Where Is The Semantic Web? – An Overview of the Use of Embeddable Semantics in Austria

Wilhelm Loibl

Institute for Service Marketing and Tourism
Vienna University of Economics and Business, Austria
wloibl@wu.ac.at

Abstract

Improving the results of search engines and enabling new online applications are two of the main aims of the Semantic Web. For a machine to be able to read and interpret semantic information, this content has to be offered online first. With several technologies available the question arises which one to use. Those who want to build the software necessary to interpret the offered data have to know what information is available and in which format. In order to answer these questions, the author analysed the business websites of different Austrian industry sectors as to what semantic information is embedded. Preliminary results show that, although overall usage numbers are still small, certain differences between individual sectors exist.

Keywords: semantic web, RDFa, microformats, Austria, industry sectors

1 Introduction

As tourism is a very information-intense industry (Werthner & Klein, 1999), especially novel users resort to well-known generic search engines like Google to find travel related information (Mitsche, 2005). Often, these machines do not provide satisfactory search results as their algorithms match a user's query against the (weighted) terms found in online documents (Berry and Browne, 1999). One solution to this problem lies in "Semantic Searches" (Maedche & Staab, 2002). In order for them to work, web resources must first be annotated with additional metadata describing the content (Davies, Studer & Warren., 2006). Therefore, anyone who wants to provide data online must decide on which technology to use. Those who intend to build applications that use the offered data have to know what information is available and in which format. This work provides answers to these questions. First, the sectors within the tourism industry are compared to each other. Then, the tourism industry as a whole is contrasted with all other business sectors in Austria. The remainder of this section gives a short introduction into Semantic Web technologies. Section 2 explains the research methodology used to obtain the preliminary results presented in section 3. Managerial implications for all involved stakeholders are outlined in section 4, while section 5 provides an overview of the main issues which still remain to be addressed. Section 6 summarises the preliminary results.

1.1 The Semantic Web

In contrast to syntax, which is only grammatical rules, semantics is about meaning. Additional information describing the meaning of a document can be put into an external document using ontologies. A ontology is a list of terms and relationships between these terms which are used to formally describe a certain domain (Antoniou & van Harmelen, 2008). A domain in this regard is a collection of all entities about a certain subject (Hjorland & Albrechtsen, 1995). The online tourism domain therefore is composed of all information entities related to travel (Xiang, Wöber & Fesenmaier, 2008). Specific ontology languages like RDF or OWL are used to describe the classes of objects and the relationships. Feilmayr and Pröll (2009) provide an overview of tourism related ontologies. Another possibility for adding metadata to online resources is by inserting it directly into the code of a web page.

The two main representatives of these embeddable formats are microformats and RDFa. Microformats reuse existing HTML code by inserting `class`-attributes with specific values. These values (the vocabulary) are defined in a profile (Lewis, 2010). The main advantages are that they use only existent technology and are therefore easy to learn (Yu, 2011). Each microformat is introduced by a certain root-element inserted as value of a `class`-attribute. RDFa creates new attributes using the RDF schema mechanism. These vocabularies may be exchanged and intermixed which makes RDFa extensible. Another dvantage of RDFa over microformats is that it is a W3C standard (Yu, 2011). Additional embeddable formats are embeddable RDF (eRDF) and Microdata. Embeddable RDF is a subset of RDF which can be placed in XHTML and HTML. Although RDFa and Microdata superseded eRDF, it has been mentioned here because it may still be used on some websites (Davis, 2012). Microdata is a mechanism which allows to embed machine-readable data in HTML documents (Hickson, 2012).

2 Research Methodology

In order to answer the questions outlined in section 1, a list of website-URLS for each industry sector in Austria was compiled using the online search service "Firmen A-Z"[1] offered by the Austrian Chamber of Commerce. This search yielded lists for the sectors "banking", "craft", "trade", "industry", "information and consulting", "tourism" and "transportation". "Tourism" was further divided into "leisure and sport", "gastronomy", "health", "accommodation", "culture" and "travel agencies". Each list was fed into a crawler[2] which, in a first step, gathered all web pages crawlable for the respective website. In a second step, the code of the individual web

[1] http://firmen.wko.at/Web/SearchSimple.aspx

[2] The technical documentation of the Spyglass crawler by W. Loibl can be found at http://tourism.wu-wien.ac.at/java/javadoc/spyglass/

page was analysed by searching for certain code-patterns. These patterns conform to the specific root-elements used to introduce a certain semantic format. Microformats use predefined values for the `class`-attribute to convey semantic information. The root-elements for hCalendar, hProduct, hMedia, adr, hListing, hAtom, hCard, hReview, hNews, hResume, hRecipe and geo were used to identify the respective microformat (Celik, 2012). In order to detect microdata, the crawler looked for `itemscope`-attributes (Hickson, 2012). RDFa was detected either through `vocab`- and `prefix`-attribues (Adida, Birbeck, McCarron, & Herman, 2012) or through the indication of one or more XML namespaces used for defining which vocabulary was used (Adida, Birbeck, McCarron, & Pemberton, 2008). Embedded RDF is indicated by a link to the eRDF profile (found at http://purl.org/NET/erdf/profile) in the `head`-element (Hebeler, Fisher, Blace, & Perez-Lopez, 2009, p. 392). A total of 58,837 websites with 307,365 web pages were analysed. After removing double counts (e.g., businesses which were listed as "tourism" companies as well as in the sector of "transport") 40,604 websites with 232,096 web pages remained.

3 Preliminary Results

The preliminary results for the tourism industry shown in table 1 already suggest that overall usage numbers are still relatively small. Nevertheless, certain differences between the individual subsectors exist.

