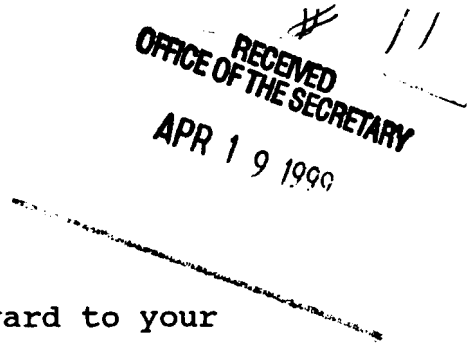


Author: glen gray <vcact00f@csun.edu> at Internet
Date: 4/13/99 12:22 PM
Priority: Normal
TO: RULE-COMMENTS at OS1
Subject: File No. S7-9-99

Dear Sirs:

I understand that you are accepting comments with regard to your modernization efforts of the EDGAR system.

I ask that you consider allowing filers to file UNOFFICIAL versions of their filings in an XML format similar to your allowing UNOFFICIAL versions of filings in a PDF format. The purpose of this is to allow the financial community to experiment with the possibilities of XML and expose the public to XML.

My colleagues and I have been conducting research regarding locating financial information on the Web and we have come to one conclusion: it is very problematic. XML could significantly improve the search process--paticularly automated searches. I have also attached a recent paper that we wrote on the topic. If the SEC accept XML for EDGAR, then XML be used more on corporate Web sites.

The accounting industry is working on setting standards for XML based financial reporting. Your action of allowing UNOFFICIAL filings will greatly help XML standards for accounting evolve.

Selecting XML as an official filing format for your current modernization would have been an incorrect choice as the technology is still maturing. However, the rapid adoption of XML clearly signals that XML is probably going to be in the future of the EDGAR system.

I appreciate your consideration.

Best regards,
Glen L. Gray, PhD, CPA

**********
Glen L. Gray, PhD, CPA * (818) 677-3948 (Voice)
Department of Accounting & MIS * (818) 677-4903 (FAX)
California State Univ. Northridge * (818) 677-2461 (Department office)
18111 Nordhoff Street * glen.gray@csun.edu
Northridge, CA 91330-8372 * http://www.csun.edu/~vcact00f/index.htm
**********

#13

# THE ELECTRONIC DISSEMINATION OF ACCOUNTING INFORMATION - RESOURCE DISCOVERY, PROCESSING AND ANALYSIS

57-9-99

*A Paper Submitted for Consideration at the 1999 EAA Congress*


Roger Debreceny

Nanyang Technological University

Nanyang Avenue

Singapore 639613

rogerd@netbox.com

Glen L. Gray

California State University at Northridge

glen.gray@csun.edu

Tony Barry

Australian National University

tony@info.anu.edu.au

# ELECTRONIC DISSEMINATION OF ACCOUNTING INFORMATION - RESOURCE DISCOVERY, PROCESSING AND ANALYSIS

## ABSTRACT

Corporations have rapidly adopted the Internet and in particular the World Wide Web (the "Web") protocol to communicate accounting information to stakeholders. Approximately 80% of major US corporations are making some type of financial disclosure on the World Wide Web. There is a clear demand for accurate, reliable and prompt delivery of financial information as an input to stakeholder decision models. Should humans or by intelligent software agents be able to retrieve financial information on the Web with high levels of accuracy, considerable opportunities would be opened for innovative and cost-effective analysis and use of accounting information. Accounting information on the Web, however, is inconsistently presented from corporation to corporation. It is difficult to find either by direct navigation to corporate "home pages" or by use of the major full-text search engines. The reality of the Web is that it falls far short of a reliable communications medium for accounting information. This is because there is no standard for metadata representation of accounting information on the Web that could improve the accuracy of searches. Further, accounting data points within pages on the Web cannot be parsed with any reliability.

A model for the presentation of accounting information on the Web is set out. The model uses the putative Web standards for an "eXtensible Markup Language" (XML) and a metadata container set (the "Dublin Core"). Taken together, the application of these tools to accounting information would mean that both humans and intelligent software agents could operate with a high degree of accuracy and reliability.

## KEYWORDS

Internet; World Wide Web; Communication; Metadata; Financial Reporting

# 1 INTRODUCTION[1]

The electronic dissemination of accounting and financial information via the Internet, particularly the World Wide Web (the Web), has been growing rapidly. Corporations have adopted the Web as an inexpensive and timely means to communicate financial and corporate information with stakeholders (Koreto 1997; Deller, et al. Forthcoming).

Disseminating accounting information by posting it on a corporate Web site is easy and cheap. However, it is much more difficult for humans or intelligent software agents to locate desired accounting information. In a broad sense, the Web is a giant loosely inter-linked *data warehouse* that contains an unprecedented amount of information. Just as corporations use data mining techniques to extract or discover relationships in their internal data warehouses (Adriaans, et al. 1996; Barquin and Edelstein 1997; Berson and Smith 1997), individual consumers can engage in "Web mining" (Etzioni 1996, 65) to extract or discover relationships in the financial data available on the Web. The problem is, as the Web grows, the difficulty of locating and navigating to specific information on the Web also grows (Kim and Hirtle 1995; Ciolek 1996; Berghel 1997; Chen and Rada 1996). For example, the seemingly simple task of finding a corporation's "official" Web site can be a major undertaking. The AltaVista search engine, for example, provides a list of 5,175 Web sites in response to a request for "AT&T" + "annual report". Philip Morris, the tenth largest company in the U.S., does not have an official corporate Web site, yet a search on "Philip Morris" + "annual report" results in 580 listings.

As this paper will illustrate, the underlying problems with these searches and Web mining in general are twofold. First, schemas (such as HTML meta tags) that could be used to identify or classify data on the Web are neither robust nor well structured. Secondly, financial reporting Web sites use existing schemas inconsistently, if at all.

Not until there are robust and widely, and consistently, implemented schemas can the power of the Web as an accounting information dissemination media be fully realized. This paper draws upon research in database accounting, data mining and metadata management to propose a model for the identification of accounting attributes within Web financial statements (attribute identification) and the location of

---

[1] Our thanks for assistance provided by Dr Paul Pacter of the International Accounting Standards Committee and Dr Renato Ianella of the Distributed Systems Research Centre, Brisbane and for the views of the commentator and participants at the 1998 Annual Meeting of the American Accounting Association.

accounting data on the Web (resource discovery). The remainder of the paper proceeds as follows. Section 2 discusses the potential benefits of electronic dissemination of accounting and financial information and discusses the current inhibitors to achieving those benefits. Section 3 introduces the concepts of Web data mining and metadata. Section 4 discusses alternative approaches to the description of metadata. The conclusion in Section 5 sets out an agenda for future research in this area and discusses the implications of this research for standards setting agencies and national corporate regulators.

