Metadata for the Masses

Christian Lieske, SAP AG, St. Leon-Rot, christian.lieske@sap.com

Dr. Matthias Kammerer, SAP AG, Walldorf, matthias.kammerer@sap.com

1 Introduction

Discussions surrounding metadata – data about data – often do not make important distinctions, such as metadata that helps during production (e.g. a production status indicator), and metadata that helps to generate revenues from sales (e.g. price information on an Internet page). Similarly, metadata for unstructured content (e.g. marketing collaterals) and metadata for structured data (e.g. product catalogues) is all too often mixed up. As possible effect of this, the uptake and impact of metadata until recently has not reached its full potential. Now, however, matured and recognized standards have found their way into business and mainstream applications. Thus, we now have metadata for the masses – including anyone in technical communications.

We will tackle the topic by first sketching that standards-based metadata really have arrived in the mainstream – and which benefits adopters can reap. This sketch is followed by an explanation of some important general concepts related to metadata. Afterwards we will look at several metadata standardization activities with particular importance for technical communications, and follow up with a discussion on the use of support for standardized metadata in mainstream applications. Finally, a case story shows that the use of standardized metadata is advantageous even in especially in projects where deadlines are tight.

2 Business Value of Mainstream Standards-Based Metadata

Metadata is useful amongst other things during production and revenue generation: A content-related status indicator such as “in review”, for example, keeps publication processes manageable; price information or a product review on an Internet page attracts customers and thus helps to sell a product or service. Standards-based metadata has even more advantages. On the one hand, standards incorporate a solid body of knowledge – thus, you do not need to build up this knowledge or to learn by trial and error. On the other hand, you often find software tools or libraries that help you to easily marry your content with standards-based metadata – think, for example, of the ways in which web servers, browsers or stylesheets control content delivery, based on standardized language tags such as “ja” (see [1]).

Due to the ubiquitous presence of Internet technology in modern business, one dimension of metadata has gained pivotal importance: labeling, cataloging, and describing resources on the Web in such a way that they can be properly searched for and processed. Today the corresponding technology stack (see [2]) – known as the Semantic Web architecture – and a set of best practices for publishing and connecting structured data on the Web – known as Linked Data (see [3]) – are already helping early adopters to reap benefits. Retailers, for example, now make use of the Resource Description Format (one of the ingredients in the Semantic Web) to accommodate price information or reviews in their web presence (see [4]) and improve the way results from this Web content are processed by Google (see [5]); the UK Government has also become active, since it believes that Linked Data generates significant primary material for digital products and services that could contribute to economic growth (see [6]).

3 Metadata Concepts

Metadata is data about data – or – more aligned with the terminology of today’s Internet: metadata
describes resources. For example: All of the PDF files available on a company’s Web site are resources, and each of them may be annotated with descriptive information, such as the name of the publisher, or copyright information. A number of reasons exist for providing resource descriptions:

1. Enable or facilitate the use of a resource – for example, making a resource easier to find
2. Help to interpret data/bridge a semantic gap – for example, by indicating that a certain string of numbers represents a price
3. Allow easy and fully automated processing by machines – for example, by indicating that certain content (e.g. a snippet from a computer program) must not be considered during spell checking

Important concepts related to metadata are: metadata models, metadata categories, and metadata encodings. The distinction more or less corresponds to the questions: “What?” and “How?”. You may, for example, need to describe a product along the lines of the name of the product in a certain country, the review status of the product’s user manual, and the product’s suggested retail price on a certain regional market. In this scenario, the data model for the product would include the data categories: Regional product name, Status of regional user manual review, and Regional retail price. Any of these data categories may be constrained with respect to possible values. The values for Status of regional user manual review could, for example, consist of the list “Ready for start”, “Started”, “Paused”, and “Finished”. The scope of the data categories could be the entity product. Of course, a data category such as Regional retail price could also apply to a different type of entity (e.g., a service). The encoding of a metadata model relates to questions such as:

– Do I use a separate database table for every data category, or do I use a single database table for all data categories?
– What element name do I choose to represent Regional retail price in an XML file?

