DATA QUALITY AND CURATION

Kevin Ashley

Digital Curation Centre, University of Edinburgh, Level 7, Appleton Tower, Crichton Street, Edinburgh EH8 9LE, UK
*Email:*kevin.ashley@ed.ac.uk

1 STATE-OF-THE-ART

Data quality is an issue that touches on every aspect of the research data landscape and is therefore appropriate to examine in the context of planning for future research data infrastructures. As producers, researchers want to believe that they produce high quality data; as consumers, they want to obtain data of the highest quality. Data centres typically have stringent controls to ensure that they only acquire and disseminate data of the highest quality. Data managers will usually say that they improve the quality of the data they are responsible for. Much of the infrastructure that will emit, transform, integrate, visualise, manage, analyse, and disseminate data during its life will have dependencies, explicit or implicit, on the quality of the data it is dealing with.

Therefore, all of the actors involved at every stage of the research data lifecycle care about quality. Yet studies show that these actors rarely agree what quality means: I want data that are comprehensive; you want data that are timely; she wants data that are accurate; they want data that are free. Each of us believes that we are talking about quality, and we are. But our metrics measure different dimensions of data quality. Sometimes these quality metrics are orthogonal to each other, and so multiple users can be satisfied by the same data source. But on other occasions the quality parameters are in conflict. It is not possible to produce comprehensive data both quickly and cheaply, for instance. One useful perspective on quality parameters is offered by the work of Wang and Strong (1996), which reduces hundreds of parameters from many domains to 15 essential domain-independent measures. At present, particular disciplines, data centres, and research groups will concentrate on a subset of these measures, often giving them domain-specific names. They are rarely explicit about the quality parameters they focus on. In many cases they also lose sight of a fundamental characteristic of most data quality metrics: that they are surrogates for the quality measures that we really care about, such as truth. Provenance, precision, and completeness are all examples of such surrogates for some use cases.

One of the consequences of this picture is that a particular set of data typically only has one set of quality parameters applied to it. We think in terms of lifecycles for data. There are many research data lifecycles to choose from, the Research Data Lifecycle Model (UK Data Archive, 2013), the DCC Curation Lifecycle Model (Higgins, 2008), and the DDI 3.0 conceptual model (DDI Alliance, 2004), are just some of the better-known examples, but all encourage a view that a single set of processes – including those relating to quality – is applied to a single set of data. The consequences of this are that it can be difficult to integrate data from multiple contexts and to apply different quality mechanisms for different uses. Some examples are given in the use cases in 'Current Challenges' below.

Tools to check or improve data quality and training for data managers in data quality tend to be highly domain-specific. They are also insufficiently mature and scalable to deal with the data volumes that are already available. This is already leading to situations where two of the generic quality parameters – timeliness and accessibility - fail to be met for much research data. Because we wish to apply some other quality metric, such as accuracy or completeness, and because the tools we have to do this are inadequate, we fail to carry out the basic tasks of letting people know that (raw) data exist and allowing them to use this data. Some of the World Data Centres (ICSU World Data System: http://www.icsu-wds.org/), as only one example, can take 4 years to make data available for reuse after it is first submitted to them because of the checking, reprocessing, and enhancement that they apply. Giving those data centres more resources might result in some improvement to these times, but fundamental process change is likely to be needed to reduce years to days, hours, or minutes – and use cases could be found for each of those timescales.

Domain-specificity is inevitable to a great extent, but some checking and quality improvement processes can be highly generic. The US National Archive's AERIC system (http://www.archives.gov/foia/privacy-program/privacy-impact-assessments/aeric.pdf) and similar tools developed for the UK's NDAD enable highly generic checks to be applied to relational databases without knowing much about their origin or purpose. Similar tools are available for use with commercial data, where the phrase 'data cleaning' is often used to describe manual and automated mechanisms to improve one or more data quality parameters. This is also a fruitful research area in computer science (Jai, 2008). Similarly, training based on generic principles such as those articulated by Wang and Strong (1996) has been developed by organisations such as the DCC (DC 101: http://www.dcc.ac.uk/training/train-the-trainer/dc-101-training-materials). iSchools (http://www.ischools.org/) around the world are also aiming to train people with generic skills in information management. The result is a cohort of data managers, curators, and librarians who can move between disciplines, share knowledge on cross-domain data quality issues, and assist in the development of generic tools to improve automated data quality verification and enhancement. Such a cohort of skilled people with a common understanding of issues such as data quality is an essential non-technical component of regional, national, and global research data infrastructure.

2 TEN-YEAR VISION

What changes in this picture can we reasonably hope for in 10 years' time? Some are already alluded to in the description of the current state. Improvements in the education and continuous professional development of those who care for data should increasingly focus on non-domain-specific skills and measures of data quality. This can be complemented by understanding of how domain-specific concerns relate to the more general metrics, to the eventual improvement of both. Similarly, the development of more generic tools and processes for validating and improving various aspects of data quality can be hoped for but will require much more specific encouragement to become a reality. This is because there is less incentive for discipline-specific data curators to develop generic toolsets. The incentive so far has been seen by national archives and university data libraries that are increasingly dealing with data from a very wide variety of disciplines and must develop non-domain-specific tools and workflows to deal with this.

We can also reasonably expect that it will be far simpler to automatically determine aspects of data quality, to support research that requires data from many sources – and potentially many domains – that have had consistent quality parameters applied. This is not the same as looking only for 'high quality' data, whatever that means. It means being able to look for data that are good enough and that are comparable. It also means being able to do so without having to interpret domain-specific quality assertions or, even worse, narrative descriptions in natural language of data quality.

