Developing a Framework for Assessing Information Quality on the World Wide Web

Shirlee-ann Knight and Janice Burn
Edith Cowan University, Perth, Australia

s.knight@ecu.edu.au, j.burn@ecu.edu.au

Abstract

The rapid growth of the Internet as an environment for information exchange and the lack of enforceable standards regarding the information it contains has lead to numerous information qual ity problems. A major issue is the inability of Search Engine technology to wade through the vast expanse of questionable content and return "quality" results to a user's query. This paper attempts to address some of the issues involved in determining what quality is, as it pertains to information retrieval on the Internet. The IQIP model is presented as an approach to managing the choice and implementation of quality related algorithms of an Internet crawling Search Engine.

Keywords: Information Quality, IQIP, Data Quality, Information Retrieval, Search Engines

Introduction – The Big Picture

Over the past decade, the Internet[1] – or World Wide Web (Technically the Internet is a huge collection of networked computers using TCP/IP protocol to exchange data. The World-wide Web (WWW) is in essence only part of this network of computers, however its visible status has meant that conceptually at least, it is often used interchangeably with "Internet" to describe the same thing.) – has established itself as the key infrastructure for information administration, exchange, and publication (Alexander & Tate, 1999), and Internet Search Engines are the most commonly used tool to retrieve that information (Wang, 2001). The deficiency of enforceable standards however, has resulted in frequent information quality problems (Eppler & Muenzenmayer, 2002).

This paper is part of a research project undertaken at Edith Cowan, Wollongong and Sienna Universities, to build an Internet Focused Crawler that uses "Quality" criterion in determining returns to user queries. Such a task requires that the conceptual notions of quality be ultimately quantified into Search Engine algorithms that interact with Webpage technologies, eliminating documents that do not meet specifically determined standards of quality.

The focus of this paper, as part of the wider research, is on the concepts of Quality in Information and Information Systems, specifically as it pertains to Information and Information Retrieval on the Internet. As with much of the research into Information Quality (IQ) in Information Systems, the term is interchangeable with Data Quality (DQ).

Material published as part of this journal, either online or in print, is copyrighted by the publisher of Informing Science. Permission to make digital or paper copy of part or all of these works for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial advantage AND that copies 1) bear this notice in full and 2) give the full citation on the first page. It is permissible to abstract these works so long as credit is given. To copy in all other cases or to republish or to post on a server or to redistribute to lists requires specific permission and payment of a fee. Contact Editor@inform.nu to request redistribution permission.

Editor: Eli Cohen

What Is Information Quality?

Data and Information Quality is commonly thought of as a multi-dimensional concept (Klein, 2001) with varying attributed characteristics depending on an author's philosophical view-point. Most commonly, the term "Data Quality" is described as data that is "Fit-for-use" (Wang & Strong, 1996), which implies that it is relative, as data considered appropriate for one use may not possess sufficient attributes for another use (Tayi & Ballou, 1998).

IQ as a series of Dimensions

Table 1 summaries 12 widely accepted IQ Frameworks collated from the last decade of IS research. While varied in their approach and application, the frameworks share a number of characteristics regarding their classifications of the dimensions of quality.

Table 1: Comparison of Information Quality Frameworks

Yr	Author	Model	Constructs		
1996	[Wang & Strong, 1996]	A Conceptual Framework for Data Quality Summary: » 4 Categories » 16 Dimensions	**Category**	**Dimension**	
			Intrinsic IQ	Accuracy, Objectivity, Believability, Reputation	
			Accessibility IQ	Accessibility, Security	
			Contextual IQ	Relevancy, Value-Added, Timeliness, Completeness, Amount of Info	
			Representational IQ	Interpretability, Ease of Understanding, Concise Representation, Consistent Representation	
	[Zeist & Hendriks, 1996]	Extended ISO Model Summary: » 6 Quality characteristics » 32 Sub-characteristics	**Characteristics**	**Sub-characteristics**	
			Functionality	Suitability, Accuracy, Interoperability, Compliance, Security, Traceability	
			Reliability	Maturity, Recoverability, Availability, Degradability, Fault tolerance	
			Efficiency	Time behaviour, Resource behaviour	
			Usability	Understandability, Learnability, Operability, Luxury, Clarity, Helpfulness, Explicitness, Customisability, User-friendliness	
			Maintainability	Analysability, Changeability, Stability, Testability, Manageability, Reusability	
			Portability	Adaptability, Conformance, Replaceability, Installability	
1999	[Alexander & Tate, 1999]	Applying a Quality Framework to Web Environment Summary: » 6 Criteria	**Criteria**	**Explanation**	
			Authority	validated information, author is visible	
			Accuracy	reliable, free of errors	
			Objectivity	presented without personal biases	
			Currency	content up-to-date	
			orientation	clear target audience	
			navigation	Intuitive design	
	[Katerattanakul et al, 1999]	IQ of Individual Web Site Summary: » 4 Quality Categories (adapted from Wang & Strong)	**Category**	**Dimension**	
			Intrinsic IQ	Accuracy and errors of the content Accurate, workable, and relevant hyperlinks	
			Contextual IQ	Provision of author's information	
			Representational IQ	Organisation, Visual settings, Typographical features, consistency, Vividness / attractiveness	
			Accessibility IQ	Navigational tools provided	
	[Shanks & Corbitt, 1999]	Semiotic-based Framework for Data Quality Summary: » 4 Semiotic descriptions » 4 goals of IQ » 11 dimensions	**Semiotic Level**	**Goal**	**Dimension**
			Syntactic	Consistent	Well-defined / formal syntax
			Semantic	Complete and Accurate	Comprehensive, Unambiguous, Meaningful, Correct
			Pragmatic	Usable and Useful	Timely, Concise, Easily Accessed, Reputable
			Social	Shared understanding of meaning	Understood, Awareness of Bias

