Information Quality

Luciano Floridi

The most developed post-industrial societies live by information, and information and communication technologies keep them oxygenated (English 2009). So, the better the quality of the information exchanged, the more likely such societies and their members may prosper. But what is information quality (IQ) exactly? The question has become increasingly pressing in recent years.[1] Yet, our answers have been less than satisfactory so far.

In the USA, the *Information Quality Act*, also known as the *Data Quality Act*,[2] enacted in 2000, left undefined virtually every key concept in the text. So, it required the Office of Management and Budget "to promulgate guidance to agencies ensuring the quality, objectivity, utility, and integrity of information (including statistical information) disseminated by Federal agencies". Unsurprisingly, the guidelines have received much criticism and have been under review ever since.[3]

In the UK, some of the most sustained efforts in dealing with IQ issues have concerned the National Health Service (NHS). Already in 2001, the Kennedy Report[4] acknowledged that: "All health care is information driven, so the threat associated with poor information is a direct risk to the quality of healthcare service and governance in the NHS". However, in 2004, the NHS Information Quality Assurance Consultation[5] still stressed that "Consideration of information and data quality are made more complex by the general agreement that there are a number of different aspects to information/data quality but no clear agreement as to what these are".

[1] The body of literature on IQ is growing, see for example Olson (2003), Wang et al. (2005), Batini and Scannapieco (2006), Lee et al. (2006), Al-Hakim (2007), Herzog et al. (2007), Maydanchik (2007), McGilvray (2008), and Theys (2011).

[2] http://www.whitehouse.gov/omb/fedreg_reproducible

[3] See more recently US Congress House Committee on Government Reform. Subcommittee on Regulatory Affairs (2006).

[4] http://webarchive.nationalarchives.gov.uk/20090811143745/http://www.bristol-inquiry.org.uk

[5] http://webarchive.nationalarchives.gov.uk/+/www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4125508

L. Floridi (✉)
School of Humanities, University of Hertfordshire, de Havilland Campus, Hatfield, Hertfordshire
AL10 9AB, UK
e-mail: l.floridi@herts.ac.uk

Lacking a clear and precise understanding of IQ properties causes costly errors, confusion, impasse, dangerous risks and missed opportunities. Part of the difficulty lies in constructing the right conceptual and technical framework necessary to analyse and evaluate them. Some steps have been taken to rectify the situation. The first *International Conference on Information Quality* was organised in 1996.[6] In 2006, the Association of Computing Machinery launched the new *Journal of Data and Information Quality*.[7] The Data Quality Summit[8] now provides an international forum for the study of information quality strategies. Pioneering investigations in the 1990s—including Wang and Kon (1992), Tozer (1994), Redman (1996), and Wang (1998)—and research programmes such as the Information Quality Program[9] at MIT have addressed applied issues, plausible scenarios and the codification of best practices. So, there is already a wealth of available results that could make a difference. However, such results have had limited impact also because research concerning IQ has failed to combine and cross-fertilise theory and practice. Furthermore, insufficient work has been done to promote the value-adding synthesis of academic findings and technological know-how.

The proliferation of taxonomies (Batini and Scannapieco 2006 offer an excellent introduction) highlights one of the main difficulties in dealing with IQ. Once IQ is analysed teleologically, in terms of "fit for purpose", IQ properties, known in the literature as *dimensions*—such as accessibility, accuracy, availability, completeness, currency, integrity, redundancy, reliability, timeliness, trustworthiness, usability and so forth—are clustered in IQ groups, known as *categories*, such as intrinsic, extrinsic, contextual, representational and so forth (Table 1 provides an illustration). However, since there are many ways of identifying and specifying dimensions and categories, the result is that the issuing maps do not overlap, some of them resemble Borges' *Celestial Emporium of Benevolent Knowledge*'s *Taxonomy*,[10] and the all-important, practical issue of how to operationalise IQ evaluation processes is disregarded. This is not just a matter of lack of logical rigour and methodological negligence, although they too play a role. The main trouble seems to be caused by:

1. A failure to identify the potentially multipurpose and boundlessly repurposable nature of information as the source of significant complications (this is particularly significant when dealing with "big data" (Floridi 2012)), because of
2. A disregard for the fact that any quality evaluation can only happen at a given *level of abstraction*.[11] To simplify, the quality of a system fit for a particular purpose is analysed at a LoA whose selection is determined by the choice of the purpose in the first place: if one wants to evaluate a hammer for the purpose of holding some paper in place on the desk, then that purpose determines the LoA, which will include, for example, how clean the hammer is; leading to

[6] http://mitiq.mit.edu/ICIQ/2013/

[7] http://jdiq.acm.org/

[8] http://www.dataqualitysummit.com/

[9] http://mitiq.mit.edu/

[10] See Borges, "The Analytical Language of John Wilkins", originally published in 1952, English translation in Borges (1964).

[11] On the method of abstraction and LoA, see Floridi (2008).

Table 1 Example of IQ categories and dimensions adapted from Wang (1998), in italics, dimensions from Batini and Scannapieco (2006)

IQ categories	IQ dimensions
Intrinsic IQ	*Accuracy*, objectivity, believability
Accessibility IQ	*Access*, security
Contextual IQ	Relevancy, value-added, *timeliness*, *complete ness*, amount of data
Representational IQ	Interpretability, ease of understanding, concise representation, *consistent representation*

3. A missed opportunity to address the development of a satisfactory approach to IQ in terms of LoA and purpose orientation.

Admittedly, all this is a bit hard to digest, so here are three examples that should clarify the point.

In the UK, the 2011 Census population estimates were examined through a quality assurance (QA) process "to ensure that users of census data have confidence in the *quality* and *accuracy* of the information" (my italics).[12] The Census Data Quality Assurance Strategy stated that

> The proposed strategy reflects a considered balance between data relevance, accuracy, timeliness and coherence. The data accuracy that can be achieved reflects the methods and resources in place to identify and control data error and is therefore constrained by the imperative for timely outputs. 'Timeliness' refers to user requirements and the guiding imperative for the 2011 Census is to provide census population estimates for rebased 2011 mid-year population estimates in June 2012. 'Coherence' refers to the internal integrity of the data, including consistency through the geographic hierarchy, as well as comparability with external (non-census ONS) and other data sources. This includes conformity to standard concepts, classifications and statistical classifications. The 2011 Data Quality Assurance Strategy will consider and use the best available administrative data sources for validation purposes, as well as census time series data and other ONS sources. A review of these sources will identify their relative strengths and weaknesses. The relevance of 2011 Census data refers to the extent to which they meet user expectations. *A key objective of the Data Quality Assurance Strategy is to anticipate and meet user expectations and to be able to justify*, *empirically*, *2011 Census outcomes*. To deliver coherent data at acceptable levels of accuracy that meet user requirements and are on time, will demand QA input that is carefully planned and targeted. Census (2011), pp. 8–9 (my italics).

Apart from a questionable distinction between information *quality* and *accuracy* (as if accuracy were something else from IQ), overall the position expressed in the document (and in the citation above) is largely reasonable. However, I specify

[12] http://www.ons.gov.uk/ons/guide-method/census/2011/how-our-census-works/how-we-took-the-2011-census/how-we-processed-the-information/data-quality-assurance/index.html

"largely" because the statement about the "key objective" of anticipating and meeting user expectations remains quite problematic. It shows a lack of appreciation for the complexity of the fit for purpose requirement. The objective is problematic because it is unrealistic: such expectations are unpredictable, that is, the purpose for which the information collected in the census is supposed to be fit may change quite radically, thus affecting the fitness itself. To understand why, consider a second example.

