Do Metadata Models meet IQ Requirements?

Felix Naumann*
Humboldt-Universität zu Berlin
Unter den Linden 6
D-10099 Berlin
Germany
naumann@dbis.informatik.hu-berlin.de

Claudia Rolker
Forschungszentrum Informatik (FZI)
Haid-und-Neu-Str. 10-14
D-76131 Karlsruhe
Germany
rolker@fzi.de

Abstract

Research has recognized the importance of analyzing information quality (IQ) for many different applications: The success of data integration greatly depends on the quality of the individual data. In statistical applications poor data quality often leads to wrong conclusions. High information quality is literally a vital property of hospital information systems. Poor data quality of stock price information services can lead to economically wrong decisions.

Several projects have analyzed this need for IQ metadata and have proposed a set of IQ criteria or attributes which can be used to properly assess information quality. In this paper we survey and compare these approaches. In a second step we take a look at existing prominent proposals of metadata models, especially those on the Internet. Then, we match these models to the requirements of information quality modeling. Finally, we propose a quality assurance procedure for the assurance of metadata models.

1 Introduction

The quality of information is becoming increasingly important, not only because of the rapid growth of the Internet (and its implication for the information industry). Also the anarchic nature of the Internet has made industry and researchers aware of this issue. As awareness of quality issues amongst information professionals grow, their demands for high

*This research was supported by the German Research Society, Berlin-Brandenburg Graduate School in Distributed Information Systems (DFG grant no. GRK 316).

quality information will increase. There is a clear need for the industry to respond to these requirements and this also represents a genuine market opportunity [Inf95].

The autonomy of WWW information sources prevents information seekers from directly controlling the quality of the information they receive. Rather, users of such information sources must resort to analyzing the quality of the information once it is retrieved and use the analysis for future queries. Research has recognized the importance of analyzing information quality (IQ) for many different applications [WS96, Red98]. As a result, several projects have emerged to find a general measure for information quality. While the application domains differ from structured multidatabases or data warehouse applications to retrieval systems for unstructured information, the approaches to measure IQ are all similar: Domain experts define a set of IQ criteria that are deemed to be important to the field, or a general set such as that of Wang and Strong [WS96] is chosen. Next, assessment methods for each criterion are developed. These methods include questionnaires for subjective criteria, calibration methods, etc. Finally some way of summarizing the results is given, so one is able to qualitatively compare whole sources, query execution plans, or pieces of information. All approaches heavily rely on metadata, especially quality metadata. IQ criteria are of no use if no score for them is found. A dimension which cannot be assessed does not contribute to a comparison of sources.

On the other hand information providers have recognized the need to describe the products they offer and provide this metadata. Obviously this provider metadata will not directly address IQ. No information source will admit their information or data to be outdated or inaccurate. It rather covers aspects of authorship, title, etc. Such particulars can only be evaluated to indirectly find IQ ratings. The creation date of a document reveals its age, the publisher may have a good or bad reputation etc.

Our goal is to bridge the gap between IQ metadata requirements and actual metadata that is already provided by many sources. To this end we first analyze the most important proposed sets of IQ criteria, i.e., the "wish list" of information brokers and information consumers (Section 2). The next section will take a look at the most wide-spread metadata

models that already exist and are used by many providers (Section 3). The main contribution of this paper is a comparison of the IQ metadata requirements with the metadata models. We show how IQ criteria can be derived from existing metadata (Section 4). The paper ends with a proposal to let metadata registries assure the quality of metadata models in the future (Section 5), and with a further outlook onto certification authorities for metadata instances with respect to their quality (Section 6).

2 Information Quality Metadata Requirements

This section will review several projects concerned with information quality. Some provide research from a global viewpoint and define IQ in a very general way. Others have concentrated either on certain quality aspects or on certain application domains for IQ. All reviewed projects have in common, that IQ is defined as some set of quality criteria, i.e., that quality is made up of many facets. All projects face the problem of assessing values for the criteria. In the scope of this work, we view these criteria as metadata for the data being analyzed. Thus, a list of criteria can be viewed as metadata requirements, or a "wish list" of criteria one would like to evaluate.