<table>
<tr><td>2000</td><td>[Dedeke, 2000]</td><td>Conceptual Framework for measuring IS Quality
Summary:
» 5 Quality Categories,
» 28 dimensions</td><td>
<table>
<tr><th>Quality Category</th><th>Dimensions</th></tr>
<tr><td>Ergonomic Quality</td><td>Ease of Navigation, Confortability, Learnability, Visual signals, Audio signals</td></tr>
<tr><td>Accessibility Quality</td><td>Technical access, System availability, Technical security, Data accessibility, Data sharing, Data convertibtlity</td></tr>
<tr><td>Transactional Quality</td><td>Controllability, Error tolerance, Adaptability, System feedback, Efficiency, Responsiveness</td></tr>
<tr><td>Contextual Quality</td><td>Value added, Relevancy, Timeliness, Completeness, Appropriate data</td></tr>
<tr><td>Representation Quality</td><td>Interpretability, Consistency, Conciseness, Structure, Readability, Contrast</td></tr>
</table>
</td></tr>
<tr><td></td><td>[Naumann & Rolker, 2000]</td><td>Classification of IQ Metadata Criteria
Summary:
» 3 Assessment Classes
» 22 IQ Criterion</td><td>
<table>
<tr><th>Assessment Class</th><th>IQ Criterion</th></tr>
<tr><td>Subject Criteria</td><td>Believability, Concise representation, Interpretability, Relevancy, Reputation, Understandability, Value-Added</td></tr>
<tr><td>Object Criteria</td><td>Completeness, Customer Support, Documentation, Objectivity, Price, Reliability, Security, Timeliness, Verifiability</td></tr>
<tr><td>Process Criteria</td><td>Accuracy, Amount of data, Availability, Consistent representation, Latency, Response time</td></tr>
</table>
</td></tr>
</table>

<table>
<tr><td></td><td>[Zhu & Gauch, 2000]</td><td>Quality metrics for information retrieval on the WWW
Summary:
» 6 Quality Metrics</td><td>
<table>
<tr><th>Assessment Class</th><th>IQ Criterion</th></tr>
<tr><td>currency</td><td>measured as the time stamp of the last modification of the document.</td></tr>
<tr><td>availability</td><td>calculated as the number of broken links on a page divided by the total numbers of links it contains.</td></tr>
<tr><td>information-to-noise ratio</td><td>computed as the total length of the tokens after preprocessing divided by the size of the document:</td></tr>
<tr><td>authority</td><td>based on the Yahoo Internet Life (YIL) reviews [27], which assigns a score ranging from 2 to 4 to a reviewed site.</td></tr>
<tr><td>popularity</td><td>number of links pointing to a Web page, used to measure the popularity of the Web page</td></tr>
<tr><td>cohesiveness</td><td>determined by how closely related the major topics in the Web page are</td></tr>
</table>
</td></tr>
<tr><td>2001</td><td>[Leung, 2001]</td><td>Adapted Extended ISO Model for Intranets
Summary:
» Adaptation of Zeist & Hendriks Extended ISO Model, applied to Intranet environments
» The grey, *italic* sub-characteristics are not considered needed to achieve IQ</td><td>
<table>
<tr><th>Characteristics</th><th>Sub-characteristic</th></tr>
<tr><td>Functionality</td><td>*Suitability*, Accuracy, *Interoperability*, *Compliance*, Security, Traceability</td></tr>
<tr><td>Reliability</td><td>Maturity, Fault tolerance, Recoverability, Availability, Degradability</td></tr>
<tr><td>Usability</td><td>Understandability, Learnability, Operability, Luxury, Clarity, Helpfulness, Explicitness, User-friendliness, Customisability</td></tr>
<tr><td>Efficiency</td><td>Time behaviour, Resource behaviour</td></tr>
<tr><td>Maintainability</td><td>Analysability, Changeability, Stability, Testability Manageability, Reusability</td></tr>
<tr><td>Portability</td><td>*Adaptability*, *Installability*, *Replaceability*, *Conformance*</td></tr>
</table>
</td></tr>
<tr><td>2002</td><td>[Kahn et al,; 2002]</td><td>Mapping IQ dimension into the PSP/IQ Model
Summary:
» 2 Quality Types,
» 4 IQ Classifications,
» 16 IQ dimensions</td><td>
<table>
<tr><th>Quality Type</th><th>Classification</th><th>Dimension</th></tr>
<tr><td>Product Quality</td><td>Sound Information</td><td>Free-of-Error, Concise, Representation, Completeness, Consistent Representation</td></tr>
<tr><td></td><td>Useful Information</td><td>Appropriate Amount, Relevancy, Understandability, Interpretablility, Objectivity</td></tr>
<tr><td>Service Quality</td><td>Dependable Information</td><td>Timeliness, Security</td></tr>
<tr><td></td><td>Useable Information</td><td>Believability, Accessibility, Ease of Manipulation, Reputation, Value-Added</td></tr>
</table>
</td></tr>
<tr><td></td><td>[Eppler & Muenzenmayer, 2002]</td><td>Conceptual Framework for IQ in the Website Context
Summary:
» 2 Manifestations,
» 4 quality categories,
» 16 Quality dimensions</td><td>
<table>
<tr><th>Quality Type</th><th>Categories</th><th>Dimensions</th></tr>
<tr><td>Content Quality</td><td>Relevant Information</td><td>Comprehensive, Accurate, Clear, Applicable</td></tr>
<tr><td></td><td>Sound Information</td><td>Concise, Consistent, Correct, *Current*</td></tr>
<tr><td>Media Quality</td><td>Optimized Process</td><td>Convenient, *Timely*, Traceable, Interactive</td></tr>
<tr><td></td><td>Reliable Infrastructure</td><td>Accessible, Secure, Maintainable, *Fast*</td></tr>
</table>
</td></tr>
</table>

