SUPPORTING WEB CONTENT QUALITY: FORMALIZING META-DATA CONCEPTS FOR THE WEB-DOMAIN

Dawn G. Gregg
School of Accountancy and Information Management
College of Business
Arizona State University
Tempe, AZ 85210

Home phone: (480) 706-0860
Fax Number: (480) 965-8392

email: dawn.gregg@asu.edu

Number of words in abstract: 150

Number of words in text (sans abstract and references): 3494

Keywords: Information Attributes, Data Resource Management, Interorganizational Systems, Information search and retrieval

SUPPORTING WEB CONTENT QUALITY:
FORMALIZING META-DATA CONCEPTS FOR THE WEB-DOMAIN

ABSTRACT

Poor web data quality can have many costs for the typical enterprise. This is true for web based data as well as for traditional data sources. This applied research investigates the use of distributed artificial intelligence approaches for supporting web content quality. A series of prototyping/validation tasks will be conducted to evolve web-based systems that can be used to maintain web content quality. First, a meta-data protocol that provides improved access to DSS will be developed and validated. Next, a formal web data model will be developed. This model will offer a framework for defining web object dependencies and representations. These two concepts are then combined to develop a prototype Web-Quality protocol. This protocol will facilitate the maintenance of web-based content by allowing meta-data about web pages and the relationship between web pages to be distributed with web pages. Intelligent agents are used to translate this meta-data into appropriate action.

1.0 BACKGROUND AND OBJECTIVES

Supporting the quality of data resources has been a continuing concern for information systems professionals. Over time techniques have been developed for maintaining the appropriate level of quality for individual databases, for data warehouses and for transaction processing systems. However, web-based systems lack the tools and procedures for data quality to be properly maintained.

This paper seeks to develop methods that can be used to improve web-based data quality. It maps data quality dimensions, as identified in prior research [Wand et. al., 1996, and Wang et.

al., 1995], to the web domain and then proposes methods that can improve each of these data quality dimensions.

Data quality problems can create many different types of problems for organizations. Table 1 lists some of the potential costs businesses might face as a result of poor web data quality.

Table 1 Potential Costs of Poor Quality Web Content

Web Site Type	Possible Quality Failures	Potential Costs
Electronic Commerce	Out of date or missing product information	Loss of Sales
Information/ Search	Poor Quality Information Broken Links	Reduction in # Hits Loss of Advertising Revenue Loss of Subscriber Revenue
Corporate Information	Lack of Consistent Image Inaccurate Data	Degradation of Brand Identity Reduced Stock Price

For example, Reuters Air Cargo Service provides information and trading opportunities to airlines, forwarders, shippers, brokers and general sales agents worldwide. One service it provides is constantly updated rates for airfreight transport. This site must continuously update these rates or they will risk losing subscribers. This research will investigate using "Meta-data" to specify relationships between different web content. Intelligent agents are used to read this meta-data and update the content. In this example, a relationship between the price on the Reuters site and the price at each airfreight provider can be specified. Then, an intelligent agent can read the meta-data, go to the airfreight site, get updated price information and update the Reuters web page automatically.

This research combines two important information systems research streams to allow the development of the web quality meta-data and intelligent agents. Both data quality and distributed artificial intelligence (DAI) play a foundational role in the development of the distributed meta-data systems proposed in this paper.

1.1 Distributed Artificial Intelligence

Advances in DAI make it possible to construct systems of intelligent agents that can be used to discover and maintain web information. DAI research is concerned with how multiple agents can cooperate, how they communicate, share information and act to achieve a common goal [Bond and Gasser, 1988, Malone and Crowston, 1994].

Early work in artificial intelligence serves as a foundation for the intelligent agent research being conducted today. In 1980, Smith developed the "Contract net protocol" (CNET) which was a high-level protocol designed for distributed problem solving. More recently, a federated coordination and communication system for agents was developed [Goul, 1997]. In a federated system, agents do not communicate directly with each other. Instead, the agents communicate only with system programs called facilitators or mediators [Wiederhold, 1992]. The federated agent communication approach allows agents to communication in large environments such as the WWW.

