We need new names: Applying existing models of Information Quality to web archives

Brenda Reyes Ayala
University of North Texas, Department of Library and Information Science
3940 North Elm, Suite C232
Denton, Texas 76203
Brenda.Reyes@unt.edu

Keywords

Web archiving; Information quality; Quality Assurance; Digital Libraries

1. INTRODUCTION

This paper explores how different academic disciplines define Information Quality (IQ) with the goal of determining if any specific model could be applied to define quality in web archives. It presents a literature review of IQ concepts and definitions in Information Science, Philosophy, and Computer Science, as well as a discussion of how these concepts might apply to a web archive. IQ was described in the literature as a multi-dimensional construct and seven dimensions of it were prominently featured: accuracy, currency, usefulness, completeness, consistency, coherence, and credibility. Of these seven dimensions, three were found to be highly applicable to measuring the IQ of a web archive: accuracy (which includes completeness), usefulness, and coherence.

2. THE CONCEPT OF INFORMATION QUALITY: GUIDELINES, THEORIES, AND MODELS

2.1 IQ in Information Science

The field of Information Science has produced several models of IQ. Taylor incorporated IQ as part of his value-added model, which describes the interaction between users and formal information systems. As Taylor defines it, quality is “a user criterion which has to do with excellence or in some cases truthfulness in labeling” [4]. Quality has the values of accuracy, comprehensiveness, currency, validity, and reliability as described below:

1. Accuracy is a guarantee of a true copy, but is independent of the truth value of the information.
2. Comprehensiveness is the value added by the completeness of coverage of a particular subject or discipline.
3. Currency is the value added by the recency of the data acquired by the system and the capability of the system to reflect current modes of thinking in its access vocabularies.
4. Validity is the degree to which the information or data presented to users can be judged as sound.
5. Reliability is the trust a user has in the consistency of quality performance of the systems and its outputs over time. A system is reliable when it maintains an accepted level of accuracy, comprehensiveness, and currency. Taylor states that reliability is the summation of many aspects of quality.

Furthermore, Taylor’s definition of currency could be problematic for web archives. If the purpose of a web archive is to preserve older websites for future use and study, as a type of historical record, then it is not so important that it contain the most up-to-the-minute information. Some users such as historians might regard a web archive to be more valuable the older its contents get. It seems that the notion of currency in a web archive is almost the opposite of what Taylor described.

In 2002, Soo Young Rieh [3] published a study that explores how users viewed the concepts of IQ and cognitive authority on the web. To address her research questions, she studied how users navigated web sites and how they judged the information quality of what they saw. In her results, five key aspects of information quality emerged: goodness, accuracy, currency, usefulness, and importance. Rieh’s study is particularly interesting because, though it is informed by other theoretical models of IQ, her own model is derived from actual user research.

In their 2004 paper, *The continuum of metadata quality: Defining, expressing, exploiting,* Bruce and Hillman [2] addressed the issue of metadata quality in Library and Information Science. They emphasized that quality is a quantifiable and measurable concept, and presented a list of quality measures and metrics that include completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility. It is important to note that Bruce and Hillman did not seek to create a theoretical model of IQ. Instead their aim was to establish a set of feasible guidelines for practitioners. The quality measures are the following:

1. Completeness: The element set describes the target object as completely as feasible and is applied to the target object population as completely as possible.
2. Provenance: The persons who created the data, and their level of expertise, is known. Information about how the metadata was created, extracted, and transformed is included.
3. Accuracy: The information provided in the values is correct and factual and lacks typographical errors, uses standard abbreviations, and so on.

4. Conformance to Expectations: The metadata contains those elements that the community would reasonably expect to find. It does not contain "false promises."

5. Logical Consistency and Coherence: Elements are conceived in a way that is consistent with standard definitions and concepts used in the subject or related domains and are presented to the user in consistent ways.

6. Timeliness: Metadata is in synchronization with the target object and has been recently reviewed and verified.

7. Accessibility: Metadata can be read and understood by users.

2.2 IQ in Philosophy

In the past two decades, philosophers have been paying special attention to defining IQ. As in IS, many philosophers also see IQ as multidimensional; however, some have put forward an additional dimension of IQ known as "fit for purpose," which denotes anticipating and meeting user requirements. Luciano Floridi accepts that IQ is multi-dimensional and is composed of facets such as accuracy, objectivity, accessibility, security, relevancy, timeliness, interpretability, and understanding. But his main point is that past work on IQ has misrepresented the concept of fit for purpose, which has sometimes been treated as a one-dimensional or absolute concept. He argues that the concept of fit for purpose is bi-categorical. High-quality information is:

1. Optimally fit for the specific purpose/s for which it is elaborated (purpose-depth)

2. Easily re-usable for new purpose/s (purpose-scope)

Floridi argues that there is an important tension between these two aspects. Often the better a piece of information fits its original and intended purpose, the less likely it can be reused for another purpose, and vice versa. To address this issue, Floridi proposes that traditional dimensions of quality such as accuracy and timeliness be measured along the concepts of purpose-depth and purpose-scope. Our concept of "fit for purpose" would then change. For example, a pre-Copernican book on astronomy would have low information quality if its purpose was to teach its audience about the nature of the galaxy, but it would have high information quality if its purpose was to teach us about the historical development of Ptolemaic astronomy.

