Tracking Editing Processes in Volunteered Geographic Information: The Case of OpenStreetMap

Carsten Keßler, Johannes Trame and Tomi Kauppinen

Institute for Geoinformatics, University of Muenster, Germany
carsten.kessler | johannestrame | tomi.kauppinen @uni-muenster.de

Abstract. With an increasing number of applications building on OpenStreetMap, data quality is becoming a pressing issue. Data provenance gives useful hints that facilitate data quality assessments based on the features' persistence. However, this requires a detailed analysis of the editing history and the corresponding contributors. In order to make this provenance information explicit, we introduce a provenance vocabulary for OpenStreetMap and show how to annotate OpenStreetMap data using this vocabulary. We specify the different types of patterns that can be found in the provenance of features. This paper therefore lays the foundation for approaches to assess data quality that build solely on the intrinsic information collected in the OpenStreetMap database, using the trustworthiness of features as a proxy function for data quality.

Keywords. Provenance, Editing Patterns, Social Semantics, Volunteered Geographic Information, OpenStreetMap

1 Introduction

Volunteered Geographic Information (VGI) [7], such as found in OpenStreetMap (OSM), is increasingly attracting attention for professional use. Applications that build on OSM data include wayfinding [15] and location-based services[1], and the data is being augmented and combined with other information [3]. With a growing number of applications building on OSM, data quality [6] becomes an important issue. The capturing process for commercial geodata is aligned to quality guidelines, allowing safe statements about accuracy, consistency, lineage and completeness. This is not possible for OSM—and most other VGI—due to the large number of lay contributors. Quality control is put in the hands of the community, following the Wikipedia approach: Missing, outdated or fraudulent information is assumed to be fixed by other members of the community.

While these community-based approaches to quality control work well on the whole, applications often require an assessment of the quality of a specific feature. Trust and reputation models have been proposed as proxies for data quality [5,14] in the absence of the metadata and quality metrics that come

[1] See http://wiki.openstreetmap.org/wiki/List_of_OSM_based_Services.

with commercial geographic information. Sztompka defines trust as "a bet" an individual makes "about the future contingent actions of others" [17, p. 25]. A user of VGI therefore makes a bet about its quality based on the contributors' reputation, putting *informational trust* [4] in the data they use.

This paper lays the foundation for such trust and reputation models for OSM by making the data provenance explicit. We introduce a provenance vocabulary and show how to annotate OSM data using this vocabulary. We analyze the editing patterns that emerge during the collaborative mapping process. Such patterns emerge when single features in OSM develop over time, with input events from different users. By comparing different versions of a feature from the OSM history, we can identify recurring patterns for the correction of tags and geometry, which can then serve as input for a trust and reputation model.

In the next section, we review relevant related work. Section 3 presents a provenance vocabulary for OpenStreetMap. Section 4 builds on this ontology to specify recurring editing patterns, followed by concluding remarks in Section 5.

2 Related Work

Analyses on provenance in OpenStreetMap are constrained by the information covered by the OSM data model, which defines the basic types *node*, *way* and *relation*. Nodes are single points with information on lat/lon, a unique ID, the current version number, the last editing user, timestamp, and the ID of a *changeset*, i.e., a collection of edits. Finally, a set of *tags*, consisting of simple key-value pairs, describes the node thematically; see, e.g., `http://www.osm.org/browse/node/740777363`. *Ways* are defined correspondingly, with a set of additional *nd* elements that refer to the nodes that define its geometry. Polygons are defined as ways whose first and last nodes are identical. The number of relations in OSM is still comparatively small, so that we do not consider them.

Trust is a phenomenon inherent in online communities. Bishr and Kuhn [5] proposed to use trust as a proxy measure for quality of geospatial data, pointing out that quality of information is subjective and reflects fitness for use. Previous work on *content-driven reputation systems* focused on textual, unstructured contents, e.g., for Wikipedia [1,11]. Such systems increase a user's reputation if her edits are persistent. Vice versa, a user's reputation decreases if their changes are revised quickly. These approaches compute assessments of the data quality based on data *provenance* [2], eventually combined with explicit user feedback. While metadata standards such as Dublin Core cover some aspects of data provenance, there is recent research on ontological approaches for representing [8], querying [9] and publishing [10] provenance information. By distinguishing the (abstract) data from a concrete serialization, the provenance vocabulary [8] allows for a detailed capturing of all steps that yield a file, including the actors involved. In the following, we extend this provenance vocabulary with elements specific to OpenStreetMap. We show how to annotate OSM data based on this vocabulary and show how these annotations can serve as input for a trust model.

3 Feature Provenance

Contributions to OpenStreetMap are organized in *changesets* that contain new, updated, and deleted features that have been edited by a specific user in one session. OSM does not compute any differences between consecutive versions of a feature, but always stores full copies. Any information on what has been changed by a specific user therefore has to be derived from the implicit provenance information in a feature's history. This *recordable provenance information* [8] should be made explicit to enable provenance-based querying of the OSM data. We follow a *data-oriented approach* [16], as we focus on the origins of specific data items, instead of the processes that generate the data. The provenance vocabulary[2] allows to make the lineage of any online data explicit [8]. It defines *actors* that perform different kinds of *executions* that eventually lead to different kinds of *artifacts*. In the following, we extend this vocabulary to cover the provenance information in OpenStreetMap, and discuss our design decisions.