Making machine-readable assertions about data quality will also have other benefits; it will allow such assertions to be carried with data throughout the increasingly complex supply and use chains, which are a natural part of global research data infrastructures. Such quality assertions become part of the data's provenance. This is particularly important for the free movement of data in and out of the academic research domain. Not all research data are produced by research, and data produced by research can have value outside it. Such movement can be damaging to domain-specific metadata, but generic quality assertions (like general descriptive terms such as Dublin Core) are more likely to survive such transitions.

It will be far more straightforward to allow multiple releases of the same data set with different quality parameters applied – rapid release of uncorrected data, later release of highly-curated data – and to do so in ways that do not risk confusion arising as to which of the releases is 'real', since all of them have validity. This pattern of release occurs in a small number of scientific disciplines at present but not in ways that translate easily to generic research data infrastructures.

There will be a larger variety of generic tools to assess data against various quality metrics, allow them to be appropriately labeled, and to improve some aspects of quality in an automated fashion. These tools will allow one other aspect of quality – timeliness or currency – to be improved for many data releases. The tools will be capable of examining and labelling collections of (sometimes heterogeneous) data as well as individual datasets.

3 CURRENT CHALLENGES

Almost every aspect of the vision relates to a challenge or obstacle in the present day. Some of these are illustrated here through present-day use cases.

The Multiple Quality Problem
A government department collected information about traffic flow through a mixture of automated systems and human observation. Much of the work was carried out by private organisations through a variety of government contracts. The data were used to make predictions about future traffic flows and hence guide investment in transport infrastructure. Some years later the data are being made available via the National Archive. It has become apparent in the intervening years that the original data were very flawed, due to a mixture of poor methodology, malfunctioning instruments, and corruption. The National Archive makes this original data available – it is a true record of the information that governments used to make (poor) decisions. Other researchers realise that many of the errors are systematic and can be corrected automatically – they produce a new version of the data that is more accurate (in that it reflects true historical traffic patterns) but is not an accurate record. Both groups claim that the other is misleading researchers as to the quality of their data, but in reality neither data set is 'better' than the other one. They serve different communities of use.

The Career Problem
A young researcher in soil microbiology finds that his/her career begins to involve curating data more than creating data. He/she becomes very good at it. In time, funding for the field is not so easy to find. The researcher is now an excellent data curator with many transferable skills that could be applied in other disciplines. But he/she does not realise this and neither do others because their skills are expressed in terms relevant only to soil microbiology. Frustrated at his/her inability to get work in this field, he/she gives it all up and becomes a landscape gardener. Meanwhile, expensive data are being lost because of a lack of skilled data managers.

The Cloud Vision Problem
Microsoft and others have succeeded in their vision of building cloud compute and data infrastructures that allow working researchers to access hundreds of multi-disciplinary data stores using the spreadsheet on their desktop. Unfortunately the analysis the researcher wants to carry out is dependent on the precision and coverage of the underlying data sets. In the absence of domain-neutral ways of expressing these quality metrics, the researchers are faced with two courses of action. They can do the analysis anyway and end up with bad science; or they can employ a number of domain specialists to investigate each of the data sources, interpret their descriptions and domain-specific quality assertions in a way that makes sense to these researchers, and then use only those sources that meet the criteria for this analysis (note: these are not THE high-quality sources; they are the sources of appropriate quality for this piece of research.) The latter approach loses all of the benefits of the research data infrastructure that the researchers have access to.

The Portable Tool That Isn't Ported
Commercial suppliers have developed a variety of data cleaning tools targeted on a common business problem, managing customer lists and data about them. The techniques they have developed are actually applicable in many other domains of greater interest to research, but they will not be adopted there. There is no commercial incentive for the data cleaning application developers and no awareness of their work or methods by the relevant research communities. Their wheel will be reinvented many times – until someone tries to sue someone else for patent infringement.

4 RESEARCH DIRECTIONS PROPOSED

The research directions required are those exemplified by the problems articulated above: the development of non-domain-specific ways of expressing quality assertions or measures via metadata; the development of more general tools for assessing the quality of data and improving the quality of data and data collections. There are also changes in practice and infrastructure that are more properly characterised as development rather than research. Foremost among these is the education of data managers, curators, specialists, and carers (the terms vary, but the jobs are much the same) in generic techniques of data quality and other aspects of data curation. Human infrastructure is

even more important than physical infrastructure at present. Without human infrastructure that understands the general as opposed to domain-specific requirements for research data, we cannot build generic infrastructure at sufficient scale.

5 REFERENCES

DDI Alliance (2004) DDI 3.0 conceptual model. MIT Libraries. Retrieved from the World Wide Web, June 30, 2013: http://libraries.mit.edu/guides/subjects/data-management/cycle.html

Higgins, S. (2008) The DCC Curation Lifecycle Model. *International Journal of Digital Curation 3*(1). Retrieved from the World Wide Web, June 30, 2013: http://ijdc.net/index.php/ijdc/article/view/69/48

Jia, X. (2008) *From Relations to XML: Cleaning, Integrating and Securing Data*. PhD thesis, the University of Edinburgh. Retrieved from the World Wide Web, June 30, 2013: http://hdl.handle.net/1842/3161

UK Data Archive (2013) Research data lifecycle model. Retrieved from the World Wide Web, June 30, 2013: http://www.data-archive.ac.uk/create-manage/life-cycle

Wang, R, & Strong, D. (1996) Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, Armonk. Retrieved from the World Wide Web, June 30, 2013: http://web.mit.edu/tdqm/www/tdqmpub/beyondaccuracy_files/beyondaccuracy.html

(Article history: Available online 30 July 2013)