Site Metadata on the Web

Erik Wilde
School of Information
UC Berkeley
dret@berkeley.edu

ABSTRACT

The navigation structure of Web sites can be regarded as metadata that can be used for interesting applications in *User Interface (UI)* design and *Human-Computer Interaction (HCI)*, as well as for *Information Retrieval (IR)* tasks. However, there currently is no established format for site metadata, which makes it hard for Web sites to publish their structure in a machine-readable way, which could then be used by HCI and/or IR applications. We propose a model and a format for site metadata that is built on top of an existing format and thus could be deployed with little overhead by publishers as well as consumers. Making site metadata available as machine-readable data can be used for improving user interfaces (informing user agents about the context of the page they are displaying) and better information retrieval (allowing search engines to use sitemap information for better ranking and display of the results).

Categories and Subject Descriptors

H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia—*Navigation*

General Terms

Design, Standardization

1. INTRODUCTION

The URI structure of a Web site (often referred to as a *site map*) is an important aid for navigating the content of a site. Many Web sites make the site structure available through *site navigation*, often implemented visually as horizontal and/or vertical menu bars, or less frequently also through a dedicated Web page representing the site map, listing all of the site's available pages. However, there currently is no machine-readable format for this information, which we call "site metadata." This paper discusses the challenges and the potential benefits of such a format, and proposes a way to augment the *sitemaps.org* format with site metadata.

Site Metadata on the one hand greatly improves the interaction of humans with a site, because many tasks on a site require accessing more than one page on the site. On the other hand, even though explicit navigation often is provided through Web page design, IR can be used to algorithmically infer site metadata for tasks other than direct user interaction with a Web site. Google's search results, for example, occasionally include a small "site map" (called "sitelinks") for highly ranked search results (Figure 1 shows an example). Allowing Web sites to publish site map data in a machine-readable way thus could augment HCI as well as IR tasks regarding Web page structures.

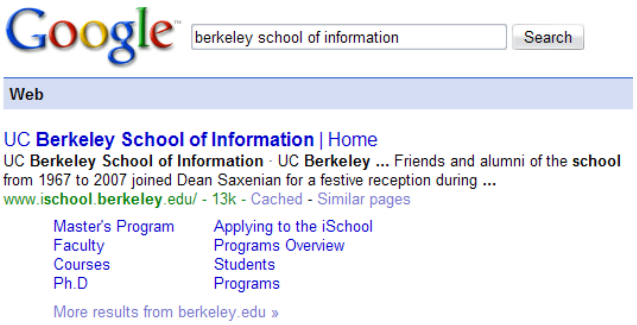

Figure 1: Algorithmically Determined Sitelinks

Sections 2 and 3 give a short overview of the possible benefits of explicit site metadata on the Web, and Section 4 summarizes this potential. Section 5 then describes the data model that we have defined so far, and Section 6 then makes a proposal for augmenting an already existing format with site metadata based on that model.

2. NAVIGATION SUPPORT FOR HUMANS

While usability and accessibility are important subjects in the context of individual Web pages, usability and accessibility of Web sites (i.e., a structured and interconnected set of Web pages) is a topic that is discussed less frequently. HTML itself has the ability to include `<link>` elements in the document head which can express a number of document relationships between HTML documents, but the available relationship types indicate that the focus of this feature is to support single logical documents which are represented by more than one HTML document. Furthermore, most browsers do not support this HTML feature.[1] And since it is defined in HTML itself, it cannot be used easily to cover HTML as well as non-HTML media types.

[1]Only Opera natively support navigating `<link>` elements; for Firefox and IE there are extensions supporting this functionality.

Second Workshop on
Human-Computer Interaction and Information Retrieval (HCIR 2008)
October 23, 2008, Redmond, WA

The *Web Content Accessibility Guidelines (WCAG)* [1] also do not discuss in great detail how to make the navigational structure of Web sites accessible, they mainly focus on making document structures accessible. WCAG technique G62 recommends to provide a site map, but talks of that site map as an HTML page, which means that the sitemap is not machine-understandable.

On today's Web, the navigational structure of a Web site is usually represented visually by common "design patterns" for Web-based user interfaces, and in most cases the actual data is provided by a *Content Management System (CMS)* on the back end, which propagates the design pattern with site data.[2] Even though there is a small number of these design patterns describing the vast majority of Web sites, this still leaves navigational structures in the realm of Web information not described in a machine-understandable way.

There is only little research about how better orientation within a Web site could help users to better navigate and utilize the site. One study conducted by Danielson [3] suggests that constantly visible site maps do have a positive effect on how people can utilize a site in terms of more effective navigation and a better overview of the available resources on a site.