	[Klein, 2002]	5 IQ Dimensions (chosen from Wang & Strong's 15 Dimensions.	**IQ Dimensions**	**Preliminary Factors**
			Accuracy	Discrepancy, Timeliness, Source/Author, Bias/Intentionally False Information
			Completeness	Lack of Depth, Technical Problems, Missing Desired Information, Incomplete When Compared with Other Sites, Lack of Breadth
			Relevance	Irrelevant Hits When Searching, Bias, Too Broad, Purpose of Web Site
			Timeliness	Information is Not Current, Technical Problems, Publication Date is Unknown
			Amount of Data	Too Much Information, Too Little Information, Information Unavailable

An analysis of Table 1 reveals the common elements between the different IQ Frameworks. These include such traditional dimensions as accuracy, consistency, timeliness, completeness, accessibility, objectiveness and relevancy.

Table 2 provides a summary of the most common dimensions and the frequency with which they are included in the above IQ Frameworks. Each dimension also includes a short definition.

Table 2: The Common Dimensions of IQ/DQ

	Dimension	**# of times**	**Definitions** *1[Wang & Strong; 1996]
1	Accuracy	8	extent to which data are correct, reliable and certified free of error *1
2	Consistency	7	extent to which information is presented in the same format and compatible with previous data *1
3	Security	7	extent to which access to information is restricted appropriately to maintain its security *1
4	Timeliness	7	extent to which the information is sufficiently up-to-date for the task at hand *1
5	Completeness	5	extent to which information is not missing and is of sufficient breadth and depth for the task at hand *1
6	Concise	5	extent to which information is compactly represented without being overwhelming (i.e. brief in presentation, yet complete and to the point) *1
7	Reliability	5	extent to which information is correct and reliable *1
8	Accessibility	4	extent to which information is available, or easily and quickly retrievable *1
9	Availability	4	extent to which information is physically accessible
10	Objectivity	4	extent to which information is unbiased, unprejudiced and impartial *1
11	Relevancy	4	extent to which information is applicable and helpful for the task at hand *1
12	Useability	4	extent to which information is clear and easily used
13	Understandability	5	extent to which data are clear without ambiguity and easily comprehended *1
14	Amount of data	3	extent to which the quantity or volume of available data is appropriate *1
15	Believability	3	extent to which information is regarded as true and credible *1
16	Navigation	3	extent to which data are easily found and linked to
17	Reputation	3	extent to which information is highly regarded in terms of source or content *1
18	Useful	3	extent to which information is applicable and helpful for the task at hand *1
19	Efficiency	3	extent to which data are able to quickly meet the information needs for the task at hand *1
20	Value-Added	3	extent to which information is beneficial, provides advantages from its use *1

IQ in the context of its use

In order to accurately define and measure the concept of Information quality, it is not enough to identify the common elements of IQ Frameworks as individual entities in their own right. In fact, Information Quality needs to be assessed within the context of its generation (Shanks & Corbitt, 1999) and intended use (Katerattanakul & Siau, 1999). This is because the attributes of data quality can vary depending on the context in which the data is to be used (Shankar & Watts, 2003). Defining what Information Quality is within the context of the World Wide Web and its Search Engines then, will depend greatly on whether dimensions are being identified for the producers of information, the storage and maintenance systems used for information, or for the searchers and users of information.

The currently accepted view of assessing IQ, involves understanding it from the users point of view. Strong and Wang (1997) suggest that **quality of data cannot be assessed independent of the people who use data**. Applying this commonly to the World Wide Web has its own set of problems. Firstly, there are no quality control procedures for information uploaded onto the Web and secondly, users of the information have to make judgments about its quality for themselves (Rieh, 2002), creating a uniquely subjective environment where one user's quality could be of little or no value to another user. This makes quality dimensions such as relevancy and usefulness not only enormously important but also extremely difficult to gauge.

IQ and Information Search Behaviour

Understanding IQ from the point of view of the user (or searcher) of Information, involves understanding the processes of Information Retrieval on the Internet. More often than not, Information Retrieval (IR) involves using a Search Engine, a specific set of *keywords* or concepts – which make up a user's *query*, followed by a decision process where the user makes *value judgements* concerning the results returned by the Search Engine to their query. These value judgements involve the user making choices according to concepts such as accuracy, currency and usefulness (Rose & Levinson, 2004).

Rose & Levison (2004) advocate that a user's perception of what is accurate, current, important or useful is not only determined by *what* information they are searching for, but by *why* they seek it. The reality that two information searchers can use the same query to convey different meanings or search goals is one of the issues that make developing search engine algorithms which facilitate a searcher's information needs such a difficult proposition. A proposition that would be made immeasurably easier if the search engine could better understand the *intent* of a query.

