3rd International Workshop on Spatial Data Quality, February 2020

Understanding the importance of Provenance from the Perspective of a Geospatial Decision-Maker

Authors: Xxxxxxxx Xxxxxxxxxx[1], Xxxxxx Xxxxx[1], Xxxxxxx Xxxxxxxxxxx[2], Xxxxxx Xxxxxx[3], Xxxx Xxxx[3]

Keywords: Provenance, Decision-making, Usefulness

Abstract:

Information derived from geospatial sources are used in decision-making in various sectors such as in defence (Franklin et al. 2013; Roy et al. 2017), in government (Harding 2006; Sutanta et al. 2016; Scott and Rajabifard 2017) and in non-government organisations (Crooks and Wise 2013; Quill 2018). However, decision-makers do not always have an easy way to decide whether to make use of the given information in their decisions – and if so, how much can they rely on them. A factor that may influence reliance on information for decision-making is well-documented provenance[4] of the information (Ma et al. 2014). Provenance is defined as the "*information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness*" (W3C 2010). It is frequently referred to as lineage, pedigree, parentage, genealogy and filiation (Buneman et al. 2001; Simmhan et al. 2005). There is thus a specific interest in whether presenting important factors of provenance alongside the delivered information, can assist decision-makers to be able to make informed decisions. This abstract presents the preliminary results of an investigation into this aspect of provenance[5].

A core challenge in evidence-based decision-making is to prevent information overload. It is thus important to find out what provenance factors are required, providing the decision-makers only with sufficient context without over-burdening them with excessive details. The first step of any approach to tackle this challenge includes developing a better understanding of related concepts – what is provenance, and what are the current factors suggested as being an important component of provenance. Research shows that data quality and metadata factors are of high importance to make provenance information more useful. This in turn leads to the development of a theoretical framework to underpin work on identifying which data quality and metadata factors are potentially relevant to decision-makers interested in the provenance of their data.

The analysis of the related concepts indicates that although provenance does not entirely correspond to metadata, these concepts (provenance and metadata) are usually linked (W3C 2010). Provenance is often described as the process to detect the lineage and the derivation of data (Alkhalil and Ramadan 2017). Yue et al. (2011) state that lineage and provenance often overlap, with both being used to describe the same information. Provenance also evaluates the data quality and reproduces processes (Simmhan et al. 2005; Moreau and Foster 2006; Chen et al. 2014; Closa et al. 2017).

[1] Xxxxxxxxxx xx xxxxx, xxxxxxxxxxxxx xxx xxxxxxx xxxxxxxxxxx, xxxxxxxxxx xxxxxxx xxxxxx (xxx)
[2] Xxxxxxxxxx xx xxxxxxxxx, xxxxxxxxxx xxxxxxxx xxxxxx (xxx)
[3] Xxxxx xxx xx xxxxxxxx, xxxxxxx xxxxxxx xxx xxxxxxxxxx xxxxxxxxxx
[4] Provenance information can answer to questions such as who created the information, when it was created, why it was created, when it was updated, who own the information.
[5] The presented work has been ethically approved by the UCL Research Ethics Committee until 15th July 2020.

Interoperability of diverse environments thus can be increased (W3C 2013). The level of detail described in provenance can determine how much quality can be assessed (Simmhan et al. 2005). Provenance can also identify relationships between different objects, trace them back, providing thereby the big image of a situation (Chen et al. 2014). Therefore, it can help a user to assess fitness for purpose for a specific application, by providing a description of the origin of the data as well as the processes implemented to bring data in the current form (Closa et al. 2017).

However, much of the work cited above relates to a producer centric view of provenance. To develop a more user-centric view of the problem – and address issues relating to information overload due to the complexity of current standards, interviews have then been conducted to further understand the decision-maker perspective on this challenge as well as their actual needs. For these semi-structured interviews, participants are selected from various sectors amongst the geospatial network of the research study. The selected stakeholders represent a wide range of sectors of decision-makers, making use of geospatial information products (geospatial decision-makers). Each interview was around 40 minutes long and covered topics including geospatial information, metadata and presentation techniques.

Once the interviews were transcribed, thematic analysis was selected as a user-friendly method of qualitative data analysis (Braun and Clarke 2012). This involves a six-phase approach (proposed by Braun and Clarke, 2012), including code generation and theme identification. The outputs of the analysis are examined in the NVIVO software, which supports the annotation and coding of qualitative data and presented through reports as well as scatter diagrams and other graphical representations. Preliminary findings include a set of factors identified as important, several suggestions to present them through provenance and additional challenges that can influence decision-makers' trust.