3. SITE METADATA FOR MACHINES

The *sitemaps.org* format has been invented by Google and now is being jointly developed by a number of major search engines. Despite its name it is not a site map, it is simply a set of URIs which can be provided by Web masters to provide search engine crawlers with a set of URIs they might want to crawl. The intent of the sitemaps.org format is not to provide information about a site's structure, but only to provide information about the accessible URIs.

In addition to the basic text format (a list of URIs, one per line), there also is an XML format. This format allows Web masters to specify additional information for individual resources, the last modified date, the expected change frequency, and a priority. Crawlers are free in how they use that information to control the crawling process, and most crawlers will use internal heuristics to decide how much they rely on this additional information.

4. POTENTIAL OF SITE METADATA

While the goals of using site metadata for supporting humans (Section 2) or machines (Section 3) are different, both goals could be accomplished by using the same metadata. The following list is likely to be incomplete, but lists some of the areas where site metadata could be used to provide better implementations of HCI- or IR-related tasks.

- *Unified Navigation:* If site metadata were available to browsers, they could provide unified controls for navigating sites, making it unnecessary for users to adjust to the various ways in which sites implement site navigation.[3] Browser navigation not necessarily has to completely replace the embedded navigation, but a browser could provide additional features to better guide users through a site.

- *Accessibility:* Even though Web page accessibility is a popular topic, this is much less true for Web site accessibility, i.e. the ability for users to navigate a Web site without having to search through embedded navigation controls. Site metadata can greatly improve site accessibility, because it allows browsers to explicitly provide navigation features, without the need to "find" the embedded navigation controls of Web pages.

- *Crawling:* The sitemaps.org format already has most important information that allows crawlers to adjust their strategy to a site's resources. However, more navigational data (such as the various "levels of hierarchy" on a Web site) might also be useful input for determining crawl sequences.

- *Ranking:* Based on a site's structure, ranking can be better informed because hits could be ranked according to specificity (a hit in a page "lower" in the hierarchy is likely to be more specific, whereas a hit in a "higher" page is more likely to be on an overview page). As for crawling, ranking could use this information as additional input to already existing strategies and algorithms.

- *Search Result Clustering:* In a way similar to that shown in Figure 1, site metadata could be use to cluster search results according to a site's structure, or to show where in a site's structure a hit occurred. Again, site metadata would most likely only be one input into such a feature.

While the HCI-oriented tasks (unified navigation and accessibility) make use of the site metadata on a per-site basis, the IR-oriented tasks are based on using the aggregated site metadata of a large number of sites. As usual, Web masters might be tempted to try to game algorithms by supplying site metadata that should improve a sites visibility in a search engine. Site metadata in such a scenario might become just one more factor in what is often referred to as *Search Engine Optimization (SEO)*, which comprises a number of legitimate and useful ways to improve a sites usability for search engines, but sometimes also includes strategies which run against the intentions of search engine providers and have to be detected and compensated for.

Machine support by site metadata is already partially supported by the sitemaps.org format, but there is only very little support for site navigation for humans. One notable exception is the *Standard-Navigation* (formerly known as *Standard-Sitemap*) Firefox add-on shown in Figure 2. It uses a custom XML format which supporting Web sites are supposed to supply, and then uses that data in a browser sidebar. The add-on even has the option to hide the embedded navigation on a Web page (which has to be marked up with specific HTML code), so that navigation controls will only be displayed in the sidebar, and not also as embedded controls in the Web page.[4]

[2]The *Web Modeling Language (WebML)* [2] supports an elaborate model of how to describe datasets and Web interfaces for them.

[3]In a simple way this already is possible if a site uses a well-design URI structure, where the navigation hierarchy is reflected in the URI hierarchy. In this case, simple browser extensions such as the Firefox *Go Up* extension allow users to go up one level on the site by using a browser button.

[4]Browsers not using the add-on will not recognize the special markup for the embedded navigation controls and will therefore not hide them.