Web-based applications exploit a wide range of AI and DAI developments. Today AI is embedded in agents operating in heterogeneous networked computing environments and used for search, retrieval and analysis of previously unimaginable quantities of data [O'Leary, 1997]. This research includes intelligent search engines, intelligent browsers, and intelligent agents, web page labeling schemes and information brokering systems [Bowman et. al., 1994]. Recent intelligent agents have been designed to find and classify information on the WWW [Acerman et. al., 1997, Etzioni, 1997, Kautz et. al., 1997a,b, Krulwich, 1997, and Burke et. al. 1997]. These agents do not attempt to map or understand the entire web. Instead, they attempt to process specific types of content about which the agent has some prior knowledge. This enables these agents to synthesize the web content from the limited domain and provide a useful service to the user.

Internet resource discovery agents are largely autonomous rather than collaborative. That is they use intelligence to find and decode web content rather than to work with other agents. A full scale meta-data system will require both autonomous agents, ones that can watch for a change in a web page or search for specific meta-data; and collaborative agents, ones that work with both users and other agents to infer appropriate high-level goals and take appropriate action to achieve these goals [Nardi, Miller and Wright, 1998].

1.2 Data Quality

Data quality problems can occur in any system where data is stored. Early research on data quality measured errors in specific management information systems (MIS). In 1982, Morey found a 25% error rate in incoming transactions and an 11.21% error rate in stored MIS records for a Marine Corps Manpower Management System. In 1986, Lauden studied a large interorganizational computer based information system for the criminal justice system and found that 50 to 80% of the computerized criminal records were inaccurate, incomplete or ambiguous.

Following these studies, several researchers have attempted to define the dimensions upon which data quality should be measured. Data quality has been defined in terms of accuracy (recorded value in agreement with the actual value), timeliness (recorded value is not out of date), completeness (all values for certain variables are recorded), and consistency, (the representation of the data value is complete in all cases) [Ballou and Pazer, 1987; Huh, Keller, Redman, and Watkins, 1990]. An additional attribute, accessibility, was proposed in an ontological classification scheme proposed by Wand and Wang [1996]. The six data quality attributes discussed in this are derived from this classification scheme and are presented in Table 2.

Table 2: Data Quality Dimensions [Wand and Wang, 1996, and Wang, Reddy and Kon, 1995]

Data Quality Dimension	Characteristic
Accessibility	Data is available to (easily found by) the user,
Completeness	Every meaningful state of the specified real world system can be represented.
Believability	The extent to which data can be counted on to be correct.
Currency	Data are current if they do not reflect outdated information
Accuracy	Data agrees with an identified source to a desired precision.
Consistency	Data are consistent if they do not conflict with one another.

Research on data quality problems in databases have indicated that the social and economic impact of poor-quality data costs billions of dollars [Wang, Storey and Firth, 1995]. These costs include costs due to customer dissatisfaction, increased operational costs, lowered employee job satisfaction, and poorer tactical and strategic decision making [Redman, 1998]. There is no reason to expect that poor content quality on the Internet is any less costly.

The WWW is a new type of data repository containing information millions of people use every day. One data quality problem associated with the WWW is the frequent failure of hyperlinks. Several studies have sited broken hyperlinks as "one of the most serious problems facing the WWW today" [Ingham, Caughey and Little, 1996, Kehoe, Pitkow, and Rogers, 1998]. Although solutions for dealing with broken links are well known, 57.7 percent of respondents to a recent web survey cite broken-hyperlinks as a serious problem with the WWW [Kehoe, Pitkow, and Rogers, 1998].

However, broken hyperlinks represent only one type of data consistency problem that can occur on the web. Data quality problems along the dimensions found in other types of data repositories are also present on the web. Strong, Lee and Wang [1997] argue that a data quality problem is any difficulty encountered along one or more quality dimensions that renders it completely or largely unfit for use. Nielson [1998] found many commercial web sites have data quality problems that fit this definition.

Some research has already been done related to improving data quality on the WWW. One approach is to centralize the data quality function to try to manage web data in the database style. "ARENEUS" [Mecca et. al., 1998] supports queries, views and updates to web data. This system creates "Wrappers" that describe page content. These Wrappers are stored in a Wrapper Library that can then be used to provide database like access to web data. One shortcoming of the ARENEUS approach is that the meta-data is housed in a central repository, which limits access, reduces local autonomy, and limits expandability of the system [Ozsu and Valdurez, 1991].