# 2 INFORMATION NEEDS

### User Needs

With electronic dissemination of accounting information on the Web, users both in person and supported by intelligent software agents, could directly monitor the performance of corporations. At the lowest level of functionality, a user could manually navigate to the Web page of a corporation that contained the desired information. At a higher level of functionality, the user could request a particular information from a search engine (e.g., "what were the total 1998 sales for Pinault-Printemps-Redoute ?"). At the highest level of functionality would be the automation of the whole process of compiling and analyzing financial information. For example, as envisaged by commentators such as Elliott and Wallman (Elliott 1994; Wallman 1997), WebBots[2] could retrieve accounting data from Web sites to use in investment models. The WebBots could rank alternative investments or build theoretical portfolios. The agent could automatically notify investors when investment rankings or portfolio moved beyond the bounds of designated parameters. The users' own Web browsing patterns can also provide the WebBots with inputs upon which to act (Cheung, et al. 1997). The collective intelligence of the Internet can also provide agglomerated analysis of accounting information; to rank investments; to raise green or red flags; or as a tool for enhanced shareholder democracy[3]. In short, WebBots could be used to monitor and respond to the

---

2 A full discussion on intelligent software agents – known also as autobots, softbots, webbots, and robots – is beyond the scope of this paper. Reference can be made to the review in the Communications of the ACM in July 1994 (e.g. Riecken 1994; Maes 1994; Norman 1994) and to Chorafas 1997; Bradshaw 1997; Huhns and Singh 1997. An accounting perspective of WebBots is made by Baker and Witmer 1997.)

3 There is a burgeoning literature on the collective ranking and analysis of World Wide Web and other Internet resources. For an overview, see the survey published in the Communications of the ACM in March 1997 (e.g Resnick and Varian 1997; Rucker and Polanco 1997 (mining of bookmarks); Terveen, et al. 1997 (organised sharing of recommendations) and Kautz, et al. 1997 (anonymous and open social networks). See also Hal Varian's Web resources at http://sims.berkeley.edu/

financial and non-financial condition of a corporation. Unfortunately, as the following discussion demonstrates, the current state of the Web would make this WebBot development nearly impossible.

**The Current Situation**

Many corporations are now publishing accounting information on the Web. Yet, corporations publish accounting information in almost as many forms as there are corporations. Several of these formats do not lend themselves to indexing by search engines or other automated analytic tools, upon which humans and intelligent agents rely. For example, many corporations publish their financial information in PDF format, effectively cutting off the contents from search engines. Other corporations use dynamic databases to serve their Web content. Under most conditions, a search engine cannot reference information created dynamically from a corporate database[4].

Further, because many third parties as well as the corporations themselves publish financial information about corporations, search engines, which rely on full text indexing of Web contents, have very low levels of precision in discovering desired information. Cathro (1997) estimates that the precision of current search engines is less than one percent.

Finding information by following hypertext links from the corporate home page can also be difficult. The terminology used for financial reports varies widely among corporations. In addition, in many cases the boundaries between the financial statements and the rest of the corporate pages are blurred (Gray and Debreceny 1997).

Given current search problems, a WebBot would have to be *very* robust to complete an acceptable Web search. Consider the manual process that a person would follow to find annual report information for Company X:

1. Locate Company X's Web site
2. Locate the appropriate link on the Company X's Web site to the annual report
3. Locate the appropriate place in the annual report that lists the needed financial information.

---

[4] The http://www.searchenginewatch.com/ site describes the implications for search engines of alternative forms of serving of information on the World Wide Web. Web pages which are dynamically generated can be identified by a '?' in the URL. A URL of the form http://www.gm.com/cgi-bin/shareholder/sh_page.cgi?e697_01 normally indicates that it has been generated as a result of interaction with a database. The content of dynamic pages cannot be stored in a full text database such as AltaVista.

Although these three steps look straightforward in theory, the process is difficult in practice. Table 1 provides findings related to conducting the above three steps for the ten largest Fortune 500 companies in the USA.

**Table 1 Selected Search Statistics for Fortune 10 Companies**

| Fortune 500 Rank | Company | Home Page | Hits on Company Name Using Alta Vista Seat Engine | Where Home Page Appeared in First 30 Hits | Where Any Company Site First Appeared in First 30 Hits | Steps From Home Page to Annual Report | Hits on Company Name + Annual Report | Where Official Site Appeared in First 30 Hits | Comments |
|---|---|---|---|---|---|---|---|---|---|
| 1 | General Motors | www.gm.com | 5,977 | N/I [1] | N/I | 2 | 1,844 | N/I | PDF only |
| 2 | Ford Motor | www.ford.com | 23,105 | N/I | 4 | 3 | 1,309 | 20 | PDF only |
| 3 | Exxon | www.exxon.com | 2,505 | 17 | 4 | 2 | 1,413 | N/I | |
| 4 | Wal-Mart | www.wal-mart.com | 90 | N/I | N/I | 2 | 229 [5] | 5 | PDF only |
| 5 | AT&T | www.att.com | 295,237 | N/I | 1 | 3 | 5,175 | 1 [3] | HTML & PDF |
| 6 | IBM | www.ibm.com | 141,818 | N/I | 17 [6] | 2 | 2,024 | 1 | |
| 7 | General Electric | www.ge.com | 622 | N/I | N/I | 1 | 751 | 2 [7] | |
| 8 | Mobil | www.mobil.com | 1,859 | 9 | 9 | 2 | 144 | N/I | |
| 9 | Chrysler | www.chrysler.com | 10,357 | N/I | N/I | [4] | 1,787 | 23 | |
| 10 | Philip Morris | none | 17,118 | N/A [2] | N/A | N/A | 580 | N/A | |

(1) N/I = not included; i.e., Home page did not appear in first 30 hits.
(2) N/A = not applicable (no Web page for this corporation).
(3) Went to PDF page, not the first page of annual report.
(4) Chrysler's annual report is not on their Web site.
(5) Searching on "Wal-Mart Corporation" + "Annual Report" yielded only 3 hits none of which included Wal-Mart's official annual report. The 229 hits came from dropping "Corporation."
(6) An IBM Denmark site.
(7) Hits financials, not the front page of the annual report.

**Step 1--Finding the Web Site**

The first task is finding the company's Web site. Fortunately, one can easily guess the Web addresses for most of these ten companies. For example, General Motors is *www.gm.com*. However, the Web address for Dayton Hudson (Fortune #28) is *www.shop-at.com/daytons/homepage.html*. An intelligent agent would be unable to know that General Motors abbreviates to GM and thence to *www.gm.com,* but that the Ford Motor Company does not abbreviate to FMC and is at *www.ford.com* and not *www.fmc.com*. An alternative to guessing the Web site from the corporation name is to use a specialist database of corporation and brand names. For example, *RealNames*, at *www.realnames.com* attempts to provide a service that correlates corporate, trade and product names, with Web sites. A search on *RealNames* displayed all ten of the largest Fortune 500 companies, on the first page of references for relevant corporation. For guaranteed accuracy, *RealNames* requires registration of individual names by the relevant company. None of the corporations we searched had formally registered with the *RealNames* service. Users and agents cannot rely on RealNames for consistent correlation of corporate name with Web site.