It is the intent of a user’s query that determines the mental coat hangers by which users make value judgements relating to the quality of a search engine's return on their query. Although the majority of research into IQ continues to reaffirm the widely held belief that these coat hangers are judgements relating to accuracy, usefulness, currency and the like; research within the IR discipline includes concepts such as user-motivation (Barnett, 1999), user self-efficacy (Yee et al, 2004) and other user cognitive processes (Quinn, 2003) as important variables in a user's perception and judgements relating to IQ. The focus on IQ from the perspective of Information Retrieval is a relatively new research area, but is absolutely critical if Information Retrieval Systems are to become effective tools for retrieving quality information from the ever burgeoning World-wide Web.

From a systems perspective, the idea is no longer to simply build a Crawler that can weave its way through the different electronic formats on the Web in order to find content related to a user's query, but one that can apply quality related algorithms to both the Crawling and Ranking strategies of a query search (Tsoi, Forsali, Gori, Hagenbuchner & Scarselli, 2003). Those algorithms would need to go beyond the PageRank strategies employed by many crawlers today, combining an ability to "tunnel" through lower ranked pages and quality criteria to return fewer, but better, results per user-query.

Quantifying Information Quality

Defining IQ with the View to Measuring It

Despite the sizeable body of literature available on Information Quality, relatively few researchers have tackled the difficult task of quantifying some of the conceptual definitions IQ. In fact, a general criticism within the IQ research field is that most approaches lack methods or even suggestions on how to assess quality scores (Naumann & Rolker, 2000). Naumann and Rolker

(2000) go on to suggest that the actual *assessment* of IQ dimensions is difficult because the notion of quality is subjective. This is further complicated by the dynamic nature of the Web, where a page can be edited at will (Hawkins, 1999), or even vulnerable to sabotage, leading to frequent changes in their "quality status".

Developing Metrics for IQ in Information Retrieval

The challenge of this current research is to not only to develop metrics that can assess IQ, but to make them tangible enough to develop into Crawling type algorithms.

Zhu and Gauch's (2000) approach is a relatively simple one, where current crawling technology is enhanced with logical algorithms that quantify characteristics such as currency or availability.

Table 3: Zhu & Gauch's approach to developing tangible assessment methods for IQ :

Assessment Class	IQ Criterion
currency	measured as the time stamp of the last modification of the document.
availability	calculated as the number of broken links on a page divided by the total numbers of links it contains.
information-to-noise ratio	computed as the total length of the tokens after pre-processing divided by the size of the document:
authority	based on the Yahoo Internet Life (YIL) reviews [27], which assigns a score ranging from 2 to 4 to a reviewed site.
popularity	number of links pointing to a Web page, used to measure the popularity of the Web page
cohesiveness	determined by how closely related the major topics in the Web page are

Naumann and Rolker's (2000) approach is more complex, using a three-fold assessment for the quality of an information source, according to the *subjects*, *objects* and *processes* involved in Information Retrieval.

The premise of this model is based on two basic assumptions:

1. ***The Quality of Information is influenced by three factors***:
   - the *perception* of the user,
   - the *information* itself, and
   - the *process of accessing* the information

   and
2. ***The Information Retrieval process involves three entities***:
   - the *user*,
   - the information, and
   - the retrieval system

Both the influences and the processes involved with Information Quality and Retrieval are used to assign quality scores within three contexts, Subject, Process or Object criteria. The scores are used to create metadata that is used to assign a Page Rank for the information source when it is listed in the results of a user's query. Figure 1 demonstrates Nauman and Rolker's (2000) model for classifying the IR entities, IQ factors (or influences) and IQ assessment contexts.

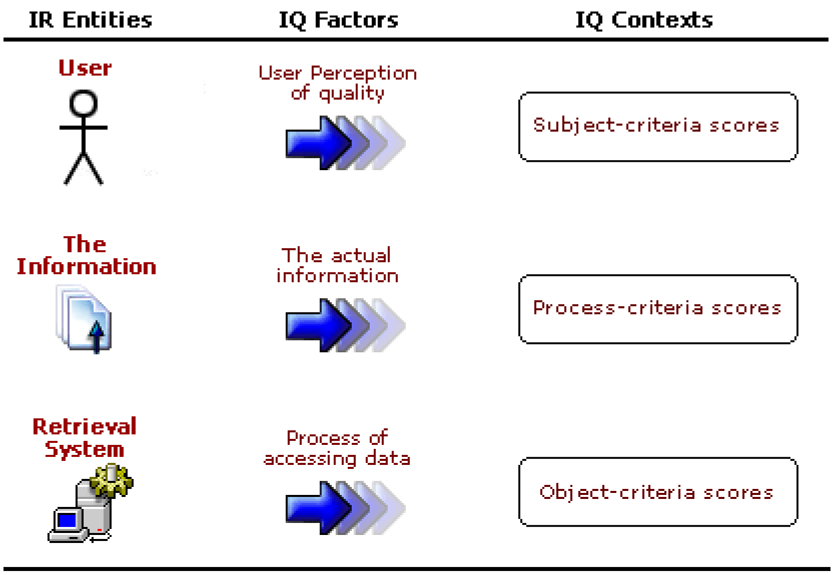


Figure 1: Extension of Nauman & Rolker Model for building quality related metadata of an Information Source

By grouping the entities and factors involved with both IQ and IR into *Subject*, *Object* and *Process* Criteria, Naumann and Rolker (2000) are then able to easily identify IQ criterion and assign assessment methods to them. Table 2 lists the IQ criterion identified by Nauman and Rolker (2000) and suggested methods for assessment.