What follows is a short summarization of the mentioned projects. Instead of listing each set in each section, we have summarized the IQ criteria of the projects in Table 1. The actual criteria names may slightly differ, but have been adapted appropriately. We have classified the criteria into four sets: *Content-related* criteria concern the actual information that is retrieved. *Technical* criteria measure aspects that are determined by soft- and hardware. *Intellectual* criteria are made up of very subjective criteria like believability. *Instantiation-related* criteria concern the presentation of the information.

2.1 TDQM

Total Data Quality Management is a project at MIT, aimed at providing an empiric foundation for data quality. Wang and Strong have empirically identified fifteen IQ criteria regarded

by data consumers as the most important [WS96]. The authors have classified these criteria into "intrinsic quality", "accessibility, "contextual quality", and "representational quality". Their framework has already been used effectively in industry and government. To our best knowledge this is the only empirical study in this field, and has thus often been used as a research basis for other projects (see below).

2.2 IQ criteria for molecular biology information systems

Based on the criteria of the TDQM model we have adapted the set to suit the integration of molecular biology information systems (MBIS) in a mediator-based architecture [NLF99]. Due to the nature of this architecture and the underlying relational model the TDQM criteria were modified: Two criteria (response time and price) were added to account for the Internet setting of the approach, some criteria were interpreted in a new manner to account for the integration aspect of the approach. Criteria such as objectivity or concise representation were dropped since in a relational data model a query result is simply a table.

For the process of planning queries against such a distributed and heterogeneous system three classes of criteria were distinguished: Source-specific, query-specific and attribute-specific criteria.

2.3 Notions of service quality

Weikum has developed a different classification of IQ-criteria [Wei99]: He distinguishes system-centric, process-centric, and information-centric criteria. The set of criteria in [Wei99] was put together in an informal manner with no claim for completeness. However in our eyes, Weikum does provide several new criteria such as 'latency', which play an increasingly important role in new information systems, especially in WWW settings. Each criterion is thoroughly discussed, again in an informal manner.

2.4 DWQ

Data Warehouse Quality (DWQ) is an Esprit funded project to analyze the meaning of data quality for data warehouses and to produce a formal model of information quality to enable design optimization of data warehouses [JV97]. Again the approach is based on the empirical studies of Wang and Strong [WS96]. However, the focus lies on data warehouse specific aspects such as the quality of aggregated data. The authors develop a model for IQ metadata management in a data warehouse setting.

2.5 SCOUG

Measurement of the quality of databases was the subject of the Southern California Online User Group (SCOUG) Annual Retreat in 1990. The brainstorming session resulted in a checklist of criteria which fall into 10 broad categories [Bas90]. These criteria are the mostly referenced ones within the database area. Although the focus lies on the evaluation of database performance (including categories like documentation and customer training) its similarity to the above described quality measures is obvious.

2.6 Chen et al.

With a focus on World Wide Web query processing, Chen et al. propose a set of quality criteria from an information server viewpoint [CZW98]. In their setting a user can specify quality requirements along with the query. Under heavy workload, the WWW server must then simultaneously process multiple queries and still meet the quality requirements. To this end, the authors present a scheduling algorithm that is based on the time-relevant criteria such as response time or network delay. The other IQ criteria are only briefly discussed.