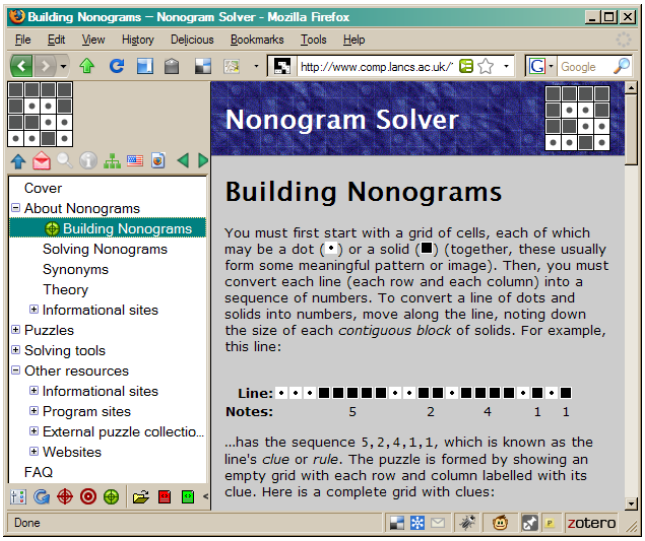

Figure 2: Standard-Navigation Sidebar

The approach of this add-on is to completely remove embedded navigation from Web pages, so that all navigation can be controlled through the sidebar. It is at least questionable whether this is a goal that will be shared by a substantial share of Web designers. We believe that it's more useful to think of browser-based controls for navigation as supplemental features for whatever the Web designers choose to embed within their Web pages. It then remains to be seen (and tested) how useful a more unified towards navigation actually is, and how much there will be a general trend towards outsourcing navigation controls from Web page content to browser controls.

5. SITE METADATA DESIGN

At first sight, the design of a site metadata model might seem almost trivial. A simple sitemap usually can be modeled as a tree representing the hierarchical structure of a Web site. For very simple sites, this model might be complete or at least sufficient, but when looking at Web sites, it quickly becomes apparent that site metadata can be much more complex in structure than just a simple tree with one kind of relation between resources. The following issues illustrate some of the potential complications of real-world site metadata:

- *Sets vs. Sequences:* While some sites might want to model their hierarchical structures as sets, other might want to model them as sequences. Moreover, in the case of sequences, the actual sequence can sometimes depend on factors which vary with resource variants (such as page titles, which will vary by language).
- *Variants:* Resources (navigation targets in the site structure) might exist in different variants, and the variants might use different dimensions of variation. Typical examples are languages (multilingual Web sites) and media types (resources might be available as HTML and PDF). While all of these resources are equivalent on a conceptual level, concrete clients will most likely only use one of them, depending on user preferences and client capabilities.
- *Versioning:* Versions can be regarded as a special type of variant because they have the built-in assumption that there is a chronological sequence of versions. Complex version models might be non-linear, for example when a page is split into multiple pages and thus the versioning structure becomes a tree (in general, versioning graphs are directed acyclic graphs).
- *Non-Tree Structures:* While many sites indeed are tree structured, there are also sites where the navigation structure "reuses" pages in various locations, so that the effective navigation structure can either be regarded as a tree with duplicate pages in it, or as a directed acyclic graph.
- *Dynamic Structures:* Advanced Web sites sometimes customize navigation structures based on criteria such as a personal profile, histories, preferences, and popularity of pages with recent visitors. With these sites, site metadata is determined by many different factors and the navigation aspects of site metadata have to be specifically determined for each client. However, there is no reason why the dynamic generation of embedded navigation controls could not also drive the generation of site metadata.
- *URI-less Navigation:* While many sites do have individual URIs for different pages in their navigation structure, there are also sites which do not have URIs for these pages. The two most common cases for this are frame-based sites, and sites where embedded code (popular examples are Ajax and Flash) handles navigation without reloading pages.

The above list of issues probably supports the way in which most Web sites would want to publish their site metadata, but it might also exclude some sites which have even more sophisticated models of their site's structure. Also, because the *Hypertext Transfer Protocol (HTTP)* [4] provides functionality beyond the simply retrieval of resources, some of the complexity of the above list could be deferred to HTTP.

For example, the detection of variants could be deferred to HTTP content negotiation, which allows Web servers to advertise that a resource is available in different variants. But many Web sites do not use HTTP-based language selection, they simply provide different resources without any machine-readable information about their conceptual equivalence. If a site metadata model should also support these sites, then variants must be included in the model.

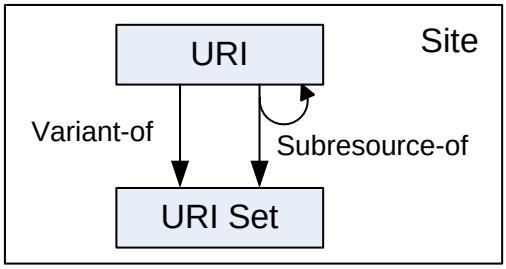

Figure 3: Site Metadata Model

Based on these considerations, we have designed the site metadata model shown in Figure 3. We decided to not

include versioning information, because it complicates the data model, and there were only few use cases where version information was a required component of the data model.

A site is escribed by a number of URIs and URI sets. possible relationships between URIs are hierarchy levels (expressed by the *subresource-of* relation), and if a resource is represented by multiple variants, a URI set is used (associated by the *variant-of* relation). URIs are associated with URI sets by specifying the dimension(s) of variation and the respective value(s). Optionally, URIs and URI sets can have position values, which are used to determine a sequence of resources, if sites want to use sequences rather than sets.