Table 1. Overview of the preliminary results for the tourism industry

Sectors	Sites	Pages	Format/Site	Avg. Format/Page	Format/Page
Gastronomy	2,252	12,273	6.26%	10.52%	6.36%
Leisure and Sport	1,516	8,791	6.27%	15.99%	6.65%
Health	276	1,727	5.43%	14.77%	7.82%
Accommodation	3,341	19,025	5.45%	13.35%	5.84%
Culture	77	492	6.49%	9.35%	7.52%
Travel Agencies	461	3,138	5.86%	8.73%	5.48%
Sum (with doubles)	7,923	45,446	5.87%	12.79%	6.21%
Sum (no doubles)	6,823	38,602	5.75%	12.53%	5.84%

Column 4 of tables 1 and 2 shows how many websites used semantic formats in them, while column 6 does the same for the individual web pages. The total number of semantic markup found divided by the total number of websites is provided in column

5. Especially sectors with less websites use the adr-microformat. Gastronomy, accommodation and leisure businesses are the only users of microdata. The most often used format is RDFa ranging from 43.48 % (culture subsector) to 74.45 % (travel agencies) of all formats.

The summarised results for all Austrian business sectors suggest that the top users of the Semantic Web are not in the tourism but in the transport sector. Especially the use of hCalendar and Microdata is very high within this industry.

Table 2. Overview of the preliminary results for all Austrian business sectors

Sectors	Sites	Pages	Format/Site	Avg. Format/Page	Format/Page
Banking	244	829	2.87%	3.74%	3.74%
Tourism	6,823	38,602	5.75%	12.53%	5.84%
Transport	2,679	16,083	5.86%	13.20%	6.03%
Industry	1,779	10,890	5.45%	9.48%	5.91%
Information	11,734	66,652	5.62%	11.16%	4.99%
Craft	14,802	84,718	5.13%	10.00%	4.45%
Trade	15,312	89,590	5.37%	10.20%	4.68%
Sum	53,373	307,364	5.42%	10.76%	4.94%
Sum (no doubles)	40,604	232,096	5.33%	10.62%	4.74%

RDFa is by far the most often used format, ranging from 93.55 % in the banking sector to 63.18 % in the industry sector. With 21.9 % the industry sector is the biggest user of Microformats, especially hCard (8.24 %) and hAtom (8.62 %). In the tourism industry about 70.84 % of all semantic formats are RDFa, 9.34 % are Microdata and 19.82 % are various Microformats, mostly adr (8.43 %), hCard (5.27 %), hAtom (3.63 %) and hCalendar (2.15 %). This clear prevalence of RDFa over other semantic formats is in stark contrast with recent findings by Bizer, Mühleisen, Harth, & Stadtmüller (2012) who suggest a dominance of microformats over RDFa.

4 Managerial Implications

This work is intended to give an overview of what type of semantic information is available and in which format. Knowing this, businesses can invest in technology which is already widely employed. Using widely deployed systems help keeping

investment cost low. As the tourism industry would greatly benefit from these technologies, it is startling to see that seemingly so little has been done yet.

5 Known Issues and Conclusion

There are three main issues still to be addressed. First, the search pattern for discovering RDFa information has to be improved. Second, better methods for finding differences between the individual sectors must be applied. Last but not least, as the preliminary results show that RDFa is the prevalent technology, it would be interesting to know which vocabularies are used in the different sectors.

The preliminary results presented in this paper show that the overall use of semantic formats is still in its infancy. Additional research has to uncover the reasons behind this phenomenon facilitating counteractions.

References

Antoniou, G., & van Harmelen, F. (2008). *A Semantic Web Primer*. The MIT Press.

Berry, M. W., & Browne, M. (1999). *Understanding Search Engines. Mathematical Modeling and Text Retrieval*. Society for Industrial and Applied Mathematics (SIAM).

Bizer, C., Mühleisen, H., Harth, A., & Stadtmüller, S. (2012). Web Data Commons. Retrieved from http://webdatacommons.org/

Davies, J., Studer, R., & Warren, P. (2006). *Semantic Web Technologies*. Wiley.

Davis, I. (2012). Embedded RDF. Retrieved from https://github.com/iand/erdf

Feilmayr, C., & Pröll, B. (2009). Ontologiebasierte Informationsextraktion im eTourismus. In M. Lassnig and S. Reich (Eds.), *eTourismus - HMD - Praxis der Wirtschaftsinformatik*, Volume 270.

Hickson, I. (2012). HTML Microdata. W3C Working Draft. Retrieved from http://www.w3.org/TR/microdata/

Hjorland, B., & Albrechtsen, H. (1995). Toward a new horizon in information science: Domain-analysis. *Journal of the American Society for Information Science*, 46(4):400–425.

Lewis, E. (2010). *Microformats made simple*. New Riders (a division of Pearson Education).

Maedche, A., & Staab, S. (2002). Proceedings of the 9th international conference for information and communication technologies in tourism: Enter 2002. innsbruck, austria. In K. Wöber, A. Frew, and M. Hitz (Eds.), *Applying Semantic Web Technologies for Tourism Information Systems*.

Mitsche, N. (2005). *Information and Communication Technologies in Tourism*, chapter Understanding the Information Search process within a Tourism Domain-specific Search Engine. Vienna: Springer.

Werthner, H., & Klein, S. (1999). *Information technology and tourism: A challenging relationship*. Vienna: Springer.

Xiang, Z., Wöber, K., & Fesenmaier, D. R. (2008). Representation of the online tourism domain in search engines. *Journal of Travel Research*, 47(2):137–150.

Yu, L. (2011). *A Developer's Guide to the Semantic Web*. Berlin/Heidelberg.