the basis of three types of indicators:

1. Regular expressions.
2. Data extracted from other services: DigiDocSpider can send the page which is analysing to several validation services available on the Internet (xhtml, css, accessibility, etc.), collect the result of those external analyses and include it on its data base.
3. Manual indicators.

DigiDocSpider is a configurable tool. Nowadays, for each of the pages, it compiles automatically more than 100 indicators relative to three aspects: accessibility, rankings on search engines (causes and effects) and quality of xhtml code.

It starts working with the input to the programme of a set of URLs, to extract information about elements from the code html or from external validation services that are of some interest. The output of this application is a report that can take different forms, among them, the ranking that allows the URLs analysed to be ordered according to their quality.

In order to analyse the quality of xhtml code, DigiDocSpider provides users with a set of initial indicators relating to the errors most commonly found in Web's source code. Among them, we find, for instance: the absence of heading

tags (<h1>, etc), the insertion of an image <img> with no “ALT” attribute, the lack of the <title> element on the head of Web pages, etc.

The authors have already carried out (and published) two studies where this tool has been used to measure the quality of source code and determine the stage of evolution of the present Web to the Semantic Web. The first research was undertaken in December 2005 [3], on three thematic groups of Web sites related to distribution of knowledge.

The whole analysis and the data obtained by DigiDocSpider can be found in (Codina and Rovira, 2006), although the conclusions drawn from it can be summed up in these three statements: 1. There was a high rate of encoding errors; 2. Metadata were quite scarce; 3. RDF encoding was almost inexistent. Therefore, the analysis confirmed that in the analysed Webs, migration to the Semantic Web had not yet started.

Another research [4], carried out in October 2007, analysed the Web sites of the digital repositories from the European Union countries registered in Roar (259 Web sites). The tables with the data of all the areas analysed can be found in the URL: http://www.observaWeb.com/repositorios.htm. Following the analysis of those data it could be concluded that, as for the accessibility and quality of source code, in general terms the results of this research conveys that accessibility to repository Web sites in Europe should be improved.

There were found meaningful deficiencies in encoding xhtml and css, the presence of deprecated elements by the W3C, and the incomplete or wrong encoding of images, forms, links and tables tags. There was a relatively high number of errors in html encoding, css as well as use of obsolete labels. All these deficiencies make evident the scarce (or void) migration to the Semantic Web, since basic components are missing (XML and RDF layers). These results contrast with the ones obtained when studying the aspects of Web positioning, which are especially positive reaching high standards of luminosity and visibility.

The list of indicators of DigiDocSpider and examples of several analysis and rankings carried out with it can be found in: http://www.observaWeb.com

4.2. *DigiDocMetaEdit*

DigiDocMetaEdit is a metadata extractor created by DigiDoc Laboratory (http://www.documentaciondigital.org/laboratorio.htm). It has been developed as a free application according to GPL license. Its implementation has been done in a modular way so that it is easily extendable and sustainable. Initially it has been structured in two main modules:

Extraction module: the objective of this module is the extraction of the contents from <meta> tags of the page to be edited. This module, programmed in Perl, has a list of meta tags to be extracted that are easily extendable by changing the source code.

Presentation and edition module: the objective of this module is to present the meta tags extracted from forms in order to allow their edition and to generate standardized fragments of metadata according to diverse standards. On its first version, the code was standardized following XHTML 2.0, RDF 1.0, Dublin Core, and Microformats Dublin Core. This module has been programmed using Javascript and HTML, and it is available in Spanish, Catalan and English.

Regarding its functioning (a beta version of this editor can be seen in http://www.metaeditor.net), this tool allows to select the documents to be analysed in two ways: 1. The user introduces a URL address; 2. The user supplies a HTML file that is kept on his hard disk. Once analysed the document or item, DigiDocMetaEdit gives the user the descriptive metadata that has automatically extracted from it, like the title, description, keywords or language. The user will check this data through an easy and clear interface, and will modify or complete them if necessary.

Besides, DigiDocMetaEdit allows the addition of other informative metadata to the application, such as collaborators, publisher, type of resource, intellectual property rights, date of creation, or any other resources related to the one that is under analysis.

If the user does not introduce any URL or document, DigiDocMetaEdit will show all its fields void. In this case, the user can optionally specify some or all of the informations required by the various fields of the application, and this will generate the encoded metadata according to each standard for such information. Having the result, the user just has to copy on his document the generated code fragments.

4.3. *Future work*

The tools presented, DigiDocSpider and DigiDocMetaEdit, are still Beta versions to be developed. At this moment, both applications are being improved: regarding DigiDocSpider, by enlarging the module for extracting meta tags. Besides, it is planned the addition of new standards for generating metadata to the module.

DigiDocMetaEdit is also been improved with the addition of a module that allows the suggestion of keywords. What is more, such module is expected to

evaluate and point out to the users the keywords for which their documents are better positioned. This way, system users will know whether their documents fulfil their expectations of positioning, and if they do not, users will have with this tool a way of improving their documents.

5. Conclusions

The Semantic Web is nowadays a real idea force, in the sense that it is an idea that has been able to mobilize energies (and illusions) and that, no doubt, will bring about positive results during the next years. However, it is difficult to guess the level of actual adoption of this initiative.

That situation has promoted the creation of the tools presented in this work. In particular, the aim of DigiDocSpider is to try to alleviate the existing uncertainty about the level of adoption of the Semantic Web, and to contribute to its adoption by providing a way of evaluating, improving and correcting the source code of html documents.

The metadata extractor DigiDocMetaEdit was created starting from the results from the first analysis of DigiDocSpider, that confirmed the scarce presence of metadata related to Web contents. DigiDocMetaEdit is intended for promoting the use of metadata, facilitating their generation in an automatic way using an easy interface.

We hope that future extensions and improvements of these tools will contribute to the process of migration from the present Web to the Semantic Web.

6. Acknowledgments

This work has been supported by the research project “Web Semántica y Sistemas de Información Documental” of the Ministerio de Educación y Ciencia, reference HUM 2004-03162 / FILO.

7. References

1. T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web”, *Scientific American*, vol. 284, nº 5, May 2005, pp. 34-43 (2001).
2. L. Feigenbaum, I. Herman, T. Hongsermeier, E. Neumann and S. Stephens, “The Semantic Web in Action”, *Scientific American*, vol. 297, nº 6, Dec. 2007, pp. 90-97 (2007).
3. L. Codina, and C. Rovira, “La Web semántica” in Tramullas, Jesús, Eds. *Tendencias en documentación digital*, chapter 1. Trea, (2006).

4. C. Rovira, M. Marcos and L. Codina, “Repositorios de publicaciones digitales de libre acceso en Europa: análisis y valoración de la accesibilidad, posicionamiento web y calidad del código digital”. *El profesional de la información* 16(1), En. 2007, pp, 24-38, (2007).