DATA QUALITY IN THE INTEGRATION AND ANALYSIS OF DATA FROM MULTIPLE SOURCES: SOME RESEARCH CHALLENGES

J. L. Harding

Ordnance Survey, Adanac Drive, Southampton, UK - Jenny.Harding@ordnancesurvey.co.uk

KEY WORDS: Data quality, Geographic information, Analysis, Use context

ABSTRACT:

This paper describes preliminary work to investigate what it means to manage data quality in a simple data integration and analysis prototype for research purposes, where input datasets are from a range of different sources. Consideration is given to how standard elements of spatial data quality (as in ISO 19115:2003) apply in the context of the prototype, which is based on a relatively straight forward ‘house hunting’ scenario. Based on initial findings the paper aims to position further work, identifying a series of research questions around needs for improved data quality management and communication in analytical processes involving geographic information. While not providing solutions or raising novel issues it is hoped the paper may serve to add support for more applied and user focused data quality research in the area of analytics.

1. INTRODUCTION

Many applications of geographic information involve some kind of decision making or deriving insight based on the information available. From viewing a map for trip planning, to carrying out complex spatial analyses of geographic data, a fundamental consideration is – how fit are the data for the specific use and user? Often the problem or decision requires data from different sources to be used together. Within navigation systems for instance, route network, place name and address data may be integrated with live traffic information. Using such systems, a decision on route choice is likely to be more critical for an ambulance driver than for a tourist and the consequences more severe if the combined routing information is wrong in some way. Providing information about data quality that is relevant to use context and communicated in a meaningful way, can allow users to assess reliability and fitness for their purpose.

Moving beyond established GIS applications to uses of geographic information in what is loosely referred to as “Big Data” analytics and other analysis where geographic information is not necessarily the central focus, the end user is interested in answers to queries or insights from bodies of data but might not themselves be involved in the data processing. In this way users of analytical output may be even more removed from the source data, but could still find information concerning data quality helpful to assess reliability of analytical output and thereby enable them to reduce uncertainty or risk in their use of the output. Trends to realise new value and insight from data through Big Data analytics and Business Intelligence Systems, involving data integration from multiple sources, data mining and visualisations of outputs for example, all present questions around quality of inputs and outputs and how significant are aspects of data quality for a given use context.

System developers are recognising that applying data quality processes are important to success in Big Data analytics (see for example http://info.talend.com/4pillarsbigdata.html) and with location being increasingly seen as a valuable dimension within Business Intelligence systems, there is growing focus on how geospatial standards can be utilised to enhance these systems (e.g. OGC, 2012). While a survey of industry managers and analysts suggested that data quality was not perceived as a major obstacle to the uptake of business analytics when compared to managerial and cultural factors (LaValle et. al., 2011) , it remains an area about which increased awareness is needed to inform information use and reduce risk of misuse (Gervais et. al., 2009). For providers of analytical services (where data delivered to client users are in the form of query/analysis results) this means being able to understand and manage data quality starting with input data, through the analytical processes to the output data quality and presentation of that data.

This paper describes preliminary work to investigate what it means to manage data quality in a simple data integration and analysis prototype for research purposes, where input datasets are of disparate types and from a range of open data sources. Consideration is given to standard elements of spatial data quality (as in ISO 19115:2003) and how they apply in the context of the prototype, which is based on a relatively straight forward ‘house hunting’ scenario. While not a critical application, misleading output information would be of little value to users and undermine trust. Based on initial findings the paper aims to position further work, identifying a series of research questions around needs for improved data quality management and communication in analytical processes involving geographic information.

2. RESEARCH APPROACH

2.1 Data Processing Aims

An overall aim of the research prototype was to demonstrate the feasibility of using linked data approaches (see e.g. http://data.gov.uk/linked-data) to integrate geo-referenced data of different types from different sources and use NoSQL database technologies to store large volumes of data and retrieve spatial query results efficiently via a web-based user interface. The benefits of these technologies include their scalability and flexibility to handle evolving content in both structured and unstructured formats. In outline, the development process involved: acquiring datasets of third party data relevant to a selected user scenario (see 2.2) together with relevant internal datasets (from Ordnance Survey); converting these data into linked data 'triples' consisting of expressions made up of named nodes and the relationship between them in RDF (Resource Description Framework; W3C, 2004) to establish linkages between the datasets; loading to a 'triple store' and providing a user interface to create queries across the linked data store as well as to present information returned in response to user selected criteria.