In the UK, postcodes for domestic properties refer to up to 100 properties in contiguous proximity. Their original purpose was to aid the automated sorting of the mail. That was what the postcode information was fit for (Raper et al. 1992). Today, they are used to calculate insurance premiums, designate destinations in route planning software and allocate different levels of public services, depending on one's location (postcode) in such crucial areas such as health and social services and education (the so-called postcode lottery). In short, the information provided by postcodes has been radically repurposed, and keeps being repurposed, leading to a possible decline in fitness. For instance, the IQ of postcodes is very high when it comes to delivering mail, but rather poorer if route planning is in question, as many drivers have experienced who expect, mistakenly, a one-to-one relation between postcodes and addresses. The same holds true in the US for the Social Security Numbers (SSNs), our third and last example. Originally, and still officially, SSNs were intended for only one purpose: tracking a worker's lifetime earnings in order to calculate retirement benefits. So much so that, between 1946 and 1972, SSNs carried the following disclaimer: "For social security purposes not for identification". However, SSNs are the closest thing to a national ID number in the USA, and this is the way they are regularly used today, despite being very "unfit" for such a purpose, especially in terms of safety (United States Federal Trade Commission 2010).

The previous examples illustrate the fact that one of the fundamental problems with IQ is the tension between, on one hand, purpose–depth and, on the other hand, purpose–scope. Ideally, high quality information is information that is fit for both: it is optimally fit for the specific purpose/s for which it is elaborated (purpose–depth) and is also easily re-usable for new purpose/s (purpose–scope). However, as in the case of a tool, sometimes the better, some information fits its original purpose, the less likely it seems to be repurposable, and *vice versa*. The problem is that not only may these two requirements be more or less compatible, but that we often forget this (that is, that they may be) and speak of purpose-fitness as if it were a single feature, synonymous for information quality, to be analysed according to a variety of taxonomies. Recall the statement from the Census Data Quality Assurance Strategy. This is a mistake. Can it be avoided? A detailed answer would require more space than is available here, so let me offer an outline of a promising strategy in terms of a bi-categorical approach, which could be implemented through some user-friendly interfaces.

The idea is simple. First, one must distinguish between the purpose/s for which some information is originally *produced* (P-purpose) and the (potentially unlimited) purpose/s for which the same information may be *consumed* (C-purpose). These two categories resemble what in the literature on IQ are known as the "intrinsic" vs. "extrinsic" categories. In our previous example, one would distinguish between postcodes as information fit for the purpose of mail delivery—the P-purpose—and postcodes as information fit for other uses, say driving navigation—the C-purpose. This bi-categorical approach could be introduced in terms of a simple Cartesian

Table 2 Example of bi-categorical IQ analysis

IQ dimensions	IQ categories	
	Mail delivery	Mail delivery
Accuracy	1	0.8
Objectivity	1	1
Accessibility	0.9	0.9
Security	1	1
Relevancy	1	0.9
Timeliness	1	1
Interpretability	0.8	0.7
Understanding	1	0.9

space, represented by P-purpose = x and C-purpose = y, in such a way that, for any information I, I must have two values in order to be placed in that space. This in turn allows one to analyse a variety of dimensions, such as accuracy, objectivity, accessibility, etc. in a purpose-oriented way (see Table 2 for an illustration).

Second, one could then compare the quality of some information with respect to purpose P and with respect to purpose C, thus identifying potential discrepancies. The approach lends itself to simple visualisations in terms of radar charts (see Fig. 1, for an illustration based on the data provided in Table 2).

The result would be that one would link IQ to a specific purpose, instead of talking of IQ as fit-for-purpose in absolute terms.

There are many senses in which we speak of fit for purpose. A pre-Copernican, astronomical book would be of very bad IQ, if its purpose was to instruct us on the nature of our galaxy, but it may be of very high IQ if its purpose is to offer evidence about the historical development of Ptolemaic astronomy. This is not relativism; it is a matter of explicit choice of the purpose against which the value of some information is to be examined. Once this methodological step is carefully taken, then a bi-