One alternative to educated guesses, or lookup services such as *RealNames,* would be for the WebBot to examine the results of an existing general-purpose search engine such as Yahoo, Excite or AltaVista. The WebBot would then attempt to select what appears to be Company X's official Web address. Table 1 shows one of the immediate problems with that approach. Column 4 contains the total number of pages reported for each company name using AltaVista. For example, AltaVista listed 5,977 pages for General Motors Corporation. Also problematic is that none of the first 30 listings linked to GM's home page, as shown in the Column 6.

After doing the initial search, the WebBot faces another problem. Whereas the WebBot could easily search AltaVista's 5,977 listed pages, how would it know when it had found the "official" GM Web site? GM's Web site (home page) itself does not say that it is GM's "official" Web site. An interesting example of the official-page problem is Phillip Morris, which does not have its own official corporate Web site. The site, *www.phillipmorris.com*, is a repository of corporate information on tobacco litigation underway in the USA. Yet, AltaVista listed 17,118 pages. Even when securities regulators, such as the SEC, maintain a list of "official" corporate names, as the SEC does with its "conformed name" (CCN) there are still differences between the official and generally used name (Kambil and Ginsburg 1998, 92).

**Step 2--Finding the Annual Report**

Once Company X's Web site is located, the next step is find the annual report. As Table 1 illustrates, only one to three steps (or links) are needed to move from a Corporate home page to the first page of the annual report. However, the different terms used to describe hyperlinks creates special problems. Of the ten companies included in Table 1, General Electric was the easiest to follow because it had a hyperlink labeled "annual report". Most of the other sites had hyperlinks with labels such as "investor relations" or "company information". Clicking on these hyperlinks brings the user to another page that would usually have "annual report" as one of its hyperlinks. Of the ten companies, Mobil's Web site was among the more challenging. The home page had several broad sounding hyperlinks. The correct hyperlink was "This is Mobil." From there, it was easier. The "This is Mobil" page had a hyperlink labeled "Financial & Shareholder Information."

Chrysler's web site also posed problems. Although Chrysler does have an official Web site, it does not mention the annual report, suggesting that Chrysler has not disseminated its annual report on the Web. However, additional research established that its annual report appears on another Web site, that of Bowne Internet Solutions, which handles Chrysler's investor relations.

Most people can cope with the ambiguity caused by the differences in Web sites. We can use other clues such as context or grouping of hyperlinks, and we can apply heuristics to quickly reduce the search space (e.g., don't follow the hyperlink labeled "buy a new Chrysler"). All of the search heuristics that skilled financial statement users employ — some of which we probably could not even articulate because they are so intuitive — would be quite a challenge for the development of a WebBot.

To test if the existing search engines could help locate annual reports AltaVista was asked to search for +"[corporate name]" +"annual report." As Column 8 in Table 1 illustrates, even the AND search can result in thousands of hits. As the Column 9 shows, with a few exceptions, the "official" annual report was not included in the first thirty hits. One exception was IBM whose current annual report was the first listing.

In an interesting reversal of manual vs. automated searches, although it was very difficult to find manually Chrysler's annual report, it was included as 23 of the 1,787 pages listed by Alta Vista. Unfortunately, how would the WebBot know that this annual report *not* located on the Chrysler Web site is the *official* Chrysler annual report?

**Step 3--Finding Specific Financial Data**

As hard as it may be to automate finding Company X's annual report, consider the difficulty of automating the parsing of specific financial information from those annual reports. For example, assume an analyst posed the question: "Rank the big three US automobile manufactures by gross profit margin for 1997?" We know the formula for gross profit margin is as shown in (1):

$$GPM = \frac{\text{Revenue} - \text{Cost of Goods Sold x } 100\%}{\text{Revenue}} \qquad (1)$$

The parsing algorithm for a WebBot would have to be intelligent enough to determine what terms and numbers on the applicable Consolidated Statement of Earnings were equivalent to Revenue and Cost of Goods Sold. Generally, the actual dollar amounts are shown in multi- column financial statements with the most current year being the first number to the right of the applicable term. The comparative years are, normally, in columns further to the right of the current period. The parsing algorithm would use a combination of pattern matching (e.g., revenue numbers are usually near the top of the Statement of Earnings) and synonyms (e.g., revenue and total revenue can *potentially* mean the same thing). As stated before, adding to the search difficulty is that some companies, such as General Motors and Ford, only provide financial statements in special PDF files, which require special software to view. PDF files require a different set of algorithms to parse.

In summary, Web mining for financial information manually is an inefficient process — and the process is going to grow more inefficient as more pages are added to the Web. The ambiguities and inconsistencies that contribute to the inefficiencies make developing some form of automated WebBot extremely challenging. The primary cause of these problems is the lack of a robust standardized schema to provide a structural overlay to all the data that is already available on the Web. The next section takes a closer look at developing such a schema.

# 3 INFORMATION PROCESSING

**Web Mining vs. Data Mining**

While the Web can be viewed as one giant data warehouse that can be subject to Web mining, there are important differences between an internal corporate data warehouse and the Web. The corporate data warehouse draws from disparate data sources including one or more database management systems (Berson and Smith 1997). Consequently, a corporate data warehouse could have thousands of tables in relational DBMSs and tens of thousands of data attributes. There are many data quality issues in building

data warehouses (Gardner 1998, 60). One issue is inconsistent definitions of similar tables and attributes in the databases that make up the warehouse. Yet this problem at the corporate level is relatively insignificant compared to building a data warehouse from Web resources. A corporate data warehouse is built from, at least in theory, well understood entities and attributes. In contrast, a Web data warehouse is made up of a collection of "Document-like Objects" (DLOs) that conform to the limited HTML structure along with other objects such as graphics, sounds, video and other multimedia elements. The current generation of Web DLOs are essentially buckets of text strings, which must be parsed to extract desired data. This is no easy task. It is common, for example, for financial statements to be in the same DLO as Management's Discussion and Analysis (MD&A) or the report of the external auditor. There is no schema support within such multi-purpose DLOs.

For the corporate data warehouse, the data-mining engineer relies on his or her knowledge of the schemas of the heterogeneous databases that make up the warehouse. Increasingly, the data-mining engineer relies on a higher level of understanding about the nature of the data relationships within and between schemas -- the *metadata* of the schemas. To be able to achieve the same level of functionality with a Web data warehouse of financial statements requires mirroring the essential elements of the corporate data warehouse model. This includes enhancing of DLOs' semantic representation by the creation of the equivalent of database schemas and the declaration of metadata descriptors. In the next section we turn to the representation of the equivalent of database schemas in Web-based DLOs.