Category	IQ Criteria	TDQM	MBIS	Weikum	DWQ	SCOUG	Chen
Content-related Criteria	Accuracy	Yes	Yes	Yes	Yes	Yes	Yes
	Documentation					Yes	
	Relevancy	Yes	Yes		Yes		Yes
	Value-Added	Yes				Yes	
	Completeness	Yes	Yes	Yes	Yes	Yes	Yes
	Interpretability	Yes			Yes		
Technical Criteria	Timeliness	Yes	Yes	Yes	Yes	Yes	Yes
	Reliability			Yes			
	Latency			Yes			Yes
	Performability			Yes		Yes	
	Response time		Yes	Yes			Yes
	Security	Yes		Yes	Yes		
	Accessibility	Yes	Yes	Yes	Yes	Yes	
	Price		Yes	Yes		Yes	
	Customer Support					Yes	
Intellectual Criteria	Believability	Yes	Yes	Yes	Yes	Yes	
	Reputation	Yes	Yes		Yes		
	Objectivity	Yes					
Instantiation related Criteria	Verifiability			Yes			
	Amount of data	Yes	Yes				Yes
	Understandability	Yes	Yes				
	Concise represent.	Yes					
	Consistent represent.	Yes	Yes	Yes	Yes	Yes	

Table 1: Metadata Requirements for Information Quality

3 Metadata Models

Metadata models have been developed for many different purposes. One of the first applications was that of modeling bibliographic information for libraries. Recently the problem of describing information in general through metadata has received much attention. The abundance of information that is nowadays accessible through the Internet and WWW makes it necessary to describe the provided information in a concise, uniform and easily understandable, and interpretable way. Without such a description, an information seeker will drown in non-relevant information and may even not find the desired information, even though it is available.

In the following sections we present several projects that attempt to set up a common metadata model for WWW information in documents and gain general acceptance in the Internet community. We have tried to cover the most important projects, and have summa-

rized the attributes of these metadata models in Table 2. The attribute names may slightly differ, but have been adapted appropriately.

3.1 Dublin Core

The Dublin Core Metadata initiative has developed a metadata element set intended to facilitate the discovery of electronic resources [Dub99]. It evolved from a series of workshops with participants from many different application domains. The element set is wide spread across many types of information systems, from digital libraries to museums and many other electronic document collections. Dublin Core is especially wide-spread in HTML-Documents where the META tag is used: `<META NAME="DC.Title" CONTENT= "MyTitle">`

3.2 STARTS

In the Stanford Proposal for Internet Meta-Searching (STARTS) project a list of required metadata fields for documents is proposed [GCGM97]. It is based on the *use attributes* of Z39.50/GILS (see Sections 3.3 and 3.4). The list was developed by researchers and practitioners from large Internet companies in a number of workshops. In 1997 the Dublin Core standard (see Section 3.1) was integrated.

STARTS also proposes a list of metadata fields to describe the query capabilities of an information source. These fields help solving the problems of source selection and rank-merging the results. While this metadata may also be relevant to assessing IQ in some situations, it is not considered here.

3.3 Z39.50 Attribute Set BIB-1

Z39.50 is an ANSI and ISO standard that describes the communication between a client and a metadata server mainly with respect to searching. Originally, it was developed for the communication interoperability of libraries.

Z39.50 is independent of any application area. A profile specifies how to use the various

functions defined by Z39.50 in a specific application area. A profile also specifies which attribute set to use. The Attribute Set BIB-1 [Z3995] describes bibliographic metadata and comprises 100 attributes. BIB-1 allows to describe bibliographic data by several identification schemas and keyword lists. Each schema/keyword list corresponds to one BIB-attribute, e.g., there are 13 subject attributes each of them referring to a different keyword list. In Table 2 these attributes are summarized in content.

3.4 Z39.50 Profile GILS

GILS [Eli99a] stands for Global Information Locator Service or for Government Information Locator Service. Originally, the latter one was understood under this synonym and was developed from an initiative in the United States. The Environment and Natural Resources Management Project of the G7 adopted the Government Information Locator Service as a model for the Global Information Locator Service. From the perspective of standards and technology there is no difference between them.

GILS is not only a means to describe books or datasets, but also to provide information about people, events, meetings, artifacts, rocks etc. The Z39.50 Profile Version 2 comprises 91 attributes [Eli99b]. The level of these attributes is very detailed and so they are summarized in content in Table 2, e.g., 12 GILS attributes correspond to the distributor attribute in Table 2.