Table 4: Classification of IQ Metadata Criteria [Naumann & Rolker; 2000]

Assessment Class	IQ Criterion	Assessment Method
Subject Criteria	Believability	User experience
	Concise representation	User sampling
	Interpretability	User sampling
	Relevancy	Continuous user assessment
	Reputation	User experience
	Understandability	User sampling
	Value-Added	Continuous user assessment
Object Criteria	Completeness	Parsing, sampling
	Customer Support	Parsing, contract
	Documentation	Parsing
	Objectivity	Expert input
	Price	Contract
	Reliability	Continuous assessment
	Security	Parsing
	Timeliness	Parsing
	Verifiability	Expert input
Process Criteria	Accuracy	Sampling, cleansing techniques
	Amount of data	Continuous assessment
	Availability	Continuous assessment
	Consistent representation	Parsing
	Latency	Continuous assessment
	Response time	Continuous assessment

Eppler and Muenzenmayer (2002) provide a helpful list of potential IQ related problems associated with individual WebPages, using the IQM (Information Quality Measurement) methodology. The problems (Web-Indicators) are identified within the context of an IQ dimension (IQ-Criterion), and the type of Web Application Tool that can be used to measure the extent of the problem are listed in Table 5.

Table 5: Measuring IQ-criteria for the website context with relevant indicators and adequate tools [Eppler & Muenzenmayer; 2002]

IQ-Criterion	Web-Indicator	Measurement Tool
1. Accessibility	# broken links # broken anchors	Site Analyzer
2. Consistency	# of pages with style guide deviations	Site Analyzer
3. Timeliness	# of heavy (over-sized) pages/files with long loading times	Site Analyzer
4. Conciseness	# of deep (highly hierarchic) pages	Site Analyzer
5. Maintainability	# of pages with missing meta-information	Site Analyzer
6. Currency	Last mutation > six months	Site Analyzer
7. Applicability	# of orphaned (not visited or linked) pages or user rating	Site Analyzer in combination with Traffic Analyzer, User Surveys
8. Convenience	Difficult navigation paths: # of lost/interrupted navigation trails	Traffic Analyzer, Web Mining Tools
9. Speed	Server and network response time	Server & Network Monitoring Tools, or Site Analyzer
10. Comprehensiveness	User rating	User Surveys
11. Clarity	User rating	User Surveys
12. Accuracy	User rating	User Surveys
13. Traceability	# of pages without author or source	Site Analyzer
14. Security	# of weak log-ins	Site Analyzer/Port scanner
15. Correctness	User ratings	User Surveys
16. Interactivity	# of forms # of personalizable pages	Site Analyzer

Developing a Framework for Quality assessment

Leung (2001), like Naumann and Rolker, concentrates on the user-application process in order to develop a method to assess quality. In Leung's (2001) case, the focus specifically concerns an Intranet environment. However, many of the governing principals and decision making processes outlined are useful when developing a way to assess the quality of information on Internet Web Pages. Leung suggests that any metric initiative must address the needs of its potential users (Leung, 2001) and should be objective, cost effective and informative. These suggestions can be summarised in the following framework.

1. identify the user
2. identify the metric application(s)
   (the applications and process that make up the system)
3. identify the dimensions to be assessed
4. prioritise the dimensions to be assessed by applying an
   Importance, *Urgency* and *Cost* metric to each dimension.
5. Develop specific assessment metrics for prioritised dimensions

Leung (2001) developed user surveys to measure the quality of the Intranet System involved in the study, which was appropriate for the dimensions, applications and general technology being assessed. In the case of developing Crawler algorithms however, beyond collecting information about the user and their experience with Internet information retrieval, surveys may prove less effective. The type of assessment required needs to be both ongoing and automated. Nevertheless, the principals of identifying the user, the technology environment and the individual IQ dimensions, followed by prioritising the dimensions and developing technology based assessment metrics is methodologically sound.

The next section of this paper will address this framework in more detail, applying the principles of identifying the user (1), the application/environment (2) and appropriate dimensions of quality (3 & 4) in order to propose tangible quality related metrics for an Internet Crawler.

IQIP: A Proposed Model

The proposed approach we will follow for the execution phase of the project can be summed up as follows:

IQIP; Identify, Quantify, Implement and Perfect.

Figure 2 illustrates the IQIP, Identify – the user, environment and task; Quantify – prioritise appropriate dimensions of Information Quality using a 'Dimension Score'; Implement – the chosen IQ dimensions into the Web Crawler; and Perfect – improve the crawler through system and user feedback.

The Model is explained in detail below.

Identify:

The model proposes that there are 3 entities that need to be identified and understood.

The user: The end-user should be known so that cognitive, sociological and quality choice processes are better understood (Rose & Levinson, 2004). Understanding what motivates users is imperative because it grounds the conceptual ideas of Information Quality into a context (Johnson, 2003) by which it can be assessed.

For the purpose of this project, one of the user groups will be Information Professionals – namely Librarians. It is proposed that they will be used in the Topic Classification and Topic related algorithm testing phases of the Crawler's development. Classification of queries and associated meanings can be built using both automated system feedback and librarian user-group feedback. This is used to refine the focused crawl behaviour of the system (Tsoi, Forsali, Gori, Hagenbuchner & Scarselli, 2003).