Credibility of the output results would in part depend on qualities of the various source data and any data transformation in the creation of linked data or in the analysis process invoked by a user defined query. It was therefore important to understand as far as possible the quality of source data, any changes to data quality in the data transformation processes, and be able to communicate information about the quality of output information in a way meaningful to the user given the type of query or analysis enabled through the user interface. In this way both 'internal data quality' of source data as provided by the dataset creators and 'external data quality' of outputs as communicated to the end user through the interface of the prototype were within scope of this study.

2.2 The User Scenario

Prototype development focused on a simple scenario of a family moving to a new area and needing information about different neighbourhoods to help them focus their house hunting. Based on setting a range of criteria reflecting their priorities, the user would be presented with a list of unit postcodes (areas encompassing on average around 15 addresses) which meet these criteria within the limits of their geographical search area. For purposes of this research, with emphasis on technical proof of concept, the research project team identified likely criteria of interest in a family house hunting scenario. These included: proximity to schools; proximity to points of interest including pubs and food outlets (for convenience or to avoid), doctors' (GP) surgeries, supermarkets.

2.3 Data Quality Elements Relevant to the Scenario

Qualitative assumptions were made, as follows, about minimum levels of data quality that matter in the use context:

Positional accuracy: For the schools and other points of interest, an address point location would be sufficient to determine presence within a postcode. For other criteria, for example values for house prices, crime rates, deprivation levels, a value or range of values applicable at postcode resolution would be sufficient. In this way spatial granularity of data was more significant for some of the required data elements than positional accuracy of individual data points; that is to say the positional information of much of the attribute data needed to be related to postcodes, areas of coarser spatial granularity than the more precise (to nearest metre) position of an address.

Attribute accuracy: It was important that all criteria were represented by attribute values true to reality to a level of detail appropriate to the scenario. For example in the case of schools, attribution as primary or secondary education needed to be correctly applied.

Temporal validity: For all criteria the most up to date data was required. Actual creation or capture date and update schedules were expected to be different between datasets, so as a minimum requirement the creation date of the dataset and assurance that the data was of the latest data release needed to be known. Data for house prices, for example, might be misleading if more than 2 years old, whereas levels of deprivation are subject to slower change and older data (as long as the most recently published) may be still relevant. Data for some criteria such as house prices and crime statistics might relate to a specified time period, in which case the bounding dates to which they apply need to be known.

Logical consistency: Each dataset needed to be logically consistent within itself and according to its data model or specification in order to facilitate translation to the linked data format (RDF) used in the prototype. Even if used without translation to RDF, logical consistency errors would impede loading to a database and running data queries or analyses.

Lineage/provenance: At a minimum for the user scenario, knowing where the data has come from in terms of source or creating organisation can help provide a basis on which the user can judge their level of trust in the output information. Understanding how that data was created is probably unnecessary in this particular scenario, but data transformations that happen within the prototype application could improve or degrade source data quality and maybe significant enough to communicate to the end user. Recording lineage within the data process is therefore an important internal consideration that may require some level of description alongside data that is output by the prototype.

Completeness: For all criteria it is important to know that query output is based on complete data, meaning that postcodes not listed in query output are absent because they genuinely do not meet the selected criteria rather than because there are data omission errors in source datasets. It is important internally to the prototype therefore that source datasets are complete with respect to their specification, or at least to recognise where datasets may not be expected to be complete. For example if data was captured from voluntary sources it cannot be expected to be complete.

Further considerations concerned coverage (geographic coverage for England was needed) and geo-referencing system used. An internal requirement of the prototype was for all source data to be georeferenced in some way (e.g. by National Grid coordinate, latitude and longitude or postcode) or to have unique identifiers enabling linkage to a geo-referenced source (e.g. a Unique Property Reference Number - UPRN).