6. DATA FORMAT

The data model for site metadata described in Section 5 can be represented in different ways. We identified the following three methods as the most promising candidates for representing site metadata:

- *Dedicated XML Format:* It is possible to create a entirely new data format, and XML is a good choice because it has become the most widely supported foundation for the open exchange of structured data.
- *RDF:* Since site metadata is not content but metadata about content, it might be regarded as something that should be represented using *Semantic Web* [5] technologies, using the *Resource Description Framework (RDF)* as its model and syntax.
- *Extension of existing XML Format:* Instead of starting from scratch, an existing format could be extended. The most promising candidate is the sitemaps.org format.

We decided that the most promising way is to extend the sitemaps.org format, which seems to have gained some popularity (even though we could not find any data about that). Unfortunately, the extensibility (as well as the format as a whole) is very poorly documented, which makes it impossible to understand what kind of extensions the format allows. This is relevant because existing implementations might break or misinterpret data if they have built-in assumptions about the data format which have not been documented in the format itself, and which are violated by an extension.[5]

Based on the limited information about extensibility, the current format could be updated as follows: URI sets are represented by the `urlset` element, which is allowed as a child of the `urlset` document element. The `url` and `urlset` elements have an optional `id` attribute, and a subresource is identified by an `parent` attribute which specified the ID of the higher-level resource. Optionally, a subresource can carry a `pos`ition attribute for specifying a sequence of subresources rather than a set. Variants use a `variant` element as a child of the `url` element, and this element has attributes for the `urlset` (it is a variant of this URI set), the `dim`ension (such as language or media type), and the `value` for that dimension (such as a concrete language).

[5]Google claims that a well-defined extensibility model is under development, but in contrast to the data model, which is openly available and CC-licensed, the development process is closed and no information about the extensibility model is currently available.

```
<urlset xmlns="http://www.sitemaps.org/xmlns/1">
  <url id="home">
    <loc>http://www.example.com/</loc>
  </url>
  <url id="contact" parent="home" pos="1">
    <loc>http://www.example.com/contact</loc>
  </url>
  <urlset id="faq" parent="home" pos="2"/>
  <url>
    <loc>http://www.example.com/faq,en</loc>
    <variant urlset="faq" dim="lang" value="en"/>
  </url>
  <url>
    <loc>http://www.example.com/faq,de</loc>
    <variant urlset="faq" dim="lang" value="de"/>
  </url>
</urlset>
```

While the main structure of the sitemaps.org format remains the same, the addition of attributes and a new child element type to the document element might be something that is considered out of scope for extensions. If that is the case, the above example can also be represented using only new child elements of the `url` element. This kind of representation is even more verbose and less elegant, but the most important issue is that the data model (Section 5) can be represented in an extension of the sitemaps.org syntax.

7. CONCLUSIONS

In this paper, we present our work towards making site metadata available on the Web. The current sitemaps.org format has gained some popularity and is useful for the IR-oriented tasks regarding site metadata, but it ignores the benefits that are possible from an HCI perspective towards better site navigation for users. Our future work is twofold: When the revised sitemaps.org format is released, we will have a well-defined set of rules for this data format. On the other hand, we want to explore the possibilities and limitations of navigation support driven by site metadata. This exploration of the usefulness of site metadata will inform our final definition of the sitemaps.org extension; it is of course possible that our current data model will have to be revised.

8. REFERENCES

[1] Ben Caldwell, Michael Cooper, Loretta Guarino Reid, and Gregg Vanderheiden. Web Content Accessibility Guidelines 2.0. World Wide Web Consortium, Candidate Recommendation CR-WCAG20-20080430, April 2008.

[2] Stefano Ceri, Piero Fraternali, and Maristella Matera. Conceptual Modeling of Data-Intensive Web Applications. *IEEE Internet Computing*, 6(4):20–30, 2002.

[3] David R. Danielson. Web Navigation and the Behavioral Effects of Constantly Visible Site Maps. *Interacting with Computers*, 14(5):601–618, October 2002.

[4] Roy Thomas Fielding, Jim Gettys, Jeffrey C. Mogul, Henrik Frystyk Nielsen, Larry Masinter, Paul J. Leach, and Tim Berners-Lee. Hypertext Transfer Protocol — HTTP/1.1. Internet RFC 2616, June 1999.

[5] Nigel Shadbolt, Tim Berners-Lee, and Wendy Hall. The Semantic Web Revisited. *IEEE Intelligent Systems*, 21(3):96–101, March 2006.