3.5 DIF

The Directory Interchange Format (DIF) was originally developed to make scientific, US-governmental catalogues describing data groups interoperable [Glo93, Ols99]. DIF consists of 25 data fields, 6 of them are mandatory.

In a number of workshops the DIF-standard was developed and based on it the data catalogue "Global Change Master Directory" (GDMC) was created. Today the GDMC staff is the maintenance agency of the DIF-standard.

	Dublin Core	STARTS	BIB	GILS	DIF
Title	Yes	Yes	Yes	Yes	Yes
Author or Creator	Yes	Yes	Yes	Yes	Yes
Subject and Keywords	Yes		Yes	Yes	Yes
Description	Yes		Yes	Yes	Yes
Publisher/Distributor	Yes		Yes	Yes	Yes
Other Contributor	Yes				
Date	Yes	Yes	Yes	Yes	Yes
Last Review Date					Yes
Future Review Date					Yes
Resource Type	Yes		Yes		
Format	Yes				
Storage Medium				Yes	Yes
Resource Identifier	Yes	Yes	Yes	Yes	Yes
Identifier Type		Yes	Yes	Yes	
Cross References		Yes	Yes	Yes	Yes
Source	Yes			Yes	
Language	Yes	Yes	Yes	Yes	
Relation	Yes		Yes	Yes	
Coverage	Yes		Yes	Yes	Yes
Rights Management	Yes			Yes	
Document-text		Yes	Yes		
Sensor name					Yes
Parameter measured					Yes
Quality Assurance Method					Yes

Table 2: Metadata Attribute Proposals

4 Matching Requirements and Metadata Models

Having introduced both a number of desired IQ criteria sets and a number of metadata attribute sets currently in use, the question arises where and how well they meet. Is it possible to derive values for the IQ criteria from existing metadata? The answer unfortunately is 'no', at least not in a straight-forward manner. The following section discusses how and how well metadata attributes help in determining IQ criteria scores. We do not examine each criterion in detail but look into a few exemplary criteria – one from each class of Table 1. Similar arguments hold for the other criteria of the respective class.

Relevancy. Wang and Strong define relevancy as "the extent to which data are applicable and helpful for the task at hand." [WS96]. Relevancy is an often used criterion in the field of information retrieval. A document or piece of information is considered to be relevant to the query, if the keywords of the query appear often and/or in prominent positions in the document. Thus, the metadata attributes Coverage, Title, Subject/Keywords, and Description are of help in determining Relevancy. Especially Title and Subject/Keywords explicitly point out prominent representatives of the information content.

Even with the help of these attributes, determining the relevancy of information is error-prone: For instance a query for the term "jaguar" at any WWW search engine will retrieve document links both for the animal and the automobile. If the user had the animal in mind, the links to automobile sites should have been considered as not relevant.

Response Time. The response time criterion measures the delay between submission of a query by the user and reception of the complete response from the information system. The score for this criterion depends on unknown factors such as network traffic, server workload etc. These aspects are hardly predictable. Another factor is the type and complexity of the user query. Again this cannot not be predicted, however, it can be taken into account, once the query is posed and a query execution plan is developed.

A third aspect plays an important role: the technical equipment of the information server. Metadata on the equipment can be derived from the Publisher attribute and the Storage Medium attribute. Storage Medium can directly be translated to some time factor. To derive a factor from the Publisher attribute, further investigations on the publishers hardware and software are necessary, for instance by directly contacting the publisher/web-site provider.

Concluding, existing metadata attributes hardly contribute to the response time criterion. A more realistic approach to determine the scores is to (a) keep statistics on previous queries and (b) employ calibration techniques as proposed in [Spi96].

Believability. When querying autonomous information sources believability is an especially important criterion. Apart from simply providing information, a source must convince the

user, that this information is "accepted or regarded as true, real, and credible" [WS96].

The main source for believability is the author or creator of the information. Thus, the Author/Creator and the Contributor attributes are helpful in determining a score. However, this cannot be done automatically. First, a user defined mapping of authors to believability scores must be created. Obviously this mapping is very subjective and must be newly created for each user.