Nowadays, standardization and extensibility are additional concepts of high importance for the world of metadata. For example: To encode information about the language that is used in an HTML/XML document or element, a set of standards should be used (namely the attribute which typically shows up as “xml:lang” and a set of values such as “ja” (see [7])). Thus, the standardization in this area pertains to using a certain attribute and a certain set of values to encode language information. Extensibility acknowledges the fact that often the “one size fits all” approach is not valid – special contexts may not allow you, for example, to categorize an entity in terms of a predefined set of values. Thus, quite a number of standardization approaches provide mechanisms to work with non-standard values. For example: The standard for Internet Media Types uses the naming convention “x” for types or subtypes that are nonstandard (see [8]).

4 Selected Metadata Activities Related to Technical Communications

The following sketches a couple of activities from realms such as fundamental metadata, metadata related to multilingual use, special purpose vocabularies, and generic encoding formats. The sketch is skewed towards activities which already are – or with some degree of certainty will be – deployed in within technical communications.

4.1 Fundamental Metadata: Dublin Core

Technical communications these days deal with physical resources (such as books), digital materials (such as video), sound, image, or text files, and composite media like web pages. The Dublin Core set of metadata elements easily relates to all of these resources. It provides a small set of 15 data categories and values for areas such as audience and creator (see [9]). It has achieved wide dissemination as ISO Standard 15836:2009 among other things.
4.2 Metadata Related to Multilingual Use: BCP47 and W3C ITS

Consumers want to be able to read or see text or other items in their native language. Thus, metadata is needed that provides information related to the language that was used, for example, while authoring a certain document fragment. BCP47 (see [10]) currently is the standard for language related information. BCP74 is highly versatile, and defines values such as “zh-Hant-HK” (Traditional Chinese as used in Hong Kong).

The majority of text that exists in more than one language is created through translation. The translation process depends on information such as: Does this phrase need to be translated? Is this phone number valid in more than one country? Here is where the Internationalization Tag Set (ITS), a standard from the World Wide Web Consortium (W3C), comes into play (see [11]). It allows any XML-based content to be annotated with metadata answering so-called internationalization questions such as the aforementioned ones.

4.3 Special Purpose Vocabularies: vCard

“vCard is a file format standard for electronic business cards. vCards are often attached to e-mail messages, but can be exchanged in other ways, such as on the World Wide Web. They can contain names and addresses, phone numbers, e-mail addresses, URLs, logos, photographs, and even audio clips.” (quote from [12]). Thus, vCard has a different scope than Dublin Core, and hence is somewhat more granular than Dublin Core, which does not look into creator details such as the creator’s e-mail address.

4.4 Generic Encoding Formats: Microformats, RDF, RDFa, and HTML Microdata

4.4.1 Microformats

Microformats (see [13]) are based on the fact that XHTML and HTML allow for semantics via the CLASS and REL attributes. Existing microformats cover domains such as contact information, geographic coordinates, calendar events, and social relationships. Since semantics is explicitly encoded and represented according to a standard, automated processing is easier. Several browsers realize support for microformats (possibly via extension/plug-in; see [13a]).

4.4.2 Resource Description Framework (RDF)

The Resource Description Framework (RDF) provides a general mechanism for representing information about resources in the World Wide Web. It is particularly intended for representing metadata about Web resources (such as the title of a Web page). In addition, it can be used to represent information about things that cannot be retrieved on the Web (such as a person).

RDF standardizes the way to make statements/provide descriptions. A key element is the use of so-called Uniform Resource Identifiers (URIs) (e.g., file://foo#) instead of words to name things in statements/descriptions. The URIs used in RDF form a vocabulary (e.g., for typical Web page metadata such as title, author, etc.)

