Mako Hill

Stephen Harris

Origins of Reading

8 May 2002

Parsing Text into Data: Bibliographies

Markup languages like SGML showcase some of the ways that the treatment of texts as data through the use of descriptive markup can provide power and flexibility over the use of traditional procedural markup methods. DocBook SGML is simple and well defined.[1] It can not allow for any ambiguity; text enclosed within an SGML tag is *clearly* demarcated as being of a particular type of data. When a software renders DocBook source into an HTML or PDF document, the software must first parse the text into data before it can be translated into the new format and redisplayed. When a human reads the same text on a printed page, their first step is also to parse and classify the text. While reading is certainly more than parsing, this analysis raises questions about the way that computer mediated literature is read. As computers' success in parsing text and language increases, technology is able to engage in one readers more difficult and important tasks.

Parsing texts into data is one of the most invisible but important acts of reading: western readers immediately know that the lines at the top right corner of a missil are an address; they know that the name centered and following "Sincerely," is the author; they know that footnotes provide extraneous information or assides and that "(90)" is a reference to a page number. No where is this type of implicit data classification more important than in bibliographies.

Bibliographic standards use elaborate systems of syntax and typographic markup to pack massive amounts of contextual data into a very compact spaces. These systems aim to balance the need for easy readability, small space, and the need to eliminate ambiguity. Take for example, the following simple bibliographic entry in MLA format:

Lessig, Lawrence. Code and Other Laws of Cyberspace. New York: Basic Books, 2000.

Even those unfamiliar with the nuances of MLA bibliographic syntax can safely parse data communicated in the entry. At first glance, readers separate the the entry into three major parts separated by periods: "Authors Name. Title of the book. Publication information." With out too much difficulty, these readers proceed to break down these components. Although this example is simple, the MLA standard is nuanced

[1] DocBook is defined by the International standards group OASIS. Before processing DocBook, you must specify the version of the standard you are using. DocBook 4.1 uses a limited and well-documented subset of tags while 3.1 uses a slightly different set.

and capacious enough to allow authors to reference almost any type of source unambiguously. In fact, almost half of the 300 page MLA Handbook for Writers of Research Papers is devoted to documenting the creation of these reference entries.

The result is a system that is compact and unambiguous. With the handbook in hand, it is relatively easy for a reader to parse a complex bibliographic line into data pointing to a single book or work. Readers must use context, experience, and knowledge of the MLA format but the methods for parsing the text are well defined. With a well designed bibliographic system like MLA, there will never be any ambiguity.

Like MLA, DocBook SGML includes the ability to classify bibliographic information. The software renderer will take liberties to rearrange a "raw" entry into the correct order according to rendering instructions. The following is an example of a bibliography entry in DocBook:

```
<biblioentry><biblioset>
<author><surname>Lessig</surname><firstname>Lawrence</firstname></author>
<title>Code and Other Laws of Cyberspace</title>
<publisher><publishername>Basic Books</publishername> </publisher>
<pubdate>2000</pubdate>
</biblioset></biblioentry>
```

In DocBook, each piece of information is placed within tags (some tags at the top of hierarchies of subtags). With a basic knowledge of DocBook and SGML, a reader can easily parse this block of DocBook and understand that it is referring to the same book referenced in MLA format above; both entries contains the same data. If this SGML were rendered into MLA format, it would reproduce the first example. The metadata in the DocBook example is simply a more explicit version of metadata that becomes largely contextual in the rendered MLA form.

Bibliographies provide an example of only one way that markup acts to classify text as data. While the bibliographic subset of DocBook is written for electronic readers and while MLA is aimed at humans, both standards seek to place text within unambiguous contexts to define text as data. When authors or designers lay out the text on a page, they classify text as data and encode super-textual data. While reading, this information is unpacked. By parsing, understanding, and manipulating this super-textual data, DocBook rendering software takes on part of the role of reader and author and confuses traditional definitions of these roles.