3. RESULTS

3.1 Data Sources

For purposes of the proof of concept, open government datasets were sought with content potentially suitable to serve the house hunting scenario at the assumed minimum quality levels. The datasets used were mostly sourced from www.data.gov.uk but also included a commercial points of interest dataset. In addition Ordnance Survey OpenData and commercial products were used in order to present results in the user interface against a zoomable map backdrop.

Data type required	Source/provider	Data set
House prices	HM Land Registry	HMLR House Price
Deprivation levels	Office for National Statistics	Index of Multiple Deprivation
Crime statistics	UK police forces	Crime statistics
Primary and secondary schools locations	Department for Education	EduBase
GP surgery locations	Organisation Data Service/NHS	NHS organisations
Points of interest (various)	POINTX	Points of Interest
Zoomable backdrop map and gazetteer data to enable place based search	Ordnance Survey	OS OpenData™, OS OnDemand data

Table 1. Data required for the prototype and sources used

3.2 Input Data Quality and Uncertainties

Generally, across dataset sources used, the availability of metadata for data quality was very limited, either with the dataset or in associated documentation. The following summarises uncertainties about levels of quality with respect to elements listed at 2.3 above.

Positional accuracy: In most cases the type of georeferencing used in a dataset (e.g. postal address, postcode centroid) indicated the degree of positional accuracy or granularity of data to be expected. In all cases, however, correct positioning of data could not be verified unless it were to be compared with alternative sources of position for the same feature. Alternative sources would need to be of different provenance to afford an independent comparison. This was not undertaken within the prototype.

Attribute accuracy: Likewise, correct attribution could not be verified unless it were to be compared with alternative sources of the attribute for the same feature. This was not undertaken within the prototype.

Temporal validity: For most sources metadata was provided about the date range that the dataset applied to. Some also provided a dataset creation date or publication date. Uncertainty remained over whether the dataset was the most up to date available unless information about update schedule was also provided.

Logical consistency: Few errors were found in logical consistency during the conversion of source data to RDF. This conversion process was itself a way of validating logical consistency where required. Errors that did occur in source data were due to postcode syntax (e.g. S016 0AS instead of SO16 0AS).

Lineage/provenance: For all datasets, provenance in terms of source organisation or publisher name was directly obtainable together with license terms and in some cases advice on attribution statements to be used. Information about lineage in terms of the data creation process was in most cases less accessible. For some datasets however, for example the Index of Multiple Deprivation, the data creation process can be found in separate technical documentation (CLG, 2011).

Completeness: Where datasets in effect provide complete coverage of a choropleth type, a contiguous coverage of polygons would be expected with a value per attribute for each polygon. For these kinds of data therefore errors of omission or commission should be detectable on ingesting the data. Only the Index of Multiple Deprivation was of this type. For the other datasets involved, completeness could not be verified.

Coverage: All the sources used provided information on geographic coverage for the dataset, usually by country name rather than geometry of extent.

3.3 Data Processing causing change to Data Quality

Source data was changed in terms of required fields being converted to linked data, where data was not already in this format. In the case of logical consistency errors, this conversion process resulted in removal of errors and thereby improvement to logical consistency. Such change due to data processing becomes part of data lineage within the prototype. A record of this process was produced for one of the input datasets, HM Land Registry house price data, using W3C provenance vocabularies (W3C, 2013a) in order to demonstrate the potential for managing lineage in linked data structures.

3.4 Communicating quality of output query results

Given the needs of the user scenario outlined above and uncertainties in many elements of source data quality, it was decided in this prototype to focus on those elements of most certainty and relevance to the user scenario. These were the provenance and temporal validity of the data sources. Search results returned in the user interface were accompanied by a table showing: data sources used to provide the results; date of creation or publication; date range of the data if applicable.

4. DISCUSSION

4.1 Summary of Experience from the Prototype Study

By reporting data provenance and relevant dates with the results of user defined queries, the user of the prototype in this study at least has some basis for making their own judgment as to whether the information presented are suitable aids to their house hunting. Nevertheless, in terms of providing a service from acquiring source data through to delivering query or analytical results, uncertainties exist in many of the elements of source data quality as described above. In more rigorous or critical analytical scenarios, improved certainty about data quality may be necessary to enable fit for purpose outputs and enable user evaluation of risk. What can be done to reduce these uncertainties?