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dcterms="http://purl.org/dc/terms#">
  <rdf:Description
    rdf:about="http://www.example.org/infoShed.html">
    <dcterms:creator
```
The most useful type of RDF acknowledges the fact that metadata benefits (for example automatically linking data such as a person’s name with a more complete record of that person’s data) can only be harvested easily, if standard vocabularies are used wherever possible. The above exemplifies this: The first “rdf:Description” makes use of Dublin Core, whereas the second works with proprietary markup. Moreover, the example clearly relates itself to the Dublin Core notion of “creator”, and not the notion of “creator” which has been created by a specific scientific community.

### 4.4.3 RDF in Attributes (RDFa)

The Resource Description Framework in Attributes (RDFa) enables XML (eg. XHTML) to carry RDF information in attributes (see 18 [http://www.w3.org/TR/xhtml-rdfa-primer/]). To be specific, RDFa defines attributes such as about, property, or datatype – and thus allows one to make RDF statements.

Although only defining a basic set of attributes (there are all in all 10 attributes), RDFa is very powerful. With RDFa, you cannot only make statements about the current resource, but also about other resources (“The metadata features available in XHTML only allow information to be expressed about the document itself. RDFa allows the document to contain metadata about other documents and resources.” [http://www.w3.org/TR/rdfa-syntax/#sec_2.2.]).

A very simple and intuitive example for RDFa is the following:

```html
<div xmlns:foaf="http://xmlns.com/foaf/0.1/" about="#me" rel="foaf:knows">
  <ul>
    <li typeof="foaf:Person">
      <span property="foaf:givenName">Felix</span>
      <span property="foaf:familyName">Müller</span>
    </li>
  </ul>
</div>
```

Like RDF in general, RDFa encourages/enforces the reuse of existing, standard vocabularies. The example above refers to friend-of-a-friend (foaf).

### 4.4.4 HTML Microdata

“The HTML microdata mechanism allows machine-readable data to be embedded in HTML documents in an easy-to-write manner ... At a high level, microdata consists of a group of name-value pairs. The groups are called items, and each name-value pair is a property. Items and properties are represented by regular elements.” (Quote from [14]).
My name is Neil.

My band is called Four Parts Water.

I am British.

5 Use of and Support for Standardized Metadata in Mainstream Applications and Environments

RDF already has a place in the following environments amongst others

- All Semantic Web technologies
- The DITA Learning/Training spec includes a subset of elements from the IEEE Learning Object Metadata specification within its metadata, and there is an RDF binding available for that.
- Adobe’s Extensible Metadata Platform (XMP) is an example of technology that allows metadata about a file to be embedded into the file itself. XMP uses RDF/XML as the basis for its metadata representation. A number of Adobe products already support XMP.
- RSS 1.0 (“RDF Site Summary”) is an RDF vocabulary that provides a lightweight, yet powerful way of describing information for timely, large-scale distribution and reuse.
- Open Document Format (ODF; Oasis/SUN, Open Office)

Google provides Web masters with technology that allows them to control which snippets from a text are shown in a result list. At the heart of the technology there are three encodings: the microformats, RDF, and HTML Microdata mentioned above. These encodings are used, for example, to capture information like that on business cards – vCard is used for this purpose.

6 Case Study: Better With Than Without Standardized Metadata

It is fairly easy to outline the necessary steps for using standardized rather than proprietary metadata. Imagine that you have a need for metadata such as the following in your files:

<sxsd:metadata xmlns:sxsd="urn:x-slt:xliff12:metadata:1.0"
    xmlns="urn:x-slt:tsmetadata:1.0">
    <object-name>DasExecutionCompView.wdview.xlf</object-name>
    <collection>App Package 47</collection>
</sxsd:metadata>
In order to turn to standardized metadata, you only need to take two small steps:

1. Switch from a proprietary general encoding scheme to a generalized one such as RDF
2. Switch from proprietary encoding for data categories and values to standardized ones such as Dublin Core (see [16])

After having taken these steps, you are already in the world of the semantic web ... A world where applications help you to manage your metadata and to make use of it – a world where you have less hassle to explain or teach what your metadata is all about (see [17] for related issues such as metadata provenance).

Like RDF in general, RDFa encourages/enforces the reuse of existing, standard vocabularies. The example above refers to friend-of-a-friend (foaf).

7 Outlook

The recently started Thematic Network, which is funded by the European commission, may shed additional light on the current state of metadata for content (see [19]).

8 References

Acknowledgements

The authors would like to thank Prof. Dr. Felix Sasaki for commenting on draft versions of this article.