By contrast, the approach proposed in this paper is "distributed," reducing many of the problems found in centralized systems. The advantages of distributed meta-data are discussed in the next section.

1.3 Distributed Meta-data

This paper proposes using DAI approaches to support web-based content quality. Under this approach intelligent agents would be used to read and interpret meta-data related to web-based content. Meta-data is information about information. Meta-data has been used in databases to describe the relationships between tables and is currently being used on the web to provide a title, author, and description for web-pages [Elmasri & Navathe, 1994, Resnick, 1997]. Meta-data approaches allow classification of web content and can describe the relationships between content and other web resources. Meta-data can be created at the same time as a web page. It consists of specialized <tags> that can be read by intelligent agents designed for web site search and/or maintenance.

Supporting web content quality requires meta-data be developed to support the six data quality dimensions listed in Table 2. For example, meta-data can be used to identify the content and quality of web pages [Resnick, 1997; Resinick & Miller 1996]. Web pages can include meta-data related to page type (i.e. home page, product page, sales site, information site, white paper, software download etc). Meta-data can also include information related to a page's subject area. Subject areas included in the Yahoo web index are: Arts & Humanities, Business & Economy, Computers & Internet, Education, Entertainment, Government, Health, News & Media, Recreation & Sports, Reference, Regional, Science, Social Science, and Society & Culture [Yahoo, 1998]. This meta-data would be designed to improve the *accessibility* of web content. That is, it would be designed to improve end-user's ability to locate specific content or resources that are available on the WWW. It is also possible for content meta-data to provide a measure of the *completeness* of a specific web resource.

Supporting the *accuracy*, *consistency* and *currency* (per table 2) of web resources can also be accomplished using meta-data. Often content on a web page can be related to content found on other web pages, yet there is currently no mechanism for ensuring that updates to web based content are automatically propagated to related web sites. This requires that meta-data concepts be expanded to include information about the complex relationships among web-based content, thus, allowing related content to be retrieved and updated efficiently.

Meta-data should also improve end-user's ability to assess the *believability* of web content because they will have a better understanding of what the content is and of the author of the page. In addition, end-users can use the meta-data to assess whether a particular web page is adequately maintained and to determine the source for some of the critical content contained in the page.

2.0 RESEARCH QUESTION

The purpose of the proposed research is to investigate methods for organizing the content in a web-space such that discovering, tracking and updating that content is facilitated. It will evaluate how the capabilities of intelligent agents, DAI and the WWW can be integrated to create advanced information systems capable of maintaining content quality across a web-space. Based on this research purpose the following general research question is proposed:

To what extent can distributed meta-data:

1) Provide specific data related to the page type and subject area.

2) Meet the specific needs of the target end-user population.

3) Provide information related to the content type and appropriate values for specific web content or resources.

4) Specify how a resource in one file is related to resources in other files; and

5) Be designed so autonomous agents can readily use it when processing retrieval and update activities?

Answers to the above research question can be useful in the design and development of systems to manage distributed web content.

3.0 RESEARCH METHODOLOGY

This applied research investigates the use of DAI approaches for supporting web content quality. A series of prototyping/validation tasks will be conducted to evolve a web-based system that can be used to maintain web content quality.

This research begins by exploring the use of meta-data to provide specific information about the content of a web page. It proposes a protocol suite for deploying and sharing content for a specific domain, Decision Support Systems [Gregg and Goul, 1999]. This protocol suite will be validated using an exploratory experimental design with a selected group of subjects.

The meta-data concepts will then be expanded to incorporated the data quality dimensions presented in Table 2. A conceptual formal software engineering methodology will be used to define a data model tailored to the needs of the WWW environment [Vinze, Kulkarni and Gregg, 1999]. The web data model will then be converted to an Extended Markup Language (XML) document content specification protocol that allows data model constraints to be enforced. This document content specification will be validated by developing meta-data for a variety of domains.

4.0 RESEARCH PLAN

The research approach briefly described above is more fully explained in the next three subsections.

