Challenges of open data quality: more than just license, format, an customer support.

CORSAR, D. and EDWARDS, P.

2017

Challenges of Open Data Quality: More Than Just License, Format, and Customer Support

DAVID CORSAR, University of Aberdeen
PETER EDWARDS, University of Aberdeen

CCS Concepts: •**Social and professional topics** → **Quality assurance;** •**Applied computing** → *E-government;*

Additional Key Words and Phrases: Open Data

ACM Reference Format:
David Corsar, and Peter Edwards, XXXX. Challenges of Open Data Quality More Than Just License, Format, and Customer Support. *JDIQ* V, N, Article A (January YYYY), 3 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Public sector organisations world-wide are implementing Open Data initiatives, which, it is hoped, will stimulate economic growth, increase transparency and accountability, and improve engagement between data consumers (typically citizens) and data holders/publishers [Open Knowledge 2012]. Open Data is defined as data that "anyone can freely access, use, modify, and share for any purpose."[1] As developers have started using open data within their applications they are reporting quality issues with such data sets that have subsequently been addressed by the data publishers[2]. These include OpenStreetMap[3] developers correcting the location of 18,000 UK bus stops, and users identifying errors and omissions in data relating to UK registered charities. There is a growing recognition among the open data community that, in order to maximize the impact of such initiatives in terms of economic growth and increased accountability, focus must shift from publication of data to issues such as coverage, openness, and quality[4]. Definitions of quality in the open data context vary considerably; for example, the European Data Portal considers data to be of high quality if "humans can understand it and machines can manipulate it"[5] and point to the 5 Star Open Data rating system[6] as a data marque. Others, such as the G8 Open Data

[1]http://opendefinition.org/
[2]For further examples see, http://bit.ly/opendata-betterdata
[3]http://www.openstreetmap.org
[4]https://opendatawatch-public.sharepoint.com/Pages/MR-Indices.aspx
[5]http://www.europeandataportal.eu/elearning/en/module5/#/id/co-01
[6]http://5stardata.info/

The research described here was supported by the award made by the RCUK Digital Economy programme to the dot.rural Digital Economy Hub, award reference: EP/G066051/1; and by the Innovate UK award reference: 102615.
Author's addresses: D. Corsar and P. Edwards, Computing Science, University of Aberdeen, Aberdeen, UK, AB24 5UA.

DOI: http://dx.doi.org/10.1145/0000000.0000000

Table I. **Sample Open Air Quality Data.** Extract from Comma-Separated Values file downloaded from the UK DEFRA data archive, providing data measured at two monitoring sites in Aberdeen, UK.

Hourly measurement data supplied by UK-air on 25/10/2016							
All Data GMT hour ending							
Status: V=Verified P=Provisionaly Verified N=Not Verified S=Suspect							
		Aberdeen				Aberdeen Union Street Roadside	
Date	Time	Volatile PM2.5 (Hourly measured)	Status	Sulphur dioxide	Status	Nitrogen oxides as nitrogen dioxide	Status
01/01/2016	01:00:00	-3	V ugm-3 (TEOM FDMS)	No data		92.80215	V ugm-3
01/01/2016	02:00:00	-4.2	V ugm-3 (TEOM FDMS)	No data		52.66069	V ugm-3
01/01/2016	03:00:00	-2.8	V ugm-3 (TEOM FDMS)	No data		60.26527	V ugm-3

Charter[7] and the Open Data Institute Certification Badges[8] focus on the provision of metadata, data schema descriptions, use of shared data dictionaries, license used, file format, and publisher support for interacting with data users.

2. AN EXAMPLE: POLLUTANT EMISSION DATA

Pollutant emissions are a key dataset included in the Global Open Data Index assessment[9], and are currently published by 61 of the 121 countries included in the latest survey. In the UK this data is recorded at 300 monitoring sites across the country, and published as CSV files by the DEFRA Data Archive[10]; Table I provides an extract of data obtained from this service for two monitoring sites in the city of Aberdeen, UK[11]. While small, this extract illustrates the kind of quality issues frequently encountered in open data. The file does not comply with the RFC 4180 CSV specification[12], as the first three rows provide meta-information about the data, rather than a header line and data records. Consistency issues are also evident: the "Status" column provides two pieces of information: the datum's status (indicating if it has been subjected to DEFRA's data verification process[13]) and the measurement units; and missing data appears to be handled inconsistently, either with the text "No data" or an empty cell. Accuracy issues are also evident, as shown by the negative values for Volatile PM2.5, which typically indicate an error with the calibration of the device, despite these values having been verified, indicating that the PM2.5 values cannot be trusted.