Floridi's concepts of fit for purpose, purpose-depth, and purpose-scope would be applicable to the study of web archives. In the web archiving community there seems to be some confusion as to the audience (real or potential) of a web archive. Some organizations aim to capture websites for a general audience, others focus on very specific audiences such as researchers, while others do not even specify an audience. Having a clear idea of the purpose-depth and purpose-scope of a web archive might help web archivists improve their archives. Ultimately, fit for purpose might be just another version of the "usefulness" dimension as described by Rieh, but Floridi's definition is more detailed and nuanced.

Other philosophers have added to and expanded on the notions of IQ. For example, Batini, Palmonari, and Viscusi [1] make some very important points on the subject. They posit that humans evaluate IQ in two important ways:

Method 1 By using a reference version of the information

Method 2 By referring to the perceptual and/or technological characteristics of information. These characteristics depend on the type of information representation

In other words, people evaluate the quality of different types of information in different ways. For example, a person might read an article that says the capital of Spain is Barcelona. If she consults an encyclopedia and finds that the capital of Spain is actually Madrid, she might judge the original article to have poor IQ (Method 1, IQ is judged by comparison to a reference version). But if she looks at a photograph she might instantly judge it to have bad quality if she finds the image blurry or unfocused (Method 2, IQ is judged by perception). The authors put forward their own definition of IQ, with the different dimensions clustered according to their perceived similarity:

1. *Accuracy/correctness/precision* refer to the adherence to a given reference reality.

2. *Completeness/pertinence* refer to the capability to express all (and only) the relevant aspects of the reality of interest.

3. *Currency/volatility/timeliness* refer to the information up-to-dating.

4. *Minimality/redundancy/compactness* refer to the capability of expressing all the aspects of the reality of interest only once and with the minimal use of resources.

5. *Readability/comprehensibility/usability* refer to ease of understanding and fruition by users.

6. *Consistency/coherence* refer to the capability of the information to comply to all properties of the membership set (class, category,...) as well as to those of the sets of elements the reality of interest is in some relationship.

7. *Credibility/reputation*, information derives from an authoritative source.

2.3 IQ in Computer Science

In their paper, Zhu and Gauch [5] explored how quality metrics can be used to improve the performance of Information Retrieval systems. Their focus was on finding and using metrics that could be operationalized. The authors reviewed numerous quality metrics, and selected the ones they felt were amenable to automatic analysis. For their experiments, the authors operationalized the metrics as follows **??**.

1. Currency: the time stamp of the last modification of the document.

2. Availability: the number of broken links on a page divided by the total numbers of links it contains.

3. Information-to-Noise Ratio: the total length of the tokens (words) divided by the size of the document.

4. Authority: a score from the Yahoo Internet Life (YIL) reviews.

5. Popularity: the number of links pointing to a web page.

6. Cohesiveness: how closely related the major topics in the page are.

They defined the "goodness" of a site as its overall quality. Goodness can be defined as:

$$G_i = \overline{W}_i * (a''_s * \overline{T}_i + b''_s * \overline{A}_i + c''_s * \overline{I}_i + d''_s * \overline{R}_i + e''_s * \overline{P}_i + f''_s * \overline{C}_i) \quad (1)$$

where $\overline{W}_i, \overline{T}_i, \overline{A}_i, \overline{I}_i, \overline{R}_i, \overline{P}_i$ are the means of information quantity, currency, availability, information-to-noise ratio, authority, and popularity of site i across topics relevant to the query, $\overline{C}_i$, is the cohesiveness of site i, and $a''_s, b''_s, c''_s, d''_s, e''_s, f''_s$ are the weights representing the importance of each quality metric. [5, p. 291]

3. FINDINGS AND DISCUSSION

When comparing the different models of Information Quality, it is clear that, though they agree on some aspects of IQ, they are not interchangeable. Some models are highly-tailored to a specific situation or domain, such as Information Retrieval systems (Taylor, Zhu and Gauch), websites (Rieh), and metadata (Bruce and Hillman). Other models, such as those presented by Floridi and Batini et al., are more general in their approach and can arguably be classified as "middle range" theories. Because of their high-degree of generalizability and flexibility, the models put forward by the theorists in Philosophy stand as the most comprehensive and readily-applicable to web archives.