These preliminary results highlight the importance of taking into account the decision-makers' needs when presenting provenance information and will help develop a focus on the important factors that should be presented as provenance accompanying the received information. Based on this results, online surveys will be distributed to a larger number of participants that is not possible to participate in the interview study, providing immediate data validation and faster response rates (Sue and Ritter 2012; Díaz De Rada and Domínguez-Alvarez 2014). This information - i.e. the usefulness of the provenance information for information derived from geospatial information – will thus form part of an enhanced provenance framework, with the next stages of the work focussing on usability and trust. A number of low and high-fidelity prototypes will be then developed to present provenance information according to the decision-makers' preferences. The developed prototypes will be also evaluated through usability tests where the stakeholders will have to interact with several tasks, trying to assess if the provenance information is presented in a usable way as well as if their trust level is increased.

Acknowledgments

This work is funded by a PhD studentship co-sponsored by the UK Defence Science and Technology Laboratory and the UK Engineering and Physical Sciences Research Council. The authors would like to thank both funders for their support.

References

Alkhalil A, Ramadan RA (2017) IoT Data Provenance Implementation Challenges. Procedia Comput Sci 109:1134–1139. doi: 10.1016/j.procs.2017.05.436

Braun V, Clarke V (2012) Thematic Analysis.

Buneman P, Khanna S, Tan W-C, Chiew W- (2001) Why and Where: A Characterization of Data

Provenance Recommended Citation Why and Where: A Characterization of Data Provenance Why and Where: A Characterization of Data Provenance ?
Chen P, Plale B, Aktas MS (2014) Temporal representation for mining scientific data provenance. Futur Gener Comput Syst 36:363–378. doi: 10.1016/j.future.2013.09.032
Closa G, Masó J, Proß B, Pons X (2017) W3C PROV to describe provenance at the dataset, feature and attribute levels in a distributed environment. Comput Environ Urban Syst 64:103–117. doi: 10.1016/j.compenvurbsys.2017.01.008
Crooks AT, Wise S (2013) GIS and agent-based models for humanitarian assistance. Comput Environ Urban Syst 41:100–111. doi: 10.1016/j.compenvurbsys.2013.05.003
Díaz De Rada V, Domínguez-Alvarez JA (2014) Response Quality of Self-Administered Question-naires: A Comparison Between Paper and Web Questionnaires. Soc Sci Comput Rev 32:256–269. doi: 10.1177/0894439313508516
Franklin AL, Mott T, Williams THL (2013) Coproduction in the U.S. Department of Defense: Examining how the evolution of geographic information systems (GIS) expands non-traditional partner engagement. Policy and Internet 5:387–401. doi: 10.1002/1944-2866.POI345
Harding J (2006) Vector Data Quality: A Data Provider's Perspective. In: Fundamentals of Spatial Data Quality. pp 141–159
Ma X, Fox P, Jacobs K, Wample A (2014) Capturing provenance of global change information. Nat Clim Chang. doi: 10.1038/nclimate2141
Moreau L, Foster I (2006) Provenance and Annotation of Data. Chicago, IL, USA
Quill TM (2018) Humanitarian Mapping as Library Outreach: A Case for Community-Oriented Mapathons. J Web Librariansh 12:160–168. doi: 10.1080/19322909.2018.1463585
Roy SE, Kase EK, Bowman H (2017) Crowdsourcing Social Media for Military Operations. Assoc Comput Mach 23–27.
Scott G, Rajabifard A (2017) Sustainable development and geospatial information: a strategic framework for integrating a global policy agenda into national geospatial capabilities. Geo-spatial Inf Sci 20:59–76. doi: 10.1080/10095020.2017.1325594
Simmhan YL, Plale B, Gannon D (2005) A Survey of Data Provenance Techniques.
Sue VM, Ritter LA (2012) Conducting online surveys. Sage Publications
Sutanta H, Aditya T, Astrini R (2016) Smart City and Geospatial Information Availability, Current Status in Indonesian Cities. Procedia - Soc Behav Sci 227:265–269. doi: 10.1016/j.sbspro.2016.06.070
W3C (2010) What Is Provenance - XG Provenance Wiki. https://www.w3.org/2005/Incubator/prov/wiki/What_Is_Provenance. Accessed 28 Jul 2018
W3C (2013) PROV-Overview. https://www.w3.org/TR/prov-overview/. Accessed 27 Jul 2018
Yue P, Wei Y, Di L, et al (2011) Sharing geospatial provenance in a service-oriented environment. Comput Environ Urban Syst 35:333–343. doi: 10.1016/j.compenvurbsys.2011.02.006