In Scotland, emissions data is also available from the Scottish Air Quality (SAQ) website[14]. Comparison of data from the SAQ and DEFRA sources identifies a further accuracy issue: DEFRA provide data to (several) decimal places, while the SAQ service appears to round the data to zero decimal places. There are also inconsistencies in the schemas, as the SAQ dataset includes a "units" column, instead of overloading the "Status" column as shown in Table I. Such consistency issues are also evident when

[7]https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex
[8]https://certificates.theodi.org/en/about/badgelevels
[9]http://index.okfn.org/dataset/
[10]https://uk-air.defra.gov.uk/data/data_selector
[11]The complete dataset can be download at https://uk-air.defra.gov.uk/data/data_selector?q=695574#mid
[12]https://www.ietf.org/rfc/rfc4180.txt
[13]https://uk-air.defra.gov.uk/assets/documents/The_Data_Verification_and_Ratification_Process.pdf
[14]http://www.scottishairquality.co.uk/data/data-selector

comparing emissions data published by different countries: for example, the EPA in the USA[15] uses the label "ug/m3 SC" for Volatile PM2.5, while the UK uses "ugm-3."

3. IMPROVING THE QUALITY OF OPEN DATA: CHALLENGES AND OPPORTUNITIES

We must consider the extent to which the quality of data in open data portals can be improved given the variety of such data (in terms of domains covered and intended use by "anyone ... for any purpose") and the typically subjective, use oriented view of quality [Wang and Strong 1996]. We believe this highlights an opportunity for data quality researchers to consult with the open data community to document the quality issues they experience, with the aim of identifying two categories of quality metrics: those generally relevant to all open datasets; and those relevant to the various types of data that are routinely published. Examples of the former include completeness of the dataset, representational consistency (including how data and missing data are described), accessibility, conformance to file formats, and metrics developed by [Neumaier et al. 2016] to assess the quality of metadata in repositories as part of improving discoverability; an example data type specific metric would be *Volatile PM2.5 should be greater than 0*, which is potentially relevant to multiple publishers and datasets.

While both types of metrics could be used by data publishers to perform quality assurance, the prominent challenge then becomes how to support such individuals/organisations, who may not necessarily have a technical or data-specialist background, with understanding how the metrics may apply to their data, when and how any assessment processes should be utilised, and how to use the results of those assessments to improve the quality of the data they publish. This requires working with open data publishers to identify suitable intervention points within the publication process, an activity which may benefit from alignment with guides, such as the open data handbook [Open Knowledge 2012], and integration of quality tools into software frameworks, such as the W3C Digital Data Toolkit[16]. A related challenge here is how to effectively utilise the crowd (both data users, and the wider network of workers available via crowdsourcing platforms) in the quality improvement process. As discussed above, there have been instances of ad-hoc, informal feedback from data users to publishers, and more generally, [Acosta et al. 2013] have demonstrated the use of the crowd to identify and resolve quality issues that are beyond the current capabilities of machines. For example, improving the semantic accuracy of bus stop locations. However, the use of the crowd is non-trivial and is not guaranteed to produce beneficial outcomes [Markovic and Edwards], and so raises new research opportunities into the appropriate workflows and incentive/reward models for using the crowd to identify and resolve quality issues with open data.

REFERENCES

M. Acosta, A. Zaveri, E. Simperl, D. Kontokostas, S. Auer, and J. Lehmann. 2013. Crowdsourcing Linked Data Quality Assessment. In *The Semantic Web – ISWC 2013*. Springer Berlin Heidelberg, 260–276.

M. Markovic and P. Edwards. In press. The Challenge of Quality in Social Computation. *Journal of Data and Information Quality* (In press).

S. Neumaier, J. Umbrich, and A. Polleres. 2016. Automated Quality Assessment of Metadata Across Open Data Portals. *Journal of Data and Information Quality* 8, 1, Article 2 (Oct. 2016), 29 pages.

Open Knowledge. 2012. *The Open Data Handbook*. Retrieved Feb 1, 2016 from http://opendatahandbook.org/.

R.Y. Wang and D.M. Strong. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems* 12, 4 (1996), 5–33.

[15]https://www.epa.gov/outdoor-air-quality-data

[16]http://www.w3cdigitaldatatoolkit.com