The second group of users will be Post-graduate level university students, lecturers and researchers who regularly use the Internet for information search and retrieval purposes. This group of users (or 'searchers') will provide both quantitative and qualitative data about the system, through search-session monitoring and survey feedback and through user profile analysis and interview feedback respectively.

As well as completing surveys and questionnaires in relation to their own subjective perceptions of quality, users will be asked to participate in "controlled information retrieval", i.e.; specific exercises and tasks common to the groups of sub-users. In order to limit the set-task variables, these users will be working on the same equipment in the same computer-lab environment. It is acknowledged that some variables such as user personality, cognitive ability, and previous experience cannot be controlled. This however can be used to the advantage of the research in that it will be utilised to paint a rich picture of a variety of user Information Quality perceptions and Information Retrieval strategies.

In any case, the *task* will always be Information Retrieval, rather than other internet activities such as "surfing" or "entertainment".

The Environment: The true nature of the systems environment must be analysed and understood fully so that the appropriate established IQ dimensions are chosen. In this case, the environment includes the World Wide Web and a Web Page Crawler (type of Search Engine). Understanding the unique characteristics of these two environments should help identify which Information Quality dimensions are likely to thrive or be compromised within their context.

The major characteristics of the World Wide Web can be characterised as follows:

1. open, accessible (parts of it are constantly available),
2. distributed, networked and interlinked (not ONE entity but made up of multiple parts)
3. extremely large – possibly immeasurable – in content and structure

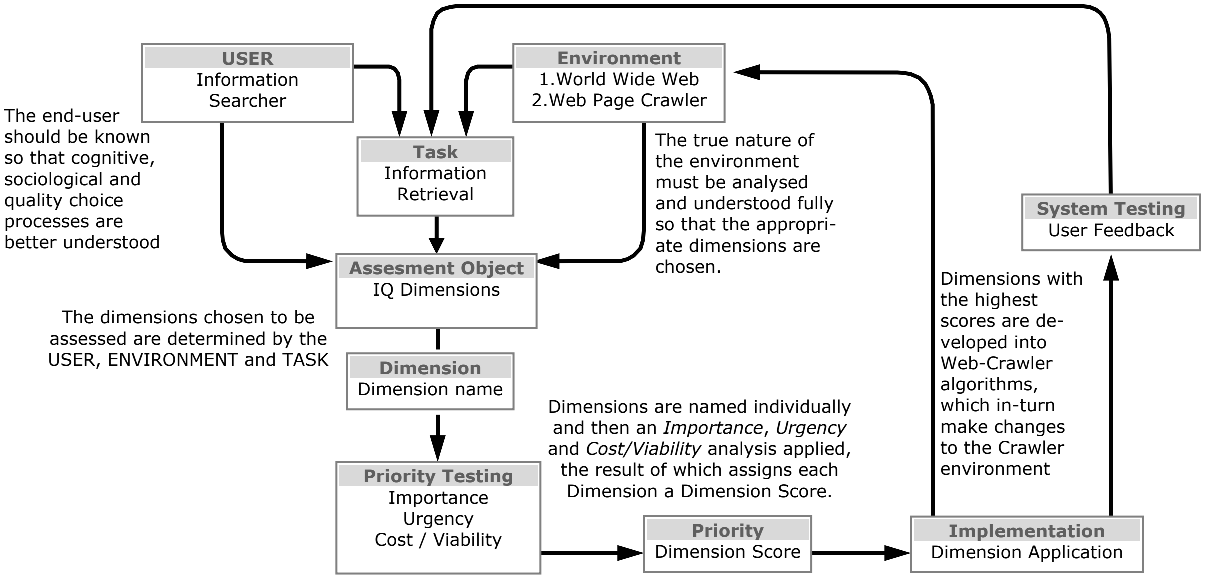
4. evolving, not-static, (Jacobs, 2002)
5. different from traditional Information Retrieval environments (Brooks, 2003)
6. having no enforceable quality or retrieval standards (Eppler & Muenzenmayer, 2002)
7. Unsafe, with component parts vulnerable to breakdown and attack

The major characteristics of Web Page Crawlers environments are typically

1. inconsistent with returns on queries (Iivonen, 1995)
2. limited in what web-formats they are able to parse, ie: can a crawler determine WHAT is inside a *.jpg image?
3. a "snap-shot" of the World Wide Web at a specific time in history (Brooks; 2003) rather than a complete index of data/information available
4. flexible and changeable at a developer level – allowing for constant improvement

The Task: The task must be understood within the context of the end-user and systems environment so that the appropriate dimensions – relevant to the task – can be quantified accordingly. In this case, the task is Information Retrieval on the World Wide Web.

To a lesser degree the IQ Dimensions also need to be identified, however in the context of the IQIP, the purpose of quality naming is not so much to establish what IQ is, but rather to develop a way to prioritise and quantify those generally accepted quality dimensions from previous IQ research literature so that the appropriate IQ elements are applied to the project.



The end-user should be known so that cognitive, sociological and quality choice processes are better understood

The true nature of the environment must be analysed and understood fully so that the appropriate dimensions are chosen.

The dimensions chosen to be assessed are determined by the USER, ENVIRONMENT and TASK

Dimensions are named individually and then an *Importance*, *Urgency* and *Cost/Viability* analysis applied, the result of which assigns each Dimension a Dimension Score.