Determining IQ scores for all intellectual criteria is a very difficult task. Not only are these criteria of extremely subjective nature. Also, one must assume that information sources will be very resourceful trying to find ways to improve believability without improving the correctness of the information itself. A common authority as proposed in the next section might help determine and control the scores.

Verifiability. When believability is not as high as it could be, the quality of information can greatly improve, if it is verifiable through a second source. The verification process can be supported by the attributes Resource Identifier, Relation, and Cross References. Relation and cross references may point to another source, where the information can be verified. A global identifier will help identification of the object or information in that other source, where it can be verified. Thus, the content of the attributes do not directly contribute to verifiability, but their existence does improve information quality.

Figure 1 summarizes the discussion above and additionally gives matches for all criteria not examined. Similar considerations have led to the each of the matchings.

5 Quality Assurance by Metadata Registries - a Proposal

Metadata registries are set up to avoid multiple development of similar metadata schemata and to ensure interoperability between the metadata schemata at both syntactic and se-

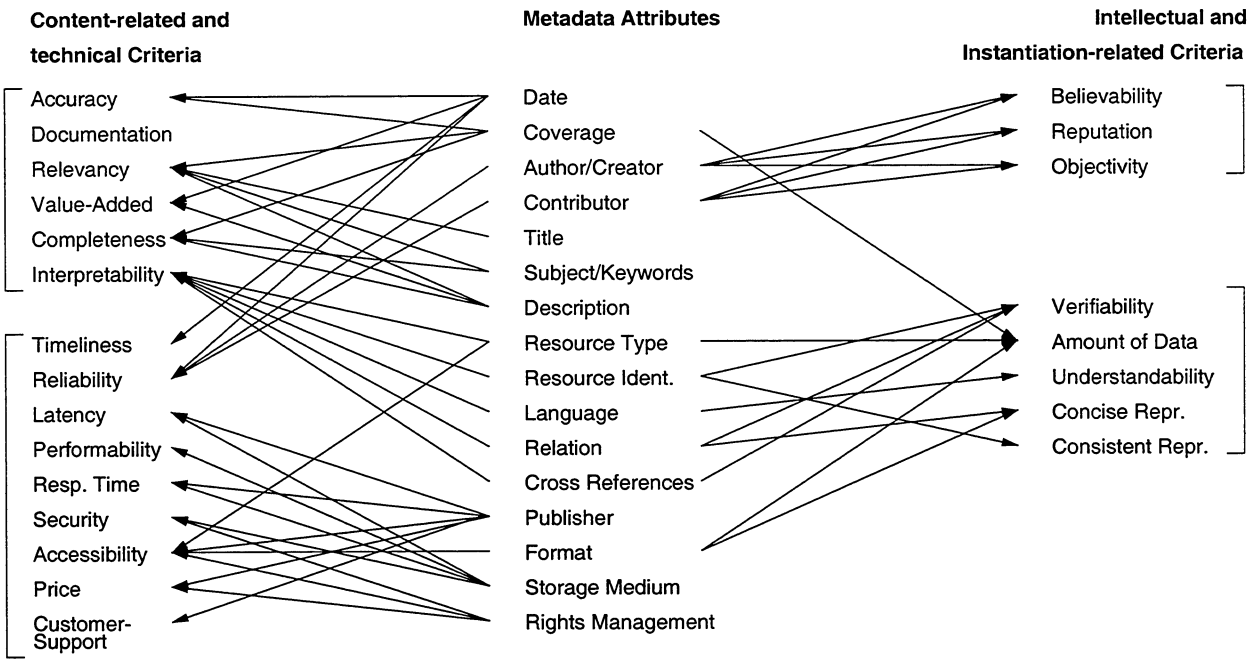

Figure 1: Matching required IQ Criteria and existing general-purpose Metadata attributes

mantic levels. In [Gai99] a metadata registry is defined as 'a publicly accessible system that records the semantics, structure and interchange formats of any type of metadata. A formal authority, or agency, maintains and manages the development and evolution of a metadata registry. The authority is responsible for policies pertaining to registry contents and operation.'

