*Language Learning & Technology*
http://llt.msu.edu/vol3num1/emerging/index.html

July 1999, Volume 3, Number 1
pp. 12-15 online
(page numbers in PDF differ and should not be used for reference)

**EMERGING TECHNOLOGIES**

# Web Metadata: More Efficient Resource Cataloging and Retrieving

**Bob Godwin-Jones**
Virginia Comonwealth University

**Contents:**



## What a Tangled Web We've Woven

For many second language teachers, Web sites have become a major source for authentic language materials. Despite English's role as the lingua franca of the Web, there continues to be tremendous growth in the number of Web pages in virtually all languages. Students are finding the Web to be an invaluable storehouse of timely data; URL citations are supplementing, if not supplanting, traditional print sources. Yet well-know problems with the use of Web resources remain. Wonderful sites here today are gone tomorrow, disappearing with no forwarding address. Topical information--the strength of the Web--is by its nature short-lived; contextualizing and annotating texts or materials from the Web can become a non-stop process, as today's news becomes tomorrow's broken link.

An equally troublesome issue lies in the difficulty of locating desirable Web sites in the first place. Out of the hundreds or thousands of hits from a search engine, how can one tell which are really applicable for the topics I'm researching, which are appropriate in terms of language level, media, or format, and -- most critically -- which are most likely to provide accurate and reliable information. Currently, from the long list typically retrieved from the search process there is little more to go on than page titles and a snippet of text. While there is little one can do about the ephemeral volatility of the Web (except to request permission to save local copies of valued pages manually or using software like WebWhacker), there is the promise of help on the horizon for more efficient searching and retrieving of Web documents.

## Metadata to the Rescue

Currently, search engines such as Hotbot or AltaVista provide for options for narrowing searches in the hopes of finding appropriate sites. However, those options which are available, such as language, are often not reliable. The search engines are not to blame. The problem is that most Web pages provide formatted content only and not information about the contents of the page (i.e., *metadata*) such as language. While "intelligent agents" can roam the Internet searching for appropriate information sources, their usefulness is limited by the forced reliance on full text analysis rather than indexed data. A helpful tool for language teachers (and others) currently under development, KWICFinder by William Fletcher, does a good job of maximizing the benefit of current Web seach mechanisms by offering more sophisticated options and automated production of a Key Word in Context abstract of each match.

Actually, there has been for some time a means available for including basic metadata information about a Web page by incorporating a <meta> tag in the HTML source, which is typically used to provide a short description and/or a list of keywords. The tag is placed in the header of the document and has the following syntax:

<META NAME = "keywords" CONTENT = "fairy tale, Grimm Brothers, child protagonist,

brother and sister, woodcutter, stepmother, forest, poverty, child neglect, witch, cannibalism">

If authors routinely included keywords with their pages, this would not only allow search engines to catalog more efficiently, it would also allow for local search options. The subject/keyword search for 19th-century German stories, for example, uses keyword descriptors to enable searches which can combine terms in a variety of ways and languages.

While helpful, keyword and brief descriptions are not enough to determine applicability, appropriateness, and reliability. What's needed is a standardized cataloging method for Web documents: a means for Web authors to include more detailed information about their Web pages. If Web documents contained rich metadata, search engines could provide many more options. In particular, Web documents appropriate for pedagogical use could provide optional information about the resource such as target age/grade, learning level, or pedagogical approach.

**The Dublin Core**

A significant step in this direction has been the development, wide-spread acceptance, and growing international use of a standard set of metadata categories known as the Dublin Core. This is a common set of 15 categories established by an international working group spearheaded by librarians. Using the Dublin Core allows for extensions of the metadata for a Web page to include significant additional information, stored in a straightforward way. An example of (selected) Dublin Core metadata:

```
<META NAME = "DC.title" CONTENT="Hänsel und Gretel">
<META NAME = "DC.creator" CONTENT="Grimm Brothers">
<META NAME = "DC.language" CONTENT="de">
<META NAME = "DC.subject" CONTENT = "fairy tale, Grimm Brothers, child protagonist,
brother and sister, woodcutter, stepmother, forest, poverty, child neglect, witch, cannibalism">
<META NAME = "DC.format" CONTENT="text/html; images/gif; audio/ra">
<META NAME = "DC.identifier"
CONTENT="http://www.vcu.edu/hasweb/for/grimm/haensel.html">
<META NAME = "DC.source" CONTENT="1857 edition of Haus- und Kindermärchen der
Brüder Grimm">
<META NAME = "DC.relation" CONTENT="Grimm fairy tale number 15">
<META NAME = "DC.rights" CONTENT="Free use for educational purposes">
```

The "DC" appended to the metadata name indicates that the term is being used as defined in the Dublin Core. The standard set of DC metatags are beginning to be widely used, including in resources for language teachers, such as the ARTFL collection of French texts (metadata example). It is also possible to extend the tag set to include more specifically pedagogical information, as has been done for the "Gem" tags used in the AskEric lesson plans (metadata example). To date, the Dublin Core element set has been translated into eighteen different languages, with more on the way. However, implementation of multi-lingual metadata is a complex issue which has yet to be fully resolved.