Dimensions with the highest scores are developed into Web-Crawler algorithms, which in-turn make changes to the Crawler environment

Figure 2: IQIP – A model to Identify, Quantify, Implement & Perfect the process of IQ dimension application to Web Crawler quality retrieval algorithms

Quantify:

The dimensions chosen to be assessed are selected from the established IQ literature, however, they are quantified – given a value and ranking – within the context of ***USER***, ***ENVIRONMENT*** and ***TASK*** (Strong, Lee, & Wang, 1997).

This is achieved using Lueng's (2001) Importance, Urgency and Cost metric. The Cost metric is extended further to include the concept of *Viability*. This is so that other "costs" – besides financial ones – can be included in the dimension analysis. In other words, the costs in the sense of what technical skills or system equipment the project team has at its disposal becomes an important part of the analysis of what IQ dimensions become a priority. It allows the team to address their limitations within the context of the project, and so able to realistically determine what can be achieved.

The Importance, Urgency and Cost/Viability metrics are used to assign each IQ dimension a "Dimension Score", which are used to:

1. Better manage the process of designing and applying algorithms.
2. Make the crawler more practical and functional, better able to meet the Information Needs of users

Implementation:

The implementation phase involves creating Web Crawler algorithms for those IQ dimensions with the highest "dimension score". In keeping with Nauman and Rolker's (2000) model of understanding quality criterion within the context of their assessment class – that is; the context in which the quality is used; algorithms are developed that trigger the Web Crawler to produce Metadata about the pages it crawls.

This metadata is used initially to include or exclude specific pages from the results of a query on the grounds of the dimensions with the highest dimension score. Subsequent algorithms can be used to group results together into clusters according to topics, or into a Page Rank according to Dimension scores.

It should be noted here however, that the initial crawling of a dataset could be considered to be a different system process than that of page ranking. This is because the "environment" – initially the WWW complete with its IQ related characteristics – has now changed to a dataset of documents that meet certain quality criteria. If this is true, then the re-crawling algorithms of these "chosen" results to further refine the search results can also be developed using the IQIP approach.

Perfect:

An important characteristic of the implementation of quality related algorithms is that as the system crawls and achieves results, those results should feedback to the crawler and improve its ability to continue crawling. The feedback is achieved two ways;

1. through automated processes of remembering and analysing successful query results, and;
2. through user-feedback from a control group of system users.

In the case of the current project, it is anticipated that the group of users who will initially "feedback" to the developers will be a control group of librarians familiar and comfortable with electronic search and retrieval. The main purpose of this type of feedback is in relation to developing algorithms that can better classify ***topic*** related content through recognising ***relevancy*** quality dimensions.

The second user-group will test both current Internet Search Engines and the project's (Tsoi Burn, & Gori, 2003) developing focused crawler (Tsoi, Forsali, Gori, Hagenbuchner & Scarselli, 2003) within the context of their perception of Information Quality on the Internet, as it relates to the process of Information Retrieval. The overall goal here is not only to *quantify what users believe to be 'Information Quality'*, but to critically analyse those *perceptions in the context of their actual Information Seeking Behaviour*.

The interface of the focused crawler will include:

- » User-profile settings,
- » Survey / Questionnaires – for user data collection
- » Feedback mechanisms regarding Crawling effectiveness

- Set Information Retrieval exercises

Users will be asked to examine their own perceptions of Information Quality in the context of their Information Retrieval. This process should become progressively more complex as the research goes on and users begin testing the actual Internet Crawler being developed as part of the "Building a Prototype for Quality Information Retrieval from the Internet" project

Conclusion

Defining Information Quality is a complex and multi-faceted issue made even more difficult in the context of information retrieval from non-validated sources such as the World Wide Web. This paper has attempted to summarise the state of research on IQ to date and summarise the most common dimensions which can be applied to measure the concept of IQ in the context of its use. Understanding IQ from the point of view of the user, however, also implies understanding the processes of information retrieval on the web prior to applying metrics to assess quality. An approach to measurement, IQIP, is proposed which encompasses identification of the user, environment and task; quantification of the quality dimensions within the context of user, environment and task; implementation of a process to assess the quality and a feedback mechanism to continually refine and perfect the quality retrieval process based on relevancy.

The next stage of this research will be concerned with the application of IQIP using a closed data set of web pages and the development of an intelligent crawler. At the same time the research team will be further refining the 'Quality' criterion and developing user interfaces which can be used to measure user acceptance and satisfaction with the quality information retrieval process from the web.

References

Alexander, J. E. & Tate, M. A. (1999). *Web wisdom: How to evaluate and create information quality on the web*. Mahwah, NJ: Erlbaum.

Barnett, A. (1999). A survey of Internet searches and their results. *Reference & User Services Quarterly. Winter 1999, 39* (2), 177.

Brooks, T. A. (2003). Web search: How the Web has changed information retrieval. *Information Research, 8* (3); April.

Dedeke, A (2000). A conceptual framework for developing quality measures for information systems. *Proceedings of 5th International Conference on Information Quality*, p.126–128.

Eppler, M J. & Wittig, D. (2000). Conceptualizing information quality: A Review of Information Quality Frameworks from the Last Ten Years. *Proceedings of 5th International Conference on Information Quality*, p.83–96.

Eppler, M. & Muenzenmayer, P. (2002). Measuring information quality in the web context: A survey of state-of-the-art instruments and an application methodology. *Proceedings of 7th International Conference on Information Quality*; p.187–196.