**Current Generation of Financial Reports on the Web**

The current generation of financial reports on the Web is published primarily in HTML, which is based on the Standard Generalized Markup Language (SGML). HTML has a limited and finite tagset, which give the Web browser information on the physical layout of the document. A tag such as a first level heading (<h1></h1>) instructs the browser to display the paragraph as left aligned and in a relatively large font. An HTML document has only a very limited level of semantic representation since it has a limited number of tags for the metadata representation of the document. Semantic information can be derived from the physical structure of the HTML DLO. It may be reasonable for a WebBot, which is digesting an HTML DLO, to assume that a first level heading tag is the title of the document and that second level headings are section headings. These assumptions do not always hold, however, because Web page developers are free to, and do, use tags in any way they desire.

The representation of accounting information within a Web DLO needs attribute identification at a level of functionality approaching that of a formal database schema or object definition. The tagset within

HTML is limited and finite. Jon Bosak of Sun Microsystems notes that HTML is not extensible – it does not allow the creation of new tags for specialist tasks. HTML provides no structure, as might be needed for processing data. Finally, and perhaps most importantly, HTML cannot be validated (Bosak 1998).

The W3 Consortium standard for a new "Extensible Markup Language" (XML)[5] provides the means of implementation of an attribute identification scheme (Bray 1997; Khare and Rifkin 1997). XML is a tightly defined but rich subset of SGML[6]. XML is designed to provide the semantic power of SGML while at the same time strictly limiting the choices that SGML traditionally allows (Connolly, et al. 1997; St. Laurent 1998; Goldfarb and Prescod 1998).

The most important elements of XML are:

- **"Validation"**[7] DLOs, if they conform to a XML Document Type Definition (DTD), can be validated for correctness.
- **"Structural complexity"** DLOs mapped in XML can be nested to arbitrary levels.
- **"An extensible tagset"** New tags can be defined within a DTD or within a DLO which links to, for example, a style sheet[8].

Whilst HTML is a SGML subset, practical implementations of HTML have, as is discussed above, allowed for a number of variations. XML, conversely, has strict rules of compliance. XML documents must be "well formed". If they are broken, a "fatal error" is recorded. The determination of whether a document is "well formed" is measured by reference to, or inclusion of, a XML-compliant

---

5 See *http://www.w3.org/XML/.*

6 There is an extensive literature on SGML. The definitive SGML reference is Goldfarb 1990. See also Herwijnen 1994 for a practical introduction. Goldie 1997 provides an example of the practical use of SGML in the production of technical literature. Derose 1997 presents an analysis of the linkage between SGML and XML. Connolly 1998 reproduces the articles contained in a special edition of the W3 Journal and is a good starting point for research on XML. Pragmatic advice on XML is contained in Holzner 1997; Tauber 1998 and St. Laurent 1998. Apart from these printed references, web sites on XML include *www.w3.org/XML, www.xml.com* and James Tauber's *www.xmlinfo.com*

7 This formulation is drawn from Dale Dougherty's overview of XML at http://webreview.com/97/05/16/feature/index.html

8 Work is ongoing in the W3C on the relationship between style sheets and XML. The current W3C standards for style sheets (CSS1 and CSS2) apply only to HTML. A working draft for XSL, an XML-based stylesheet language, was released in August 1998.

Document Type Definition (DTD). A part of the 1995 SEC DTD for EDGAR submissions is shown in Figure 1:

```
<!ENTITY % subm-front "(%pac-id;)?,
                       accession-number,
                       deletion?,
                       type,
                       confirming-copy?,
                       public-document-count,
                       references-429*,
                       period?,
                       items*,
                       filing-date,
                       date-of-filing-date-change?,
                       effectiveness-date?,
                       sros*,
                       group-members*" >
```

**Figure 1 Fragment of SEC EDGAR DTD**

As Figure 1 shows, this fragment of the DTD provides semantic information on the party submitting the EDGAR filing. The EDGAR DTD is both SGML-compliant and XML-compliant. The "structural complexity" of XML would allow, for example, an accounting DTD to introduce the concept of a set of financial statements. The set could be made up of elements which would be identical to the current set of accounting statements including the MD&A, Balance Sheet, Income Statement, Statement of Cash Flows and Notes.

The XML tagset is extensible. It allows the user defined creation of attributes. Part of the EDGAR attribute tagset is shown in Figure 2:

```
<accum-apprec-or-deprec>      ? number
<accumulated-gains-prior>     ? number
<accumulated-net-gains>       ? number
<accumulated-nii-current>     ? number
<accumulated-nii-prior>       ? number
<allowance-close>             ? number
<allowance-domestic>          ? number
<allowance-foreign>           ? number
<allowance-open>              ? number
<allowance-unallocated>       ? number
<allowance>                   ? number
<allowances>                  ? number
<apprec-increase-current>     ? number
<assets-other>                ? number
<average-net-assets>          ? number
<avg-debt-outstanding>        ? number
```

**Figure 2 Selected SEC EDGAR Data Attributes**

These additional tags could be used within the body of an XML-compliant document as shown in the hypothetical example in Figure 3:

```
<p>The problem loans at <fiscal-year_end>Dec-31-1997</fiscal-year_end>totaled $<loans-problem>10,000,000</loans-problem> or 2% of total loans at this date of $<loans>50,000,000</loans></p>
```

**Figure 3 Application of User-Defined Tags**

This would be represented in the Web browser as shown in Figure 4:

The problem loans at Dec-31-1997 totaled $10,000,000 or 2% of total loans at this date of $50,000,000

**Figure 4 Display of Extensible Tagset**

The browser hides the data attributes. The WebBot would, however, be able to read the XML source and accurately identify each attribute represented in the tagset. Tags can also be cascaded (or nested) in a tree structure so that a <cash-at-bank> tag can be cascaded within a <cash>, <current-assets> and <assets> hierarchy.

For consistency, an accounting information representation DTD would need to be defined by national or international accounting standards setters or securities regulators as the SEC has done for EDGAR[9]. A reference within an XML-compliant accounting report to the applicable international DTD would be of the form shown in Figure 5:

```
<?xml version="1.0"?>
<!DOCTYPE iasc_master SYSTEM "http://www.iasc.org.uk/dtd/iasc_master.dtd">
```

**Figure 5 Link to Hypothetical International Accounting XML DTD**

The adoption of an accounting-specific XML DTD would potentially improve the accuracy of WebBots retrieval close to the desired 100% level. The WebBot would act directly on the information nominated in the document. Inaccuracies would arise primarily from incorrect tagging by the corporation or data errors (Kambil and Ginsburg 1998, 92).

In summary, adoption of a standard XML DTD for the representation of accounting information would allow WebBots to retrieve specific accounting information with near to certain probability of accuracy. Use of a standard accounting XML DTD would also improve resource discovery, by allowing searches and the content of search engine databases to be restricted to Web DLOs that link to the standard DTD. As Kambil and Ginsburg (1998, 92), the original developers of the Web-based interface to the

---

[9] The SEC EDGAR DTD was established when EDGAR information was collected by on behalf of the SEC by LEXIS. With the EDGAR project likely to move to the private sector, the future of the EDGAR DTD is unclear. While LEXIS is not currently involved with EDGAR, most companies still file using the 1995 EDGAR DTD.