There are some metadata registries already running on the Web, for instance Metadata.Net [Dis99] or ROADS [Mic99]. Moreover, standardization organizations are currently developing a framework for metadata registries [D-L98, Fra99, Fra97].

Each metadata registry expects the metadata to be described in a standardized schema language like the following:

- An important member of these specifications is XML. XML is the 'Extensible Markup Language' [Wor99b] (extensible because it is not a fixed format like HTML). It is designed to enable the use of SGML on the World Wide Web. SGML is the Standard Generalized Markup Language (ISO 8879), the international standard for defining descriptions of the structure and content of different types of electronic document.

Documents types are specified through Document Type Definitions (DTDs). A DTD is a file (or several files to be used together), written in XML, which contains a formal definition of a particular type of document. We propose to include attributes for quality metadata in such a definition. The simple structure of a DTD will then allow to easily evaluate the quality of a source or document.

- The Platform for Internet Content Selection (PICS) specifies a labeling infrastructure to enhance HTML headers [Wor99a]. While it was originally created to attach ratings to WWW material that is inappropriate for children, the approach has been adapted to support various metadata tasks. PICS is supported by the W3 Consortium.

  Again the inclusion of additional attributes for quality metadata can assist in finding and selecting relevant information or documents.

- The Resource Description Framework (RDF) is an infrastructure that enables the encoding, exchange and reuse of structured metadata and is an application of XML [Wor99c]. It additionally provides a means for publishing both human-readable and machine-processable vocabularies designed to encourage the reuse and extension of metadata semantics among disparate information communities.

  RDF imposes needed structural constraints to provide unambiguous methods expressing semantics.

We propose that metadata registries should not only register metadata, but also should have an eye on the usefulness of the registered metadata models towards quality reasoning. The aim should be that all registered metadata models fulfill a certain level of quality by requiring a minimal set of quality criteria. Once this measure is implemented, information seekers will greatly profit from the new metadata: For instance, users will be able to choose between an accurate but somewhat slow information source and one that is fast but inaccurate to a certain degree. Information systems that integrate many sources (meta information systems) will also benefit, since they could combine sources in a way that produces qualitatively better results, and not arbitrarily combining sources as it is done today.

6 Conclusions and Outlook

With the help of metadata registries quality assurance of metadata models can be reached, as during the registration process the metadata developer could be forced to show that his metadata model covers the IQ criteria. Having quality assured metadata models is one step, but it is also important that all metadata instances of a registered and quality assured metadata model provide values for these attributes. Of course, these values must be correct and believable. To this end, a certification authority is needed which takes care of the quality of the produced metadata instances. The examination of instances with respect to their quality is an unsolved problem and can probably only be achieved by carrying out spot checks. While this procedure may seem expensive, the benefits of accessing and using certified quality information are obvious.

Whereas metadata registries and schema languages for the description of metadata exist, the task of quality assurance executed by registries and the need for quality certifying authorities are still an issue. Without such a centralized control, WWW information system designers and users must rely on the somewhat inaccurate and subjective methods described in Section 4.

Concluding, there is a long way to go for metadata models until they meet the requirements to evaluate information quality. On the other hand, it is inevitable that quality analyzers must compromise in their need for metadata. A middle ground may be provided by metadata registries. These authorities can combine and match the desires of users or systems requiring high quality information on the one side and the possibilities of the information providers on the other side.

References

[Bas90] Reva Basch. Measuring the Quality of the Data: Report on the Fourth Annual SCOUG Retreat. *Database Searcher*, 6(8):18–24, October 1990.

[CZW98] Ying Chen, Qiang Zhu, and Nengbin Wang. Query processing with quality control in the World Wide Web. *World Wide Web*, 1 (4):241–255, 1998.

[D-L98] D-Lib Magazine. Metadata Registries Workshop, April 15 - 17, 1998,Washington, DC, Summary . http://mirrored.ukoln.ac.uk/lis-journals/dlib/dlib/dlib/may98/05clips.h%tml, May 1998.