Hawkins, D. T. (1999). What is credible information? *Online, 23* (5), 86-89.

Hölscher, C., & Strube, G. (2000). Web search behaviour of Internet experts and Newbies. *Proceedings of the 9th conference on World Wide Web*, pp.81 - 101.

Iivonen, M. (1995). Searchers and searchers: differences between the most and least consistent searches. *Proceedings of the 18th ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, United States, 149.

Johnson, J. D. (2003). On context of information seeking. *Information Processing and Management, 39* (5), 735-760.

Jacobs, I. (2002). Architectural principles of the World Wide Web, W3C working draft. World Wide Web Consortium (W3C.org). Retrieved 20 March 2003 from http://www.w3.org/TR/2002/WD-Webarch-20020830/

Kahn, B. K.; Strong, D. M. & Wang, R. Y. (2002). Information quality benchmarks: Product and service performance. *Communications of the ACM, 45* (4), 84–192.

Katerattanakul, P. & Siau, K. (1999). Measuring information quality of web sites: Development of an instrument. *Proceedings of the 20th international conference on Information Systems*. Charlotte, North Carolina, United States; p.279–285

Klein, B. D. (2001). User perceptions of data quality: Internet and traditional text sources. *The Journal of Computer Information Systems; 41* (4), 9–18.

Klein B. D. (2002). When do users detect information quality problems on the World Wide Web? *American Conference in Information Systems, 2002*, p1101.

Leung, H. K. N. (2001). Quality metrics for intranet applications. *Information & Management, 38* (3), 137-152.

Naumann, F. & Rolker, C. (2000). Assessment methods for information quality criteria. *Proceedings of 5th International Conference on Information Quality*, p.148–162

Quinn, B. (2003). Overcoming psychological obstacles to optimal online search performance. *The Electronic Library, 21* (2), 142–153.

Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information Science and Technology, 53* (2), 145-161.

Rose, D. E.; Levinson, D. (2004). Understanding user goals in web search. *Proceedings of the 13th international conference on World Wide Web*, 2004.

Shankar, G. & Watts, S. (2003). A relevant, believable approach for data quality assessment. *Proceedings of 8th International Conference on Information Quality*, p.178–189; 2003

Shanks, G. & Corbitt, B. (1999). Understanding data quality: Social and cultural aspects. Proceedings of the 10th Australasian Conference on Information Systems; p785

Strong, D. M.; Lee, Y. W. & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM, 40* (5), 103–110.

Tayi, G. K. & Ballou, D. P. (1998). Examining data quality. *Communications of the ACM, 41* (2), 54–57.

Tsoi, A. C., Burn, J. & Gori, M. (2003). Building a prototype for quality information retrieval from the Internet. *Australian Research Council Discovery Application*; Proj ID: DP0452862.

Tsoi, A.C., Forsali, D.; Gori, M.; Hagenbuchner, M. & Scarselli, F. (2003). A novel focused crawler. *Poster Proceedings of the 12th World Wide Web Conference*, 20-24 May 2003, Budapest, Hungary

Tsoi, A.C., Morini, G.; Scarselli, F.; Hagenbuchner, M. & Maggini, M. (2003). Adaptive ranking of web pages. *Proceedings of the 12th World Wide Web Conference*, 20-24 May 2003, Budapest, Hungary

Wang, R.Y. & Strong, D.M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, Spring,* 5–33.

Wang, Y. (2001). Link based clustering of web search results. *Proceedings of 2nd Web-Age Information Management (WAIM) Conference*. Xi'an, China, p.225-236.

Yee, P. L.; Hsieh-Yee, I.; Pierce, G. R.; Grome, R. & Schantz, L. (2004). Self-evaluative intrusive thoughts impede successful searching on the Internet. *Computers in Human Behaviour, 20* (1).

Zeist, R.H.J. & Hendriks, P.R.H. (1996). Specifying software quality with the extended ISO model. *Software Quality Management IV – Improving Quality, BCS*, 145-160.

Zhu, X. & Gauch, S. (2000). Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, Athens, Greece. 2000. p.288–295

Biography

Shirlee-ann Knight is a Research Fellow with the School of Management Information Systems at Edith Cowan University in Perth, Western Australia. She is currently working on her PhD, "The Impact of User Perceptions of Information Quality on World-wide Web Information Retrieval Strategies", which is part of larger project between ECU, Wollongong and Siena Universities "Building a Prototype for Quality Information Retrieval from the Internet".

Although most of her time is now devoted to research, Shirlee-ann was awarded the "Coursework Supervisor of the Year - 2004" (an award nominated and voted for by students at ECU) for her supervision of the online Information Retrieval & Document Management unit. Prior to beginning her PhD, Shirlee-ann was involved in Interface Design & Content Management components of MIS's school-wide implementation of WebCT. Her research areas of interest include: WWW Quality information Retrieval, Human Computer Interaction, Internet Search Engine Information Seeking Behaviour, Web enabled e-Learning System Interface Design.

Janice Burn is Foundation Professor and Head of School of Management Information Systems at Edith Cowan University in Perth, Western Australia. She has previously held senior academic posts in Hong Kong and the UK. Her main research interests relate to information systems strategy and benefits evaluation in virtual organisations with a particular emphasis on social, political and cultural challenges in an e-business environment. She is on the editorial board of six prestigious IS journals and participates in a number of joint research projects with international collaboration and funding. She has published over 200 papers in the IS field and in 2005 will co-chair the ICIS track on ‘Valuing IT Opportunities’