SEC's EDGAR documents note, "until the EDGAR systems are reengineered with a fuller and standard set of semantic tags strictly enforced by the SEC, it is a difficult technical process to extract information".

# 4 METADATA REPRESENTATION

Section 3 discussed how XML provides a technique to tag and subsequently locate a specific item of information. However, to start searching for a specific item the user or intelligent agent must first locate the applicable Web page(s) that contain the desired items. Section 4 discusses the role of metadata in performing this latter task.

An important element of Web mining identified by Etzioni (1996) is resource discovery (e.g., finding the desired Web resources). At present, resource discovery on the Web can be divided into three categories:

- Ad-hoc hypertext links from page to page
- Full-text brute-force indexes such as AltaVista
- Structured hierarchical objectories such as Yahoo.

The inclusion of metadata information on a DLO significantly improves resource discovery. Berners-Lee (1997a) defines metadata as being "machine understandable information about web resources or other things."

At one end of a metadata continuum is the content of brute-force full-text indexes such as AltaVista. These indexes are metadata of a form, albeit without any pre-defined structure and a near absolute level of granularity. As Weibel (1995) notes, the usefulness of an index increases in inverse proportion to the size. He notes that "indexes are most useful in small collections within a given [knowledge] domain."

At the other end of the metadata continuum, a database schema is a form of metadata that relatively precisely models the real world and does so with a pre-defined structure (Elmasri and Navathe 1994, 6). A database schema is organisation and application specific and is not normally designed to be interoperable with applications outside the organisation.

There are many different formats of interoperable metadata. One of the most popular forms of Web metadata is the meta tags introduced with HTML 2.0. These meta tags are placed in the <HEAD></HEAD> section of Web page. The browser does not display this information. However, this

information is used by some of the search engines to classify and categorize Web pages. A frequent meta tag is "keywords" and could look something like:

```
<META NAME="keywords" CONTENT= "[a list of keywords would go here]">
```

The "scheme" attribute allows the tag to refer to some predefined scheme, which describes the meaning of the content. The following are tags, which refer to pre-defined schemata:

```
<META name="date" scheme="ISO" content="980115">
<META name="identifier" scheme="ISBN" content="0201310163">
```

Unfortunately, it appears that most developers of financial reporting Web sites do not know how or care to exploit these meta tags. Table 2 demonstrates the current under-utilization of meta tags. Note that two of the nine companies who have Web pages use no meta tags for either their home page or annual report page. Six do use *some* meta tags on their home page, but not their annual report page. In general, based on this small sample, those companies that do use meta tags, they do not use them very extensively.

**Table 2 The Use of Meta Tags by the Fortune 10**

| | | Meta Tags on Home Page | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Fortune 500 Rank | Company | Author | Owner | Description | Keywords | Abstract | Alias | Date | Revision | Developer | Program | Reply-to | Refresh | PIC-label |
| 1 | General Motors | | | X | X | | | | | | | | | |
| 2 | Ford Motor | | | | | | | | | X | | | | |
| 3 | Exxon | | | | X | | | | | | | | | |
| 4 | Wal-Mart | X | | | X | | | | | | X | | | X |
| 5 | AT&T | | | | X | | | X | X | | | X | | X |
| 6 | IBM | | X | X | X | X | X | | | | | | X | X |
| 7 | General Electric | | | X | X (1) | | | | | | | | | |
| 8 | Mobil | | | | | | | | | | | | | |
| 9 | Chrysler | | | | | | | | | | | | | |
| 10 | Philip Morris | | | | | | | | | | | | | |
| | | **Meta Tags on Annual Report Page** | | | | | | | | | | | | |
| Rank | Company | Author | Owner | Description | Keywords | Abstract | Alias | Date | Revision | Developer | Program | Reply-to | Refresh | PICS-label |
| 1 | General Motors | | | | | | | | | | | | | |
| 2 | Ford Motor | | | | | | | | | | | | | |
| 3 | Exxon | | | | | | | | | | | | | |
| 4 | Wal-Mart | X | | | X | | | | | | X | | | X |
| 5 | AT&T | | | | X (2) | | | X | X | | | X | | |
| 6 | IBM | | | | | | | | | | | | | |
| 7 | General Electric | | | | | | | | | | | | | |
| 8 | Mobil | | | | | | | | | | | | | |
| 9 | Chrysler | | | | | | | | | | | | | |
| 10 | Philip Morris | | | | | | | | | | | | | |

(1) Included 97 keywords
(2) Keyword was "AT&T House of Style"

One of the more interesting meta tag usage was by General Electric. They only used two meta tags, description and keywords, but they listed 97 keywords ranging from nuclear to laundry reflecting its very diversified organization.

The problem with these HTML meta tags is that neither the tags themselves nor the parameters (e.g., keywords) are standardized. That is, Web developers are free to use any tags they want and use any content terms they believe appropriate. Further, few Web page developers are aware of the current incentives to use meta tags. Not all of the search engines provide added weight to meta tags,[10] so that even if the use of the HTML meta tags increased they would still fall short of the power of some of the alternative metadata formats that are either currently available or under development. Dempsey and Heery (1997), with a library orientation, identify 23 different metadata formats. The most familiar of these formats is the MARC format[11] (Machine Readable Catalogue Format), which is used in somewhat varying forms by the Library of Congress, the National Library of Australia and other national libraries for their catalogues of books and other library materials.

**Cost Benefit Tradeoffs**

The provision of metadata on the Internet is subject to a cost-benefit relationship between functionality of the metadata format and the time taken to generate the metadata within the particular schema. Thoroughly organized metadata schemata, such as MARC, are very costly to maintain. It takes approximately 0.2 person-days to develop a new MARC record. This is an appropriate investment to make for a printed book but not for the great bulk of pages in HTML (Weibel 1995). In two hundred years, the Library of Congress catalogue has accumulated 17 million book titles. This compares with the 75 to 140 million pages on the Web indexed by the larger search engines or the estimated 200 million separate, publicly available pages available for indexing[12]. Many of the pages on the Internet have only a relatively short duration, and most contain only a fraction of the information as in a book. It would hard to justify the efforts required to prepare a MARC record for each Web page.

---

10 Searchenginewatch states that for only two (Hotbot and Infoseek) of the seven search engines does the existence of meta tags affect the ranking of web pages. See *www.searchenginewatch.com/features.htm*

11 See *http://lcweb.loc.gov/marc/marc.html*

12 Source: Estimates of number of pages indexed by the major search engine reported at: *http://www.searchenginewatch.com/*

Market acceptance of a metadata format will depend on the functionality of information discovery tools and the extent of information that has been subjected to the particular metadata schema. Successful metadata schema will be those that do not add significant development time and costs yet make their Web content more readily tractable and discoverable so that the content providers receive a return on the cost of providing the metadata. Reuse of metadata information is also likely to be important, as collections of pages within a particular knowledge domain will vary only at the margin. The pages at *http://www.aicpa.org/* will share many common content descriptors including "accounting," "auditing" and "American Institute of Certified Public Accountants" .