4.1 Development and Validation of an Open-DSS Protocol

A proposal for a protocol suite for deploying and sharing Specific DSS both within and across organizations by utilizing the WWW was developed. The model proposed for the protocol suite utilizes meta-data to fully describe DSS such that an intelligent agent can easily discover them.

The proposed protocol will be validated in a study that assesses a subject's understanding of Specific DSS capabilities based solely on the meta-data. The experiment uses a set of business situations to evaluate perceived functionality for DSS developers and potential users. The following hypothesis is offered (stated in null form):

> *H_0: There will be no significant agreement between answers for DSS evaluators and those for DSS developers.*

This part of the research will proc
will be given a brief description of the 37
protocol and asked to create protocol con
the second stage of the experiment, the st
that are likely to be related to their DSS.
for solving or contributing to the solution
business problems will then be given to a
the applicability of the DSS to the busines
answers for the developers and evaluators
agreement between the two groups. The subjects will also answer questions aimed at gathering data about the meta-data creation process, the completeness of the specification, and the clarity of the specification.

4.2 Formalizing a Web Data Model

The second stage of the research will involve the development of a formal data model that will allow web based content to be maintained in a manner consistent with organizational quality goals for the web content. It will offer a framework for defining web object dependencies and their representations. It will begin with a discussion of the concepts and constructs necessary for maintaining content quality in an organization's web-space. Next, a formal definition of a web-space data model will be developed.

Under the proposed model any web object, ***o***, can be defined in terms of driver objects, some other object in the web space. The following formalism is proposed to describe web object inter-data dependency [eg. Taha, Helal and Ahmed, 1997]:

- ***D*** is the description of the driver web object. It includes the web page URL, the location of the driver object within the web page (paragraph 3, word 3), the data type and range of expected values.
- ***P*** is the dependency predicate, which describes the relationship between the web object and its drivers. It shows the detailed method for constructing the web object value.
- ***C*** is the consistency requirements for the web object. These can vary depending on the purpose for which the content is being maintained.
- ***A*** is the procedures that should be activated to restore the web object's consistency according to ***C***.

Web objects can be in one of three states, consistent, tolerated, or inconsistent. This depends on ***P*** and ***C*** as defined for each ***O***. If ***P*** is violated due to an update of the driver web page then ***O*** is no longer consistent with ***D***. If this occurs and ***C*** is not violated, then ***O*** moves into a tolerated state. If both ***P*** and ***C*** are violated, then ***O*** is in an inconsistent state and the appropriate ***A*** is triggered. In the web domain, the ***A*** taken will depend on the nature of the changes to ***D***. If ***D*** remains the same data type and falls within the range of expected values, then update of ***O*** can be made automatically. However, if ***D*** changes dramatically, or if the structure of the driver web page web changes, or if the driver web page is deleted, then ***A*** involves notification (via email or log file) of the inconsistent state.

These constructs, operations and integrity rules can be used to help maintain web space content quality. These rules should make it possible to automate many of the processes necessary to update web-based content.

This data model will be validated by a set of experts familiar with database systems and the WWW.

4.3 Development and Validation of a Web Content Quality Protocol

The research will conclude by exploring the concept of using DAI to improve an organization's ability to maintain the quality of individual web pages. It will contain a proposal for a protocol that will facilitate the maintenance of web-based content by allowing meta-data about the pages and the relationship between the pages to be distributed with the web pages. This will be accomplished by combining the meta-data content labeling concept with meta-data derived from the data model developed in the prior phase of the research. The XML Web-Quality protocol will allow intelligent agents to translate this meta-data into appropriate action.

The proposed protocol will be validated in a study that assesses the applicability of the protocol for maintaining a variety of web sites. In this study, specific web pages from a variety of web domains will be identified. Identical web pages will then be created with the addition of appropriate meta-data. Finally, the original web pages and web pages containing meta-data will be tracked for a period of three months to determine which set of sites maintain higher levels of quality. The following hypothesis is offered (stated in null form):

H_0: There will be no significant difference between the timing of updates to web content for Web sites containing meta-data and identical sites that do not contain meta-data.