One issue involved in using a rich array of metadata is the tediousness of hand-coding the information. Some help for basic meta tags is provided in HTML editors such as FrontPage or Dreamweaver. For extended metadata, some metadata creation tools are available such as a metadata template for use with WordPerfect. An on-line tool for generating Dublin Core metadata is available from the Nordic Metadata Project in both full and minimalist versions. Given the growing acceptance of the crucial importance of metadata for the Web, more help is likely forthcoming in commercial software for creation of metadata, using the Dublin Core, as well as other medadata schemes.

**The Instructional Management System**

One of the enhancements to the Dublin Core which holds considerable promise for the educational community is the Educause-sponsored Instructional Management System (IMS) Project. Building on a base of Dublin Core elements, IMS adds a set of categories for pedagogical use. Sample of selected additional metadata tags (IMS version 0.4):

```
<META NAME="IMS.Interactivity" CONTENT="low (reading) to high (quizzes)">
<META NAME="IMS.Identifier" CONTENT="http://www.vcu.edu/for/menu.html">
<META NAME="IMS.LearningLevel" CONTENT="6-99:1"> <-- indicates age range, from 6 to 99-->
<META NAME="IMS.Objectives" CONTENT="Reading comprehension of German literary texts">
<META NAME="IMS.Pedagogy" CONTENT="Expository">
<META NAME="IMS.Platform.RequiredSoftware.Description" CONTENT="Web browser supporting JavaScript 1.0 or ECMAScript 1.0">
<META NAME="IMS.Platform.RequiredHardware.Description" CONTENT="Capability to run scripting enabled Web browser">
<META NAME="IMS.Prerequisites" CONTENT="Basic reading knowledge of German (novice to intermediate)">
<META NAME="IMS.Presentation" CONTENT="Text, images, sound, self-correcting quizzes">
```

Metadata is just one of the aspects of the IMS project; its main goal is to establish standards ways to find, use, and exchange on-line learning materials.

Although IMS metadata can be incorporated into HTML, as in the example above, the project assumes use of a recently adopted Web standard for metadata, the Resource Description Framework (RDF). RDF is designed to be a powerful and flexible means of describing metadata using the next-generation Web encoding language XML, or eXtensible Markup Language. XML is a greatly enhanced relative of HTML which was created to overcome its limitations in a number of areas. XML is not yet widely implemented by Web browsers; Internet Explorer 5.0 offers basic support. Assuming knowledge of HTML, RDF tagging (which uses XML syntax) should not be difficult to understand or use (example of IMS tags in RDF and Dublin Core examples in RDF). Using RDF it is possible not only to add more metadata categories, but to add entirely new tags. RDF/XML adds the concept of "namespace" to include a linked "schema" (optimally machine-readable) for understanding and parsing the tag set used. This ensures that use of tags from multiple schemas for the same document is possible, allowing for both "mix and match" and creation of document-specfic tags (assuming an accessible declarative schema).

**Outlook**

One of the benefits of using RDF/XML is its reliance on Unicode, a character-encoding system which supports alternative character sets. Clearly, this is of benefit to the language learning community. However, support for Unicode must also be provided in operating systems and browsers. Furthermore, for any of the metadata schemas to provide help in searching and retrieving, they need to be used widely. Only then will search engines begin to offer more options based on increased metadata information. Much of the information on the Web is of an ephemeral nature and does not need to be cataloged any better than is the case currently. However, with its growing use in instruction, the Web has also become the repository of valuable, long-term learning resources. For such Web sites, it makes great sense to include rich metadata, to allow them to be found and used more efficiently.

**Resource List**

**Metadata: Information and Tools**

- Metadata Information links and tools list from Oklahoma State University

- Making the most of meta tags Tips from cnet
- Meta tags From HTML tips in Builder.com
- Nordic Metadata Project Information and Tools for using metadata (Dublin Core)

**Metadata Schemas**

- Dublin Core Metadata Initiative Official home page
- The State of the Dublin Core Metadata Initiative - April 1999 By Stuart Weibel
- User Guidelines for Dublin Core Creation From the Nordic Metadata Project
- Dublin Core Examples in RDF Extensive list in different formats
- Using IMS Metadata Overview of IMS approach
- IMS Metadata Dictionary Complete set of metadata tag set in IMS
- GEM: The Gateway to Educational Materials Gem tag extensions to Dublin Core

**Resource Description Framework**

- Frequently Asked Questions about RDF From W3.org
- What is...RDF From Ariadne
- Resource Description Framework Official specs
- RDF and Metadata From Tim Bray, one of the XML architects
- XML and the Second-Generation Web Article from Scientific American

*All links validated on June 23, 1999*