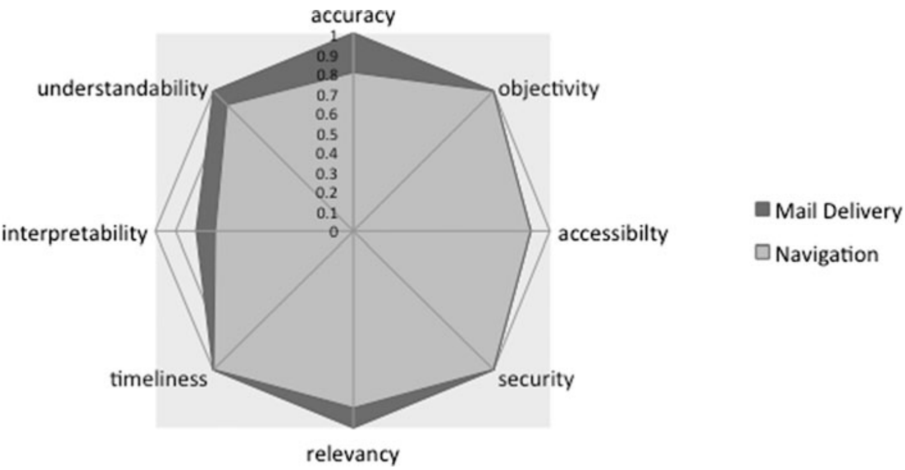

Fig. 1 Graph of a bi-categorical IQ analysis

categorical approach is compatible with, and can be supported by quantitative metrics, which can (let users) associate values to dimensions depending on the categories in question, by relying on solutions previously identified: metadata, tagging, crowd sourcing, peer review, expert interventions, reputation networks, automatic refinement and so forth. The main advantage of a bi-categorical approach is that it clarifies that the values need not be the same for different purposes. It should be rather easy to design interfaces that enable and facilitate such interactive selection of purposes for which IQ is evaluated. After all, we know that we have plenty of information systems that are syntactically smart and users who are semantically intelligent, and a bi-categorical approach may be a good way to make them work together successfully.

References

Al-Hakim, L. (2007). *Information quality management: theory and applications*. Hershey, PA: Idea Group.

Batini, C., & Scannapieco, M. (2006). *Data quality-concepts, methodologies and techniques*. Berlin: Springer.

Borges, J. L. (1964). *Other inquisitions, 1937–1952*. Austin: University of Texas Press.

Census (2011) Census Data Quality Assurance Strategy, http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/processing-the-information/data-quality-assurance/2011-census—data-quality-assurance-strategy.pdf

English, L. (2009). *Information quality applied: best practices for improving business information, processes, and systems*. Indianapolis: Wiley.

Floridi, L. (2008). The method of levels of abstraction. *Minds and Machines, 18*(3), 303–329.

Floridi, L. (2012). Big data and their epistemological challenge. *Philosophy and Technology, 25*(4), 435–437.

Herzog, T. N., Scheuren, F., & Winkler, W. E. (2007). *Data quality and record linkage techniques*. New York: Springer.

Lee, Y. W., et al. (2006). *Journey to data quality*. Cambridge: MIT.

Maydanchik, A. (2007). *Data quality assessment*. Bradley Beach: Technics.

McGilvray, D. (2008). *Executing data quality projects ten steps to quality data and trusted information*. Amsterdam: Morgan Kaufmann/Elsevier.

Olson, J. E. (2003). *Data quality the accuracy dimension*. San Francisco: Morgan Kaufmann.

Raper, J. F., Rhind, D., & Shepherd, J. F. (1992). *Postcodes: the new geography*. Harlow: Longman.

Redman, T. C. (1996). *Data quality for the information age*. Boston: Artech House.

Theys, P. P. (2011). *Quest for quality data*. Paris: Editions TECHNIP.

Tozer, G. V. (1994). *Information quality management*. Oxford: Blackwell.

United States Federal Trade Commission. (2010). *Social security numbers and ID theft*. New York: Nova Science.

United States. Congress. House. Committee on Government Reform. Subcommittee on Regulatory Affairs. (2006). *Improving Information Quality in the Federal Government: hearing before the Subcommittee on Regulatory Affairs of the Committee on Government Reform, House of Representatives, One Hundred Ninth Congress, First Session, July 20, 2005*. Washington: U.S. G.P.O.

Wang, R. Y. (1998). A product perspective on total data quality management. *Communication of the ACM, 41*(2), 58–65.

Wang, R. Y., et al. (Eds.). (2005). *Information quality*. Armonk: ME Sharpe.

Wang, Y. R., & Kon, H. B. (1992). *Toward quality data: an attributes-based approach to data quality*. Cambridge: MIT.