5.0 EXPECTED CONTRIBUTION

This development effort represents a new approach to managing web content. The meta-data labels developed as a part of this research can help improve all six quality dimensions presented in Table 2. Content information labels help to improve the *accessibility* of web information by improving an end-user's ability to locate specific content or resources that are available on the WWW. It is also possible for content labels to provide a measure of the *completeness* of a specific web resource. Meta-data maintenance labels can be used to insure the *currency*, *accuracy* and *consistency* of web content. Both types of meta-data systems should help improve the *believability* of web content because end-users will have access to the meta-data labels and will be able to use those labels to judge the likelihood that a specific piece of web content is reliable.

The power of the distributed meta-data approach to managing web data quality is that it distributed with actual web pages. This provides universal access to the content and maintenance information and contributes to the disintermediation of the WWW. These meta-data labels allow individuals or automated intelligent agents to more readily find specific information about web page content or to update the distributed web content. The distributed nature of the approach can help increase the performance speed for the system, especially for large web-sites. In addition, using a distributed approach provides greater flexibility to local webmasters and web page designers increasing local autonomy, as well as increasing expandability of the system [Ozsu and Valdurez, 1991].

6.0 POTENTIAL DIFFICULTIES

The first challenge will be to develop a data model that is judged sufficient to improve web data quality while being tractable to implement in the web domain. A second challenge involves identifying web sites to develop meta-data for that are rich enough to demonstrate the capabilities of the system, yet not too complex to be unmanageable. On balance these difficulties can be overcome within available time.

7.0 REFERENCES

Acerman, Billsus, Gaffney, Hettich, Khoo, Kim, Klefstad, Lowe, Ludeman, Muramatsu, Omori, Pazzani, Semler, Starr, and Yap, "Learning Probabilistic User Profiles - Applications for Finding

Interesting Web Sites, Notifying Users of Relevant Changes to Web Pages, and Locating Grant Opportunities, *AI Magazine* (18:2), Summer 1997, pp. 47-56.

Bond, A. H., and Gasser, L. (eds) *Readings in Distributed Artificial Intelligence,* Morgan Kaufmann, 1982.

Bowman, C. Mic, Danzig, Peter B., Manber, Udi and Schwartz, Michael F., "Scaleable Internet Resource Discovery," *Communications of the ACM* (37:8), August, 1994, pp. 98-107+.

Burke, Robin D., Hammond, Kristian J., Kulyukin, Vladimir, Lytinen, Steven L., Tomuro, Noriko and Schoenberg, Scott, "Question Answering from Frequently Asked Question Files," *AI Magazine* (18:2), Summer 1997, pp. 57-66.

Codd, E. F., "A Relational Model for Large Shared Data Banks," *Communications of the ACM* (13:6), June 1970, pp. 377-387.

Elmasri, Ramez and Navathe, Shamkant, *Fundamentals of Database Systems,* The Benjamin/Cummings Publishing Company, Redwood City, California, 1994.

Etzioni, Oren, "Moving Up the Information Food Chain: Deploying Softbots on the World Wide Web," *AI Magazine* (18:2), Summer 1997, pp. 11-18.

Gregg, Dawn G., and Goul K. Michael, A Proposal for an Open DSS Protocol, *Forthcoming in Communications of the ACM,* 1999.

Goul, M.; Philippakis, A., Kiang, M.; Fernandes, D.; and Otondo, R., "Requirements for the Design of a Protocol Suite to Automate DSS Deployment on the World Wide Web: A Client/Server Approach," *Decision Support Systems* (19:3), March 1997, pp. 151-170.

Huh, Y. U., Keller, F. R., Redman, T. C. and Watkins, A. R., "Data Quality," *Information and Software Technology* (38:2), October 1990, pp. 559-565.

Ingham, David, Caughey, Steve and Little, Mark," Fixing the Broken Link Problem: the W3Objects Approach," *Computer Networks and ISDN Systems* (8:11) May 1996, pp. 1255-1268.

Kautz, Henry, Selman, Bart and Shah, Mehul, "The Hidden Web," *AI Magazine* (18:2), Summer 1997, pp. 27-36.

Kautz, Henry; Selman, Bart and Shah, Mehul, "Referral Web: Combining social networks and collaborative filtering," *Communications of the ACM* (40:3), March 1997, pp. 63-65.