A number of developments are now underway in the government, library, Internet, and Web communities to build cost-effective frameworks for metadata. These include the Dublin Core, the W3 Consortium's RDF project and the US Government Information Locator Service (GILS)[13] . The latter format[14] is important, as it is likely to influence the development of governmental services in the USA such as the SEC's EDGAR. GILS is, however, a rather complex format as it has been heavily influenced by the MARC standard (Dempsey and Heery 1997). Many of the current Web applications of GILS are merely gateways to Z39.50[15] pipes to pre-existing databases based on MARC. Because of the required effort, it is highly questionable, unless it were to be a requirement of securities regulators, whether producers of information outside of the public sector will take the time to generate Web metadata under the GILS standard. The other two frameworks, the RDF protocol and the Dublin Core metadata container set, can be seen as an integrated solution to metadata representation. Each, however, can and does stand independently.

RDF[16] is a proposed layer for the representation of metadata of Web URLs that sits on top of XML (Berners-Lee 1997b). As XML is readily extensible, a number of potential applications are under construction, including an interface between XML and EDI, XML/EDI[17]. There is a proposal from

---

13 Overviews of these and other metadata formats and pointers to URLs are incorporated in Dempsey and Heery 1997 at *http://www.ukoln.ac.uk/metadata/DESIRE*

14 See Christian 1996 and *http://www.usgs.gov/public/gils/*

15 An international standard for common real time interfaces to databases and catalogues.

16 See *http://www.w3.org/TR/WD-rdf-syntax* and *http://www.w3.org/metadata/*

17 See the executive overview at *http://www.geocities.com/WallStreet/Floor/5815/xmlexec.htm*

Microsoft Corporation for a W3C recommendation for an XML-compliant Web data format[18]. The W3C is working on proposals for XML layers for rating of Web pages (PICS 2.0) and for privacy (P3P). RDF employs XML namespaces to provide either internal representation of metadata or, more likely, links to external metadata schema such as Library of Congress, Dewey Decimal, British Library or the Dublin Core. RDF allows the creation of *n-ary* metadata representations of a DLO.

RDF provides a neutral XML-compliant mechanism to define and encode metadata schemas. The so-called Dublin Core is a relatively simple container that has been explicitly designed for cost effective declaration of metadata elements within DLOs[19]. Weibel notes:

> *An alternative solution that promises to mediate these extremes* [of full text indexes and MARC type indexes] *involves the creation of a record that is more informative than an index entry but is less complete than a formal cataloging record. If only a small amount of human effort were required to create such records, more objects could be described, especially if the author of the resource could be encouraged to create the description. And if the description followed an established standard, only the creation of the record would require human intervention; automated tools could discover these descriptions and collect them.*(Weibel 1995)

A standards development team from the library, Web and Internet communities along with content specialists has designed the Dublin Core protocol in this spirit. The protocol has been developed at a series of international meetings, the first of which was held in Dublin, Ohio in 1995. While not all the elements of the Dublin Core have been resolved, most of the building blocks are now in place. The core elements of the Dublin Core have been largely unchanged since December 1996 and are now formalized in an IETF RFC.

Dublin Core metadata is designed to enhance resource discovery of DLOs by the creation of a schema of metadata containers that will be used by Web search engines. It has been designed to be employed by a variety of low-volume and high-volume content producers.

The Dublin Core metadata is not designed to provide access to directories, databases or other rapidly changing database-like resources. Other protocols such as X.500, LDAP, whois++ and, to a lesser

---

[18] Proposed "Specification for XML-Data" at *http://www.microsoft.com/standards/xml/xmldata.htm*

[19] The Dublin Core home page is at http://purl.oclc.org/metadata/dublin_core/ See also Weibel 1995; Cathro 1997.

extent, Z39.50 are more appropriate for the task of querying such resources. Most accounting data of the type found in periodic financial reports is contained in DLOs.

The Dublin Core has fifteen elements[20] as shown in Figure 6:

---

[20] See Dublin Core Metadata Element Set: Reference Description at http://purl.oclc.org/metadata/dublin_core_elements, Weibel, et al. 1997, Cathro 1997.

| Element | Label | Comment[21] |
|---|---|---|
| Title | *title* | The name given to the resource. |
| Author or Creator | *creator* | "The person or organization primarily responsible for creating the intellectual content of the resource." |
| Subject and Keywords | *subject* | It is envisaged that controlled vocabularies and formal classification schemas would be used |
| Description | *description* | "A textual description of the content of the resource" |
| Publisher | *publisher* | |
| Other Contributor | *contributor* | "A person or organization not specified in a *creator* element who has made significant intellectual contributions to the resource but whose contribution is secondary to any person or organization specified in a *creator* element". |
| Date | *date* | "The date the resource was made available in its present form." |
| Resource Type | *type* | "The category of the resource, such as home page, novel, poem, working paper, technical report, essay, dictionary." |
| Format | *format* | "The data format of the resource." |
| Resource Identifier | *identifier* | "String or number used to uniquely identify the resource. Examples for networked resources include URLs and URNs (when implemented)." |
| Source | *source* | "A string or number used to uniquely identify the work from which this resource was derived, if applicable. For example, a PDF version of a novel might have a *source* element containing an ISBN number for the physical book from which the PDF version was derived." |
| Language | *language* | "Language(s) of the intellectual content of the resource." |
| Relation | *relation* | "The relationship of this resource to other resources." Experimental |
| Coverage | *coverage* | "The spatial and/or temporal characteristics of the resource." Experimental. |
| Rights Management | *rights* | "A link to a copyright notice, to a rights-management statement, or to a service that would provide information about terms of access to the resource." |

**Figure 6 Dublin Core Elements**

Figure 7 shows a sample RDF metadata, using Dublin Core tags, for the second quarter 1998 results for a hypothetical corporation, RiverRidge Inc.

