The Challenge of “Quick and Dirty” Information Quality

ADRIANE CHAPMAN, ARNON ROSENTHAL, LEN SELIGMAN
The MITRE Corporation

We present a new research challenge in ‘quick and dirty’ information quality (IQ) –to quickly assess sources’ quality. We also describe its real-world importance, and suggest research directions.

1. THE PROBLEM

Traditional information quality (IQ) techniques provide CIOs and other senior IT managers an in-depth quality assessment of data assets under the control of the enterprise. The goal is not only to clean dirty data in order to improve current operational effectiveness but also to identify ways in which the organization’s data production processes can be improved so as to produce higher quality data in the future. The information quality assessment and improvement process typically takes place over months and years, and consumes large amounts of data management resources [2, 7, 8]. Prior work such as [3] that study data quality within a limited time frame assume that the data owner is undertaking a curation task.

This paper argues for a new direction in IQ research, to address a different user’s need: the composable capabilities developer who must rapidly 1) choose from among a pool of candidate data sources (often external ones beyond the control of the developer’s home organization) and, as quickly as possible, 2) do some cleaning (e.g., removing illegal values), 3) integrate, and 4) provide a lightweight, value-added application to decision makers. This is fundamentally a triage activity, involving multiple, rapid sprints that identify the next 1-2 sources that seem worth incorporating, postponing some for subsequent sprints, while ruling out others from further consideration. There are often severe time constraints, thus giving rise to the need is for *quick and dirty information quality* (QDIQ) assessments.

We have encountered the need for QDIQ among several government agencies spanning public health, disaster response, intelligence, counter terrorism, and coalition military operations. A common thread is the need to help decision makers respond to a rapidly evolving crisis. Such users typically prefer receiving timely, “good enough” information over pristine information that is late.

Example 1: After the 2010 earthquake in Haiti, the Haitian government asked the US to coordinate relief efforts across 30+ nations and hundreds of NGOs. This was an extraordinarily complex effort including delivering and distributing food, water and

Author’s addresses: A. Chapman and L. Seligman, MITRE, 7515 Colshire Drive, McLean, VA 22102 USA; A. Rosenthal, MITRE, 202 Burlington Road, Bedford, MA 01730 USA.

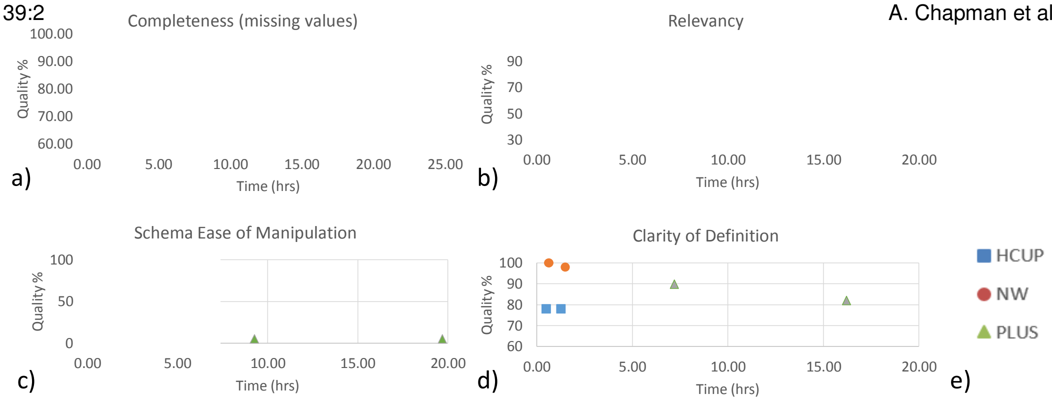

Figure 1: Classic Data Quality assessments for three datasets, PLUS, HCUP and Northwind to show the time it takes to do such an assessment and the variations in quality readings. a) Completeness; b) Relevancy; c) Schema Ease of Manipulation; d) Clarity of Definition; e) Legend for all charts.

medicine, providing medical and security services, executing rescue operations, and more [1]. Decision makers needed data to support urgent decisions. For example, “Three ships are leaving soon for Haiti; based on current needs and resources, how should I allocate what goes on those ships?” Not surprisingly, information was incomplete, heterogeneous, and of widely varying quality. The data quality challenge is: Given extremely limited time, what information would provide the greatest value to support this decision? Which datasets should be taken “with a grain of salt”?