[Dis99] Distributed Systems Technology Centre . Metadata.Net: Metadata Schema Registry and Metadata Tools & Services . http://metadata.net/, June 1999.

[Dub99] Dublin Core Metadata Initiative. http://purl.org/dc/index.htm, 1999.

[Eli99a] Eliot Christian , U.S. Geological Survey. GILS FAQ. http://www.gils.net/faq.html, 1999.

[Eli99b] Eliot Christian , U.S. Geological Survey. GILS Metadata Elements. http://www.gils.net/element2.html#table_2, 1999.

[Fra97] Frank Olken. Workshop Report: Joint Workshop on Metadata Registries . http://www.lbl.gov/~olken/EPA/Workshop/report.html, Dec 1997.

[Fra99] Frank Olken and John McCarthy . Metadata Registries: Averting a Tower of XML Babel . http://www.lbl.gov/~olken/mendel/w3c/papers/xtech99/abstract.html, January 1999.

[Gai99] Gail Clement and Pete Winn. Dublin Core User's Guide Glossary. http://webster.effem.com/dublin/glossary.htm, June 1999.

[GCGM97] Luis Gravano, Chen-Chuan K. Chang, and Hector Garcia-Molina. STARTS: Stanford proposal for internet meta-searching. In *Proc. of the ACM SIGMOD Conference*, 1997.

[Glo93] Global Change Master Directory . Directory Interchange Format (DIF) Manual. ftp://nssdca.gsfc.nasa.gov/MD_DOC/DIFMANUAL.PS2;1, April 1993.

[Inf95] Information Market Observatory (IMO). The Quality of Electronic Information Products and Services . http://www2.echo.lu/impact/imo/9504.html, September 1995.

[JV97] M. Jarke and Y. Vassiliou. Data warehouse quality design: A review of the DWQ project. In *Proc. 2nd Conference on Information Quality*, MIT, Boston, 1997.

[Mic99] Michael Day. ROADS Metadata Registry . http://www.ukoln.ac.uk/metadata/roads/templates/, Feb 1999.

[NLF99] Felix Naumann, Ulf Leser, and Johann Christoph Freytag. Quality-driven integration of heterogenous information systems. In *Proc. of the Int. Conf. on Very Large Dtabases*, Edinburgh, UK, 1999.

[Ols99] Olsen (Global Change Master Directory) . Directory Interchange Format (DIF) Writer's Guide, Version 7 . http://gcmd.gsfc.nasa.gov/difguide/difman.html, May 1999.

[Red98] Thomas C. Redman. The impact of poor data quality in the typical enterprise. *Communications of the ACM*, 41(2):79–82, 1998.

[Spi96] Myra Spiliopoulou. A calibration mechanism identifying the optimization technique of a multidatabase participant. In *Proc. of the Conf. on Parallel and Distributed Computing Systems (PDCS)*, Dijon, France, Sept. 1996.

[Wei99] Gerhard Weikum. Towards guaranteed quality and dependability of information systems. In *Proc. of the Conf. Datenbanksysteme in Büro, Technik und Wissenschaft*, Freiburg, Germany, 1999.

[Wor99a] World-Wide Web Consortium. Platform for Internet Content Selection. `http://www.w3.org/PICS/`, 1999.

[Wor99b] World-Wide Web Consortium (W3C) . Extensible Markup Language (XML) . `http://www.w3.org/XML/`, June 1999.

[Wor99c] World-Wide Web Consortium (W3C) . RDF. `http://www.w3.org/RDF/`, 1999.

[WS96] Richard Y. Wang and Diane M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal on Management of Information Systems*, 12, 4:5–34, 1996.

[Z3995] Z39.50 Implementors Group and Z39.50 Maintenance Agency. Attribute Set BIB-1 (Z39.50-1995): Semantics. `http://lcweb.loc.gov/z3950/agency/defns/bib1.html`, Sep 1995.