Metadata is back!

Bernhard Haslhofer
Department of Information Science
Cornell University
Ithaca, New York, USA
bernhard.haslhofer@cornell.edu

1. TALK SUMMARY

I chose the title *Metadata is back!* because of schema.org, which is a joint effort of Bing, Google and Yahoo to provide a collection of schemas for marking up Web pages with machine-readable data. This helps search engines to understand the meaning of information and allows them to provide rich search results to the user, which makes it easier for them to find relevant information on the Web. But schema.org is only the latest in a series of attempts of adding machine-processable metadata to the Web and the goal of this talk was to show how libraries use metadata publishing techniques and to discuss possible future directions.

Librarians have been using *metadata* for centuries. Because of the increase in the production of printed materials they suddenly had to deal with that information overload and needed novel mechanisms to find the right resources. Libraries understood the value of metadata and developed *identification* mechanisms, *controlled vocabularies*, and *cataloguing rules* for describing their resources.

With the advent of the World Wide Web some of the aforementioned information organization mechanisms, especially metadata and controlled vocabularies, were superseded by full-text information retrieval and natural language processing techniques, which work pretty well for typical Web search use cases. Many end users now favor the Web as an information platform over the library catalogue. They prefer the simplicity of Web search, which delivers the most relevant results in the first place, and increasingly rely on online social networks to fulfill their information needs.

In 2007 *Linked Data*, the little cousin of the Semantic Web, was born. Motivated by the problem that machines have a hard time in understanding the meaning of the Web pages, Linked Data focuses primarily on the data-centric aspects of the Semantic Web and is therefore less complex and easier to implement. People started to combine this simple data publishing method with the idea of "open data", which led to the Linking Open Data community project. The goal of this project is to publish and link data on the Web and use the Web to build a globally connected data network. This development also caused momentum in the library world.

After the Library of Congress and the Swedish Union Catalogue began to adopt the Linked Data principles, many other libraries followed and exposed their bibliographic metadata on the Web. Some implement Linked Data services, other published dumps of their metadata in public data sharing platforms such as `thedatahub.org`. At the time of this writing, the *bibliographic data* group comprises 73 datasets, which includes bibliographic metadata and vocabularies from the Library of Congress, the British Library, the Bibliothèque National de France, Europeana, and other major institutions in the library domain.

The technologies provided by the Semantic Web stack, such as URIs, RDF/S, OWL, SKOS, and SPARQL, enabled the transition from closed-world data representation mechanisms to open, URI-addressable resource representations and metadata descriptions. However, the existing data publishing recipes proposed by the Linked Data community require some fundamental understanding of the Web Architecture, which in practice often means that developers of Linked Data services have difficulties in distinguishing between non-information and information resources, providing the correct resource representations, and correctly redirecting requests between them.

RDFa and Microdata are promising alternative technologies because they allow developers to embed data directly into the human readable HTML representation of a Web resource. Schema.org and the Facebook Open Graph protocol are major developments into that direction and librarians should, at least in my opinion, think how they could not only adopt but also participate in the development of these technologies and, given their expertise, contribute to the definition of schemas for information entities for the Web.