4.2 Source Data Quality as Input

We probably have to accept that creators of potentially useful data cannot all be relied upon to provide quality metadata in accordance with standards such as ISO 19115. Lack of complete metadata is a common issue, as highlighted by geospatial data experts surveyed on their approach to dataset selection (Lush et. al., 2012), particularly in terms of provenance, lineage and licensing information. In addition, recommendations from within the user community, data provider reputation and data providers' comments on uncertainty and error estimates within their data were found to influence these specialist users' perceptions of quality. As found in the present study, some additional aspects of quality metadata not included in current standards would be helpful to know about source data, namely the resolution or granularity of the data, and dataset update schedules or intervals. In terms of positional and attribute accuracy, in the absence of quality statements provided with source data, an independent means of verification could help identify levels of uncertainty within the data and provide a basis for representing and communicating uncertainty to the end user, when important to output data usability. Uncertainty is inherent in much geographic data (as for example discussed by Couclelis, 2003; Duckham et. al., 2001) both in terms of position and application of classifications to real-world things , yet this aspect of data quality is not represented in standards for data quality (Goodchild, 2007).

4.3 Data Quality in Data Processing

In the case of missing quality metadata, systems for data integration and analysis need means of independently assessing some aspects of the data's quality when this is important to output data usability. Depending on the type of analysis to be carried out, uncertainty in the spatial, temporal and thematic dimensions of data used may each have impacts on the results of analysis. Zargar and Devillers (2009) review research that has linked the relative importance of these uncertainty dimensions to types of GIS operations and show how the communication of data quality information can effectively be linked to users' applications of operations. Where quality metadata is available, Devillers et. al. (2007) go further in proposing a tool based on a multidimensional cube of compiled data quality information to provide data experts with meaningful information about known spatial data quality to support the required analysis.

Reporting lineage information from data source to output results could be significant in some use contexts. The graph structure of RDF allows storing of provenance metadata using W3C provenance vocabularies (W3C, 2013a). Further prototype development is needed to test whether this is more advantageous than storing metadata separately in tables, for example based on the Data Catalogue Vocabulary (DCAT; W3C, 2013b). Also it is important to consider implications of data processing lineage for presenting attribution statements relating to analytical outputs as well as for original source data used.

4.4 Output data quality

Where available, most quality metadata associated with source data tends to be created and expressed in a producer-centric way (Goodchild, 2007) and does not necessarily assist potential users in selecting suitable data. Investigating this issue, an analysis of information collected from customer interviews and feedback emails (Boin and Hunter, 2007) found that metadata was often confusing to data consumers. Opinions on suitability of a dataset were sometimes derived from actual data content and comparisons with other information or ground truth, rather than quality metadata from the data supplier. For users of just the outputs from analytical services, data quality information needs to be communicated alongside analytical outputs with respect to relevance in the use context.

4.5 Questions arising from this study

A number of research questions relating to data integration and analytics are put forward from this short study. These may not be new and some will already be subjects of research elsewhere. The intention here however is to identify some priorities for improved data quality management and communication as part of analytical applications involving geographic information.

4.5.1 Verifying source data: Where quality metadata is lacking or insufficient for source data, how can source data be verified to identify areas and levels of uncertainty within the data? Further to this: Can source data be automatically verified against source data specifications (e.g. for positional accuracy, attribute accuracy, completeness)? In other words, how well does the data conform to its capture/creation specification? Can source data content and quality be automatically verified against other sources of data?

4.5.2 Handling uncertainty in data processing: How can uncertainty and vagueness in geographic data be handled in data integration and analysis between different data sets? How can linked data structures handle geographic data uncertainty?

4.5.3 Confidence levels and communicating data quality: How can confidence levels in *source data* quality be represented and communicated effectively for different use contexts? How can confidence levels in quality of *output data* (resulting from analyses of data integrated from different sources) be determined and communicated effectively for different use contexts? How can inherent uncertainties or vagueness in source and/or output data be represented to the user in order to inform their decision making? How can data quality information be communicated effectively in different types of use contexts – what matters, what language and what type of visualisation of quality information is meaningful to the user in order to help them assess risk of data use in decision making?