---

[21] Comments from the Dublin Core Reference Description shown in quotes.

```
<?namespace href=http://purl.org/DublinCore/RDFschema" as "DC"?>                    1
<?namespace href=http://www.w3.org/schema/rdf-schema" as "RDF"?>
<RDF:serialization>                                                                 2
        <DC:date.current>
                <RDF:Value>=19980818</RDF:Value>
                        <DC:scheme>ANSI.X3.30-1985</DC:scheme>
                </RDF:resource>
        </DC:date.current>
        <DC:title>RiverRidge Second Quarter 1998 Results"</DC:title>
<DC:creator>
                <DC:name>Miklos McCarthy</DC:name>
        <DC:email>cfo@riverridge.com"</DC:email>
                <DC:affiliation>RiverRidge Inc"</DC:affiliation>
                <DC:postal>1000 Wood Way, River Road, Dublin, OH, USA 999999"</DC:postal>
                <DC:homepage>http://www.riverridge.com/investor/"</DC:homepage>
        <DC:creator>
        <DC:description>The Second Quarter 1998 results for RiverRidge
Inc"</DC:description>                                                               3
        <DC:description>
                <RDF:Value> Financial Statements, Financial Reporting</RDF:Value>
                        <DC:scheme>ProQuest</DC:scheme>
                </RDF:resource>
        </DC:description>
        <DC:publisher>RiverRidge Inc"</DC:publisher>
        <DC:type>OrganisationInfo"</DC:type>
<RDF:serialization>
```

**Figure 7 Dublin Core Tags in RDF Container for RiverRidge Inc**

Further elaboration is required on particular attributes within the metadata:

1. This is the first example of the use of the Scheme qualifier. Dates in the date field might be written as simply as "18 July 1998" or may, as in this case, conform to a pre-defined standard, the ANSI X3.30-1985 standard. Search engines will be able to recognise the ANSI standard and provide a human readable equivalent of 19971018.

2. The complete metadata representation is contained within the "serialization" tag.

3. Each Dublin Core element can be repeated as many times as necessary. Here a plain text description is provided in the first descriptor. In the second descriptor a reference is made to the controlled thesaurus provided by UMI's ProQuest, which is in the public domain. A specialist search engine would then be able to interpret the ProQuest scheme.

When the major search engines begin to work with metadata containers such as that defined in the Dublin Core, much of the metadata will be sourced remotely from repositories of metadata tags. The metadata for each of the three-quarters of a corporation's quarterly reports will be largely identical with the only difference being the effective period of the report. Rather than embedding all the metadata elements in the DLO, a pointer will be made to the appropriate elements of the template. Metadata schemes could also be developed by organizations such as the AICPA, FASB or IASC. A link could then be made to the appropriate URL. Similarly a metadata scheme for the representation of stock exchange codes could be developed for each bourse. Once accounting-specific metadata schemes are developed and introduced,

accounting-specific search engines can be developed which would significantly enhance information retrieval. Existing technology, such as Harvester Brokers, can be used to search out this specialist information on the Web.

# 5 SUMMARY AND CONCLUSION

Accounting information on the Web is already ubiquitous. It has been brought to the Internet without the involvement of the accounting profession, accounting standards setters or securities regulators. The inconsistent presentation of accounting information by corporations, the vast scale of the Internet and the inherent limitations of HTML combine to mean that accounting information on the Web is very difficult to find and almost impossible to automatically retrieve even the most common of accounting attributes. The paper has demonstrated that current problems of attribute identification and resource discovery effectively prohibit the use of automated software agents, or WebBots, in acting upon accounting information on the Web. Yet, the prospect of automated retrieval and processing of accounting information is clearly attractive to a wide range of stakeholders.

The paper laid out a model that, if employed by the profession and corporations, would raise the confidence levels of discovery of pages and identification of attributes to a point that WebBots could be trusted to act upon Web-based financial statements. The model is based upon Web standards, most notably the W3 Consortium's Extensible Markup Language (XML) and Resource Description Framework (RDF) and the metadata community's Dublin Core metadata container framework. The model described in this paper would also be cost-effective for corporations as XML, RDF and the Dublin Core have been explicitly designed to be in the spirit of the Web--to bring information to a global community quickly and cheaply.

Such a solution goes significantly beyond national initiatives by securities regulators such as the SEC. The model would allow a corporation to make any type of accounting disclosure on the Web, not just those mandated by the SEC. The proposed model is, however, isomorphic with the 1995 SEC standards for the representation of accounting information within the EDGAR system. Indeed, the EDGAR Document Type Definitions (DTDs) would provide an immediate platform upon which national and international standards could be built.

There are a number of areas that this paper has not addressed and a number of limitations to the proposed model. First, the model assumes static reporting of unchanging accounting numbers. However, industry leaders, such as Robert Elliott of KPMG, have predicted that consumers of financial statements

will interact directly with corporate databases (Elliott 1992, 1994). The proposed model has not addressed the manner in which human or automated financial statements users might interact with underlying databases of more detailed information to provide data analysis and drill-down capabilities or to understand the assumptions upon which the financial statements were based.

There are also limitations to the proposed reporting model. The model has assumed that financial information is reported in static pages written in HTML or XML. The model, as currently laid out, does not support dynamic data within the financial reporting pages, although the pages might employ dynamic multimedia features.

This is the first paper to address issues of resource discovery or attribute identification in an accounting domain. It has, necessarily, skimmed the surface of implementation of XML, RDF and the Dublin Core. Much more research is required on issues such as how accounting attributes are to be identified or how national and international repositories of metadata might be integrated into corporate meta tagging programs. The paper has not touched on issues such as the digital signing of financial reports by corporations and auditors. Nor has it addressed international differences in financial reporting and accounting terminology.

A number of research issues arise from this research. XML and possible forthcoming standards such as XML-Data provide functionality similar to that of a database schema. The use of XML documents as a universal front-end for database accounting information systems of the type envisaged by McCarthy 1982 is an interesting prospect. The enhanced hypertext capabilities of XML also present a number of opportunities in researching user-interaction with hypertext-based financial reports.

Perhaps most important is the need to research the development of formal ontologies of accounting as a foundation for the practical implementation of structures of accounting knowledge in a DTD (Wand and Wang 1996; Guarino and Poli 1995; Guarino 1995; Gruber 1995).

# REFERENCES

Adriaans, P., D. Zantinge, and P. Adriaans. *Data Mining*. New York: Addison-Wesley Publishing Company, 1996.

Baker, W. M., and P. R. Witmer. "Intelligent agents go to work for management accountants." *Management Accounting (USA)* 78, no. 10 (1997): 32-5.

Barquin, R. C., and H. A. Edelstein, eds. *Building, Using, and Managing the Data Warehouse*. Edited by Series, D. W. I. Englewood Cliffs, NJ: Prentice Hall, 1997.

Berghel, H. "Cyberspace 2000: Dealing with information overload." *Communications of the ACM* 40, no. 2 (1997): 19-24.

Berners-Lee, T. (1997a). Metadata Architecture: Documents, Metadata and Links. 6 January. W3 Consortium. http://www.w3.org/DesignIssues/Metadata.html.

Berners-Lee, T. (1997b). W3C Data Formats. 29 October. W3 Consortium. http://www.w3.org/TR/NOTE-rdfarch.

Berson, A., and S. J. Smith. *Data Warehousing: Architecture and Technology, McGraw-Hill Series on Data Warehousing and Data Management*. Computing McGraw-Hill, 1997.

Bosak, J. (1998). XML, Java and the future of the Web. 10 March. Sun Microsystems. http://sunsite.unc.edu/pub/sun-info/standards/xml/why/xmlapps.htm.

Bradshaw, J. M., ed. *Software Agents*. Cambridge, Mass: MIT Press, 1997.