Kehoe, Colleen, Pitkow, Jim and Rogers, Juan, *GVU's Ninth WWW User Survey Report,* Georgia Tech Research Corporation: Atlanta, Georgia http://www.gvu.gatech.edu/user_surveys/survey-1998-04/, April 1998.

Krulwich, Bruce, "Lifestyle Finder: Intelligent User Profiling Using Large-scale Demographic Data," *AI Magazine* (18:2), Summer 1997, pp. 37-56.

Malone, Thomas W., and Crowston, Kevin, "The interdisciplinary study of coordination," *ACM Computing Surveys* (26:1), March 1994, pp. 87-119.

Mecca, G., Atzeni, P., Masci, A., Merialdo, P. and Sindoni, G., "The Araneus Web-Base Management System," *SIGMOD: Proceedings of the ACM SIGMOD International Conference on Management of Data,* Laura Haas and Ashutosh Tiwary, June 1-4, 1998, pp. 544-546.

Morey, Richard C., "Estimating and Improving the Quality of Information in a MIS," *Communications of the ACM,* (25:5), May 1982, pp. 337-342.

Nielsen, Jakob, "Impact of Data Quality on the Web User Experience," *Jakob Nielsen's Alertbox,* http://www.useit.com/alertbox/980712.html, July 12, 1998.

O'Leary, Daniel E., "The Internet, Intranets, and the AI Renaissance," *Computer,* January 1997, pp. 71-78.

Ozsu, M. Tamer and Valduriez, Patrick, *Principles of Distributed Database Systems*, Prentice Hall, Englewood Cliffs, New Jersey, 1991.

Redman, Thomas C., "The Impact of Poor Data Quality on the Typical Enterprise," *Communications of the ACM* (41:2), February 1998, pp. 79-82.

Resnick, Paul, "Filtering Information on the Internet," *Scientific American,* March 1997, pp. 26-32.

Resnick, Paul and Miller, James, "PICS: Internet Access Control without Censorship," *Communications of the ACM,* (39:10), October 1996, pp. 87-93.

Reuters Air Cargo Service, http://www.racs.com/acis/general/prodov.htm, downloaded May 15, 1999.

Smith, Reid G., "The Contract Net Protocol: High-Level Communication and Control in a Distributed Problem Solver," *IEEE Transactions on Computers* (C29:12), December 1980, pp. 1104-1113.

Strong, Diane M., Yang W. Lee, and Richard Y. Wang, Data Quality in Context; *Communications of the ACM* (40:5) November 1996, pp. 103-110.

Taha, Yousry, Helal, Abdelsalm and Ahmed, Khalil M., "A Stochastic Consistency Model for Data Warehousing," *Proc. AIS Americas Conference on Information Systems*, Indianapolis, IN, August 15-17, 1997, pp.455-457.

Vinze, Ajay, Kulkarni, Uday and Gregg, Dawn, "A Scientific Framework for Conducting Software Engineering Research," Working Paper, Arizona State University.

Wand, Yair , and Wang, Richard Y., "Anchoring Data Quality Dimensions In Ontological Foundations," *Communications of the ACM* (39:11), November 1996, pp. 86-95.

Wang, Richard Y., "A Product Perspective on Total Data Quality Management," *Communications of the ACM* (41:2), February 1998, pp. 58-65.

Wang, Richard Y., Reddy, M.P. and Kon, Henry B., "Toward Quality Data: An Attribute-based Approach," *Decision Support Systems* (13), 1995, pp. 349-372.

Wang, Richard Y. and Strong, Diane M., "Beyond Accuracy: What Data Quality Means To Data Consumers," *Journal of Management Information Systems* (12:4), 1996, pp. 5-34.

Wang, Richard Y., Storey, V.C. and Firth, C.P., "A Framework For Analysis Of Data Quality Research," *IEEE Transactions on Knowledge and Data Engineering* (7:4), 1995, pp. 623-640.

Wiederhold, G., Wegner, P. and Ceri, S. "Towards Metaprogramming," *Communications of the ACM* (35:11), November 1992, pp. 88-99.

Yahoo - Web Index, http://www.yahoo.com, *downloaded August 30, 1998*