We need new names: Applying existing models of information quality to web archives

Brenda Reyes Ayala[1]

[1] University of North Texas

Workshop on Web Archiving and Digital Libraries, Joint Conference on Digital Libraries (JCDL) June 23, 2016

Outline

1. Introduction
2. The Concept of Information Quality: Guidelines, Theories, and Models
3. Method
4. Findings
5. Discussion
6. Conclusions and Future Work
7. References

Goals and Research Questions

The goal of this research is to explore how different academic disciplines define Information Quality. This goal leads to the following research questions:

RQ 1 How do different disciplines characterize Information Quality (IQ)?

RQ 2 Which of these IQ models can be most-readily applied to describing IQ for web archives?

What is Information Quality?

Many scholars have attempted to define IQ

IQ is usually portrayed as a multi-dimensional construct with facets such as accuracy, timeliness, and validity.

Often described as highly subjective, dependent on both the context in which it is being applied and the audience that is viewing or utilizing the information.

Previous Work on Information Quality in Web Archiving

(Spaniol, Mazeika, Denev, and Weikum, 2009)

- Primarily concerned with the data quality of web archives, specifically crawl quality.
- Contents are considered to be coherent if they appear to be "as of" time point *x* or interval *[x;y]*.

(Denev, Mazeika, and Spaniol, 2011) presented the SHARC framework for data quality:

- *Blur*: the expected number of page changes that a time-travel access to a site capture would accidentally see, instead of the ideal view of a instantaneously captured, "sharp" site. Minimize blur.
- *Coherence*: the number of unchanged and thus coherently captured pages in a site snapshot. Here, "unchanged" denotes pages that are definitely known to be invariant throughout some time window, ideally the entire crawl. Maximize coherence.

Previous Work on Information Quality in Web Archiving, cont.

(Ainsworth and Nelson, 2015)

- Concerned with defining quality as meeting measurable characteristics.
- *Completeness* is equated to coverage; a complete web archive does not have undesired or undocumented gaps.
- *Temporal coherence* as defined by Denev, Mazeika, and Spaniol (2011). It is affected by:
- *Drift*: the difference between the target datetime originally required and the Memento-Datetime returned by an archive. Drift can be forwards or backwards in time.

Some questions from the literature

- Data quality vs. information quality?
- Are there dimensions of quality not covered by these definitions?
- Can information quality be defined independently of the technologies used to create the web archive, i.e, the crawler?

Method

- Performed a partial literature review of IQ theories and models in several disciplines.
- Focused on comprehensive theories. Operational definitions were not necessary.
- Compared and contrasted the theories and models to see which facets of IQ were most common.

What is a theory?

> *Theories are abstract entities that aim to describe, explain, and enhance understanding of the world, and in some cases, to provide predictions of what will happen in the future and to give a basis for intervention and action.*

(Gregor, 2006, p. 616)

Defining and operationalizing IQ in Computer Science

Zhu and Gauch (2000) explored how quality metrics can be used to improve the performance of IR systems. Their focus was on finding and using metrics that could be operationalized. Those are:

- Currency
- Availability
- Information-to-Noise Ratio
- Authority
- Popularity
- Cohesiveness

Defining and operationalizing IQ in Computer Science, cont.

Zhu and Gauch (2000) defined the “goodness” of a site as its overall quality. Goodness can be defined as:

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

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.

Defining and operationalizing IQ in Computer Science, cont.

The cohesiveness of a site was defined as

$$C = \frac{\frac{N*(N-1)}{2} * \frac{M*(M-1)}{2} + \sum \sum P_{ij}}{\frac{N*(N-1)}{2}} (i_1 j = N, N-1; i < j)$$

where N is the maximum number of top matching topics requested, M is the minimum of N and the number of matching topics returned, and P_{ij} is the length of the shared path between topic *i* and *j* divided by the height of the ontology.

IQ in Philosophy

Batini, Pulmonari, and Viscusi (2012) defined the following facets of quality:

1. *Accuracy/correctness/precision*: adherence to a given reference reality.
2. *Completeness/pertinence*: capability to express all (and only) the relevant aspects of the reality of interest.
3. *Currency/volatility/timeliness*: the information up-to-dating.
4. *Minimality/redundancy/compactness*: capability of expressing all the aspects of the reality of interest only once and with the minimal use of resources.

IQ in Philosophy, cont.

Batini, Pulmonari, and Viscusi (2012) defined the following facets of quality:

5. *Readability/comprehensibility/usability*: ease of understanding and fruition by users.
6. *Consistency/coherence*: 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.

Common Facets of IQ

- Accuracy
- Currency
- Usefulness
- Completeness
- Consistency
- Coherence
- Credibility

Common Facets of Information Quality in the Literature

Facets	Taylor	Rieh	Floridi	Batini et.al	Bruce et.al	Zhu et.al
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

Accuracy in a web archive: the most important?

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.

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 in a web archive: the issue of time

Currency (timeliness) is the most problematic dimension.

It is not so important that a web archive contain the most up-to-the-minute information.

Might still be useful in some contexts. For example, for small web archives or those covering very recent or ongoing events.

Usefulness in a web archive

Usefulness is a subjective construct dependent almost entirely on the audience’s assessment.

Though it 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.

Consistency and coherence

Consistency and coherence are easily applicable to web archives.

Definition: 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 topically 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.

There is more than one type of consistency or coherence. Ex: temporal coherence, topical coherence, etc.

Credibility

We simply do not have enough have enough data on web archive credibility, partly due to lack of a large user base.

We also run into the risk of confusing: the credibility of a website vs. the credibility of the archive website itself (or the institution that created the web archive).

Dimensions of IQ and their Applicability to Web Archives

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

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

Conclusions and Questions

- The notion of time and its value is different for web archives than for other objects such as websites or metadata.
- Are there some facets of IQ that can be partially operationalized? Ex: Usefulness?
- What is the difference between capture and replay and how can it affect IQ?
- If we define the IQ of a web archive only in terms of what we can reasonably operationalize, are we impoverishing our understanding?

References I

- Ainsworth, S.G, & Nelson, M.L. (2015). Evaluating sliding and sticky target policies by measuring temporal drift in acyclic walks through a web archive. *International Journal on Digital Libraries 16*(2), pp.129-144. doi: 10.1007/s00799-014-0120-4
- Denev, D., Mazeika, M., & Spaniol, M. (2011). The SHARC framework for data Quality in web archiving. *The VLDB Journal, 20*(2), pp. 183-207. doi: 10.1007/s00778-011-0219-9
- Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly, 30*(3), 611-642.

References II

- 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. doi: 10.1002/asi.10017

- Spaniol, M., Denev, D., Mazeika, Arturas, Weikum, G., & Senellart, P. (2009). *Proceedings of the 3rd workshop on Information credibility on the web*. New York, NY: Association for Computing Machinery.

- Spaniol, M., Mazeika, A., Denev, D., & Weikum, G. (2009). "Catch me if you can": Visual analysis of coherence defects in web archiving. *Proceedings of the 9th International Web Archiving Workshop (IWAW)* (pp. 27-37).

- Sutton, R., & Staw, B. (1995). What theory is not. *Administrative Science Quarterly, 40*(3), 371-384.

References III

- Taylor, R. S. (1986). *Value-added processes in information systems*. Norwood, NJ: Ablex Publishing Corporation.
- Zhu, X., & Gauch, S. (2000). 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* (pp.288-295). doi:10.1145/345508.345602