Example 2: In the week after Hurricane Sandy, two datasets describing gasoline sources with very different qualities emerged. The AAA spreadsheet, created by American Automobile Association staff calling all gas stations in the region, is very complete. Every gas station has an exact address; however, the information was days old upon release. The other dataset, the @NYC_GAS Twitter feed, is almost the exact opposite: It is very up to date, but the unstructured format makes it hard to manipulate and gas stations addresses are imprecise (e.g. Joe’s Exxon on 2nd Street). The developer must choose a source based on how and why she intends to use the data, and which qualities (e.g., timeliness, ease of manipulation) are most important for that use. While data quality metrics are objective, each user must prioritize them based on their requirements [1, 4].

Each graph in Figure 1 shows one of the many quality metrics one might compute. The X-axis shows time spent doing a traditional data quality assessment, while the Y-axis shows the quality score after a particular time. Three datasets (Figure 1e) were used. The takeaway is that understanding a dataset to understand quality is resource intensive, and there is no obvious “stopping point”. Sometimes additional assessment time will change a quality score significantly (e.g., in Figure 1d, more detailed assessments of the PLUS dataset show that the initial quality estimates were too optimistic), but sometimes additional assessment time makes no appreciable difference (Figure 1c). It is difficult to know when “enough” assessment has been done.

2. RESEARCH DIRECTIONS

This work presents several directions for IQ research, including:

- Time to “good enough”: When is the assessment sufficient to support a decision? Can we provide useful quick and dirty advance estimators for the effort needed to clean a source just enough to make it usable?
- Focusing human attention: Some crude IQ metrics can be run completely automatically, but many require human input (e.g., to describe how NULL is represented in a given system), or to compare the accuracy of a source with some known ground truth). As discussed in [5], how do we focus expensive and scarce human-in-the-loop analyst time on the highest value next questions to consider?
- Utility functions: Can we define stereotypical utility functions so an analyst can quickly describe specific needs and rapidly determine the quality requirements?
- Cost-Benefit functions for classic IQ: To focus quality improvement efforts, we need ways to estimate the utility to applications of improving the quality of a particular item.
- QDIQ dashboards: Given many data profiling statistics, what should developers be shown to help them do QDIQ? Which visualizations are most useful?
- Reuse of existing profiling tools: There are many data profiling tools [6]; can we reuse and extend existing tools to meet QDIQ needs rather than starting from scratch?
- “Integratability” as a new, potentially useful IQ metric: Most existing IQ tools consider one source at a time. However, composable capability developers need to know how much effort will be required to integrate one source with others feeding the data mashup. In addition, can we use information such as [9] to provide more details about “integratability”?
- Creating a high-quality, free, open source QDIQ toolkit

In summary, classic information quality assessment methods are time and resource intensive. This short paper has described open research issues in quick and dirty information quality assessment, and motivated the need to apply solutions to these research issues for important problems.

REFERENCES

[1] C. Cappiello, C. Francalanci, and B. Pernici, "Data quality assessment from the user's perspective," in *Proceedings of the 2004 international workshop on Information quality in information systems*. Paris, France: ACM, 2004, pp. 68-73.

[2] N. Gorla, T. M. Somers, and B. Wong, "Organizational impact of system quality, information quality, and service quality," *The Journal of Strategic Information Systems*, vol. 19, pp. 207-228, 2010.

[3] S. B. Long, L. Roberge, J. T. Lamicela, and M. Murugesan, "Balancing Data Quality Against Time and Money Constraints," *Data Warehousing, Management and Quality Sugi*, vol. 29, 2004.

[4] P. Missier and C. Batini, "A Multidimensional Model for Information Quality in Cooperative Systems," *International Conference on IQ*, pp. 25-40, 2003.

[5] P. Missier, G. Lalk, V. Verykios, F. Grillo, T. Lorusso, and P. Angeletti, "Improving Data Quality in Practice: A Case Study in the Italian Public Administration," *Distributed and Parallel Databases*, vol. 13, pp. 135-160, 2003.

[6] F. Naumann, "Data profiling revisited," *ACM SIGMOD Record*, vol. 42, pp. 40-49, 2014.

[7] T. Redman, "The impact of poor data quality on the typical enterprise," *Communications of the ACM*, vol. 41, pp. 79-82, 1998.

[8] C. C. Shilakes and JulieTylman, " Enterprise information portals," *Technical Report, Merrill Lynch, Inc.*, 1998.

[9] K. Smith, P. Mork, L. Seligman, A. Rosenthal, M. Morse, D. Allen, and M. Li, "The Role of Schema Matching in Large Enterprises," *CIDR*, 2009.