If we examine their commonalities, several dimensions of quality are visible again and again, though sometimes by different names. Accuracy, is present in five of the six models, though Taylor prefers to call it validity. Currency is also similarly present, though Floridi prefers to call it "timeliness", and Batini et al. place it inside the Currency/Volatility/Timeliness dimension. Usefulness is mentioned in three models, though Zhu and Gauch define it as the proportion of useful information and call it the "information-to-noise" ratio.

The dimensions of completeness and consistency are also featured in three models, though Taylor calls the former "comprehensiveness" and the latter "reliability." The last two important dimensions are coherence (characterized by Zhu and Gauch as "cohesiveness", and credibility, called "provenance" by Bruce and Hillman and "authority" by Zhu and Gauch. A summary of the most common facets of IQ is found in Table 1.

Of these seven dimensions of quality, some are readily applicable to web archives, while others are not. Some dimensions also lend themselves to be more easily operationalized than others. Accuracy, if defined as the level of adherence to a reference value (as Batini et al.'s model), is the most important dimension of Information Quality for web archives. In web archiving, the reference value is the original website, against which the archived version is compared. For an archived website to be considered high-quality, it must meet two criteria contain all the same information as the original website and function the same as the original, that is, a user's interactions with the archived website must be the same as with the original.

Because of the second requirement, accuracy can be said to subsume completeness. It does not matter if the information contained in the original website is factually incorrect, or if the original website contained errors such as broken links, or missing images, as these can be reproduced in the archived version without affecting the quality of the web archive. A 1:1 correspondence between the original website and the archived website constitutes perfect accuracy.

Currency (timeliness) is the most problematic dimension. As previously mentioned, if the purpose of a web archive is to preserve older websites for future use and study, as a type of historical record, then it is not so important that it contain the most up-to-the-minute information. Timeliness might still be useful in some contexts. For example, if an institution were attempting to create a web archive of very recent or ongoing events, the timeliness of the web archive might become pertinent. Zhu and Gauch [5] demonstrated in their work that currency can be easily operationalized by comparing the timestamp of the last time a website was updated to the current date. A similar operation could be carried out to measure the timeliness of a web archive.

Usefulness, as Floridi pointed out, is a subjective construct dependent almost entirely on the audience's assessment. Though the concept of usefulness could be applied to web archives, at this moment it is still difficult to assess if the real or imagined audience would find a specific web archive to be useful. The dimension of credibility is similar to usefulness because it depends on the audience.

Consistency and coherence are also quality dimensions that are easily applicable to web archives. As previously mentioned, for a web archive to be of high quality, the archived web sites must have been consistently captured and must replay consistently. Similarly, the individual archived web site must be coherent with the web archive as a whole. However, consistency is a difficult concept to measure, while coherence can be readily operationalized for web archives, particularly smaller web archives that focus on one topic. It would be difficult to ascertain if an entire web archive is consistent, though it would be less difficult for a single archived website or a small, curated web archive. As Zhu and Gauch [5] have shown coherence can be measured; in a web archive, coherence could be determined by calculating the percentage of websites that cover a specific topic.

Table 2 summarizes the points made in this discussion.

4. REFERENCES

[1] C. Batini, M. Palmonari, and G. Viscusi. The many faces of information and their impact on information quality. Symposium conducted at the AISB/IACAP World Congress 2012, July 2012.

[2] T. R. Bruce and D. I. Hillman. The continuum of metadata quality: Defining, expressing, exploiting. In D. I. Hillmann and E. L. Westbrooks, editors, *Metadata in Practice*, pages 238–256. American Library Association, Chicago, 2004.

[3] S. Y. Rieh. Judgment of information quality and cognitive authority in the web. *Journal of the American Society for Information Science and Technology*, 53(2):145–161, 2002.

[4] R. S. Taylor. *Value-added Processes in Information Systems.* Ablex Publishing Corporation, Norwood, NJ, 1986.

Table 1: Common Facets of Information Quality in the Literature.

	Authors					
	Taylor	**Rieh**	**Floridi**	**Batini, Palmonari, Viscusi,**	**Bruce and Hillman**	**Zhu and Gauch**
Facets						
Accuracy	Validity	x	x	x	x	
Currency	x	x	Timeliness	Currency/Volatility/Timeliness	x	
Usefulness		x	x			Information-to-Noise ratio
Completeness	Comprehensiveness		x	x		
Consistency	Reliability		x	x		
Coherence				x	x	Cohesiveness
Credibility				x	Provenance	Authority

Table 2: Dimensions of IQ and their Applicability to Web Archives.

Dimension	**Applicability to web archives**	**Easily operationalized?**
Accuracy (includes completeness)	High	Yes
Currency	Low[a]	Yes
Usefulness	High	No
Consistency	Low[a]	No
Coherence	High	Yes
Credibility	Low	No

[a] Can be easily applied only to small web archives, or those focused on a single topic.

[5] X. Zhu and S. Gauch. Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’00, pages 288–295, New York, NY, USA, 2000. ACM.