Bray, T. (1997). Beyond HTML: XML and Automated Web Processing. . Netscape Communications Corp. http://developer.netscape.com/news/viewsource/bray_xml.html.

Cathro, W. (1997). *Metadata: An Overview* (http://www.nla.gov.au/nla/staffpaper/cathro3.html). Canberra: National Library of Australia.

Chen, C., and R. Rada. "Interacting With Hypertext - A Meta-Analysis Of Experimental Studies." *Human-Computer Interaction* 11, no. 2 (1996): 125-56.

Cheung, D. W., B. Kao, and J. Lee. "Discovering User Access Patterns on the World Wide Web." In *KDD: Techniques and Applications*, edited by Lu, H.-J., Motoda, H. and Liu, H., 303-16. Singapore: World Scientific, 1997.

Chorafas, D. N. *Agent Technology Handbook*. New York: McGraw Hill, 1997.

Christian, E. J. "GILS What is it? Where's it going?" *D-Lib Magazine*, no. December (1996): http://www.dlib.org/dlib/december96/12christian.html.

Ciolek, T. M. "Today's WWW--tomorrow's MMM? The specter of multi-media mediocrity." *Computer* 29, no. 1 (1996): 106-8.

Connolly, D., ed. *XML: Principles, Tools and Techniques*. Travelers' Tales Inc, 1998.

Connolly, D., R. Khare, and A. Rifkin. "The Evolution of Web Documents: The Ascent of XML." *World Wide Web Journal* 2, no. 4 (1997): 119-28.

Deller, D., M. Stubenrath, and C. Weber. "A Survey of the Use of the Internet for Investor Relations in the USA, UK and Germany." *European Accounting Review* (Forthcoming).

Dempsey, L., and R. Heery. (1997). A Review of Metadata: A Survey of Current Resource Description Formats. March. UKOLN, University of Bath. http://www.ukoln.ac.uk/metadata/DESIRE/overview.

Derose, S. J. *The SGML FAQ Book : Understanding the Foundation of HTML and XML*. Translated by ISBN: 0792399439. Amsterdam: Kluwer Academic Publishers, 1997.

Elliott, R. K. "The Third Wave Breaks on the Shores of Accounting." *Accounting Horizons* 6, no. 2 (1992): 61-85.

Elliott, R. K. "Confronting the future: Choices for the attest function." *Accounting Horizons* 8, no. 3 (1994): 106-24.

Elmasri, R., and S. Navathe. *Fundamentals of Database Systems*. 2nd ed. Reading, MA: Benjamin Cummings, 1994.

Etzioni, O. "The World Wide Web: Quagmire or gold mine?" *Communications of the ACM* 39, no. 11 (1996): 65-8.

Gardner, S. R. "Building the Data Warehouse." *Communications of the ACM* 41, no. 9 (1998): 52-60.

Goldfarb, C. F. *The SGML Handbook*. Oxford: Clarendon Press, 1990.

Goldfarb, C. F., and P. Prescod. *The XML Handbook*. Upper Saddle River, NJ: Prentice Hall PTR, 1998.

Goldie, P. "Using SGML to Create Complex Interactive Documents for Electronic Publishing." *IEEE Transactions on Professional Communication* 40, no. 2 (1997): 130-8.

Gray, G. L., and R. S. Debreceny. "Corporate Reporting on the Internet: Opportunities and Challenges." Paper presented at the Seventh Asian-Pacific Conference on International Accounting Issues, Bangkok, November 1997.

Gruber, T. R. "Toward Principles For the Design of Ontologies Used For Knowledge Sharing." *International Journal of Human-Computer Studies* 43, no. 5-6 (1995): 907-28.

Guarino, N. "Formal Ontology, Conceptual Analysis and Knowledge Representation." *International Journal of Human-Computer Studies* 43, no. 5-6 (1995): 625-40.

Guarino, N., and R. Poli. "The Role of Formal Ontology in Information Technology." *International Journal of Human-Computer Studies* 43, no. 5-6 (1995): 623-4.

Herwijnen, E. v. *Practical SGML*. Boston, Mass.: Kluwer Academic Publishers, 1994.

Holzner, S. *XML Complete*. Computing McGraw-Hill, 1997.

Huhns, M. N., and M. P. Singh. *Readings in Agents*. San Francisco: Morgan Kaufman Publishers, 1997.

Kambil, A., and M. Ginsburg. "Public Access Web Information Systems: Lessons from the Internet EDGAR Project." *Communications of the ACM* 41, no. 7 (1998): 91-7.

Kautz, H., B. Selman, and M. Shah. "Referral Web: Combining social networks and collaborative filtering." *Communications of the ACM* 40, no. 3 (1997): 63-5.

Khare, R., and A. Rifkin. "Capturing the State of Distributed Systems with XML." *World Wide Web Journal* 2, no. 4 (1997): 207-18.

Kim, H., and S. C. Hirtle. "Spatial metaphors and disorientation in hypertext browsing." *Behaviour and Information Technology* 14, no. 4 (1995): 239-50.

Koreto, R. J. "When the bottom line is online." *Journal of Accountancy* 183, no. 3 (1997): 63-5.

Maes, P. "Agents that reduce work and information overload." *Communications of the ACM* 37, no. 7 (1994): 30-40.

McCarthy, W. "The REA Accounting Model: A Generalized Framework for Accounting systems in a Shared Data Environment." *Accounting Review* 57, no. 3 (1982): 554-78.

Norman, D. A. "How might people interact with agents." *Communications of the ACM* 37, no. 7 (1994): 68-71.

Resnick, P., and H. R. Varian. "Recommender systems." *Communications of the ACM* 40, no. 3 (1997): 56-8.

Riecken, D. "Intelligent agents." *Communications of the ACM* 37, no. 7 (1994): 18-21.

Rucker, J., and M. J. Polanco. "Siteseer: Personalized navigation for the Web." *Communications of the ACM* 40, no. 3 (1997): 73-5.

St. Laurent, S. *XML: A Primer*. New York: MIS Press, 1998.

Tauber, J. K. *Teach Yourself XML in 21 Days*. Sams Net, 1998.

Terveen, L., W. Hill, B. Amento, D. McDonald, and J. Creter. "Phoaks: A system for sharing recommendations." *Communications of the ACM* 40, no. 3 (1997): 59-62.

Wallman, S. M. H. "The future of accounting and financial reporting Part IV: Access Accounting." *Accounting Horizons* 11, no. 2 (1997): 103-16.

Wand, Y., and R. Y. Wang. "Anchoring data quality dimensions in ontological foundations." *Communications of the ACM* 39, no. 11 (1996): 86-95.

Weibel, S. "Metadata: The Foundations of Resource Description." *D-Lib Magazine*, July 1995.

Weibel, S., R. Iannella, and W. Cathro. "The 4th Dublin Core Metadata Workshop Report." *D-Lib Magazine* 1997.