5. CONCLUSION

Preliminary work to investigate what it means to manage data quality in a simple data integration and analysis prototype has been explored in this paper. Standard elements of spatial data quality (as in ISO 19115:2003) provide a useful basis for considering what elements of data quality are significant in a particular use context, and for identifying the presence/absence of quality metadata associated with source data, but could usefully be extended to include factors of data granularity, uncertainty in data (spatial, temporal and thematic) and update or release schedules of data. These elements as categories all have relevance to the value of geographic information within developing Big Data analytics and Business Intelligence Systems, involving data integration from multiple sources, analyses and visualisations of outputs. With users of analytical output remote from the source data and internal analytical operations, relevant data quality information presented in a meaningful way is needed to enable users to establish confidence or gauge risk in their use of the output. To this end it is important for analytics service providers to understand as far as possible the quality of source data, any changes to data quality in data transformation and analysis processes, and be able to communicate information about the quality of output information in a way meaningful for the context of use.

The paper has aimed to position further work, identifying a series of research questions around verification of source data quality, handling uncertainty in data processing and communicating meaningfully about the quality of output data. While not providing solutions or raising novel issues it is hoped the paper may serve to add support for more applied and user focused data quality research in the area of data analytics.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the work of the Ordnance Survey research team which developed the prototype referred to in this paper: Lucy Diamond, John Goodwin, Glen Hart, David Holland, Mark Pendlington, Andrew Radburn.

REFERENCES

Boin, A.T., Hunter, G.J., 2007. What communicates quality to the spatial data consumer? In: *Proceedings of the International Symposium on Spatial Data Quality 2007*. 13-15 June, Enschede, The Netherlands

CLG, 2011. The English Indices of Deprivation 2010. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/6320/1870718.pdf. Viewed 12/04/13. Department for Communities and Local Government.

Couclelis, H., 2003. The Certainty of Uncertainty: GIS and the Limits of Geographic Knowledge. *Transactions in GIS*, 7(2), pp.165-175.

Devillers, R., Bédard, Y., Jeansoulin, R., Moulin, B., 2007. Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data" *International Journal of Geographical Information Science*, 21(3), 261-282.

Duckham, M., Mason, K., Stell, J.. Worboys, M., 2001. A Formal Approach to Imperfection in Geographic Information. *Computers Environment and Urban Systems*, 25, pp89-103.

Gervais, M., Bédard, Y., Levesque, M-A., Bernier, E., Devillers, R., 2009. Data Quality Issues and Geographic Knowledge Discovery. In H. Miller and J. Han (Eds) *Geographic Data Mining and Knowledge Discovery*. CRC Press, Boca Raton, pp 99-111. .

Goodchild, M., 2007. Beyond Metadata: Towards User-Centric Description of Data Quality. *International Symposium on Spatial Data Quality 2007*, 13-15 June, Enschede, The Netherlands.

ISO, 2003. ISO19115:2003 Geographic information – Metadata. International Standards Organisation.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M., Kruschwitz, N., 2011. Big Data, Analytics and the Path from Insights to Value. *MIT Sloan Management* , 52(2).

Lush, V., Bastin, L., Lumsden, J., 2012. Geospatial Data Quality Indicators. *Proceedings of Accuracy 2012,* 10-13 July 2012, Florianópolis, Brazil

OGC, 2012. OGC White paper: Geospatial Business Intelligence (GeoBI). https://portal.opengeospatial.org/files/?artifact_id=49321. Viewed 19/04/13

W3C (2004) Resource Description Framework (RDF). http://www.w3.org/RDF/ . viewed 12/04/13

W3C (2013a) Provenance Working Group. http://www.w3.org/2011/prov/wiki/Main_Page. Viewed 12/04/13

W3C (2013b) Data Catalog Vocabulary (DCAT). http://www.w3.org/TR/2013/WD-vocab-dcat-20130312/. viewed 12/04/13

Zargar, A., Devillers, R., 2009. An Operation-Based Communication of Spatial Data Quality. *International Conference on Advanced Geographic Information Systems & Web Services*, pp.140-145.