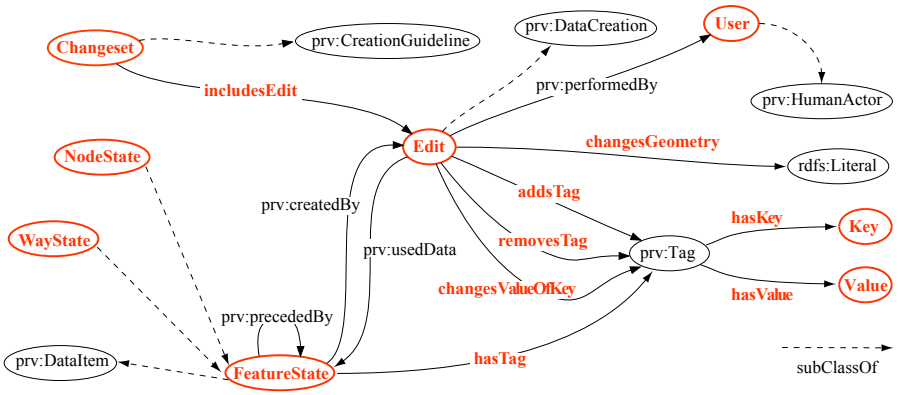

Fig. 1. Overview of the OSM provenance vocabulary. Classes and properties in red have been added to the original provenance vocabulary (`prv:` namespace).

Figure 1 gives a high-level overview of the extended vocabulary.[3] `Edit` is the central class that links changes on a specific `FeatureState` (i.e., a specific version of a node or way) to the `User` who made the change and to the corresponding timestamp.[4] `Changesets` act as containers for a collection of n edits affecting n features that were uploaded together by a specific user. Note that there is no element in the OSM data model corresponding to the `Edit` class. Individuals of this class can wrap any information on tag and geometry changes

[2] See http://trdf.sourceforge.net/provenance/ns.html.
[3] See http://carsten.io/osm/osm-provenance.rdf.
[4] The timestamp is not shown in the figure; it is attached to `prv:DataCreation`'s superclass `prv:Execution`, see http://purl.org/net/provenance/ns#.

on a specific feature that were committed by a user at one point in time. The type of the actual change is encoded in the type of the corresponding property (`removesTag`, `addsTag`, `changesValueofKey` and `changesGeometry`) pointing to *what* has changed.

Modeling `Edit` as a class enables statements about instances (e.g, who was involved and when). Moreover, the different versions of a feature still form a *provenance graph* [8], since they are connected via the `prv:preceededBy` property. Figure 2 shows an example of such a provenance graph, making the edits between the two different versions of a feature explicit:

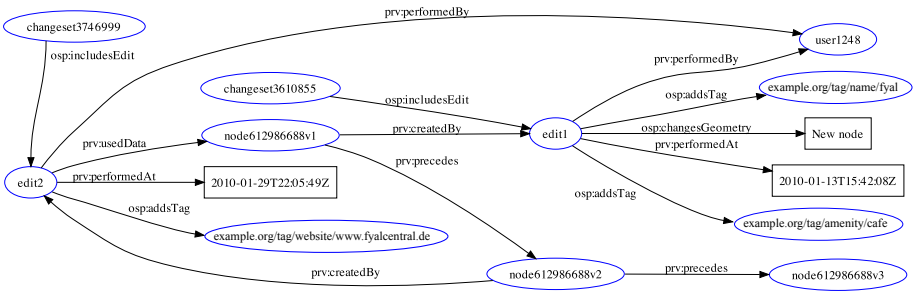

Fig. 2. Simplified provenance graph, generated from `http://carsten.io/osm/prov-example.ttl`.

4 Editing Patterns

The OSM provenance vocabulary facilitates explicit statements about the lineage of features in OpenStreetMap. It hence lays the foundations for analyses of editing event patterns in the creation of OSM data. In previous research, we have developed tools for the explorative analysis of OSM feature provenance in a Web application[5], which facilitates the identification of recurring patterns [18]. These types of patterns are crucial when assessing the trustworthiness of a feature and the reputation of the involved OSM users, respectively. We focus on patterns that can be derived from the provenance information introduced in Section 3 using logical inference and simple heuristics.

Similar to collaborative filtering approaches, the underlying hypothesis is that the community-based interaction of a large number of users results in globally observable patterns, which can be utilized to make reputation assertions on an individual level. OpenStreetMap, however, does not have any mechanism for explicit feedback. When existing information on a feature is changed, this can nonetheless be intuitively interpreted as negative feedback: if the current

[5] See `http://giv-heatmap.uni-muenster.de`

information on a specific feature is regarded as incorrect by a user, she implicitly provides negative feedback by updating this information, such as a tag or the geometry. As the patterns discussed in the following will serve as input to trust and reputation models, we also discuss each pattern's influence on these models.

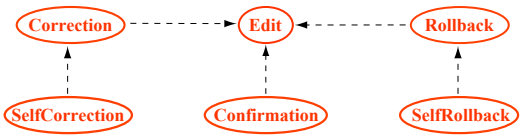

Fig. 3. Patterns as subclasses of the `Edit` class, as specified in our provenance vocabulary (see Figure 1).

Confirmations. As changes in the geometry or tags of a feature may point to quality improvements, we will treat the absence of such negative feedback as a confirmation of the feature data correctness. A user who just adds a new tag to a node, for example, implicitly confirms the correctness of the existing tags. Every edit that fulfills the following conditions will be reclassified as a confirmation:

$$\texttt{FeatureState(?f1)} \wedge \texttt{FeatureState(?f2)} \wedge \texttt{Edit(?e)} \wedge \texttt{usedData(?e,?f1)} \wedge \texttt{createdBy(?f2,?e)} \wedge \texttt{precededBy(?f2,?f1)} \wedge \texttt{!changesTag(?e,?x)} \wedge \texttt{!removesTag(?e,?y)} \wedge \texttt{!changesGeometry(?e,?z)} \rightarrow \texttt{Confirmation(?e)}$$

More sophisticated methods could even rate feature quality based on single tags or the geometry using this method. These *passive* confirmations only affect a feature's trustworthiness without immediate influence on the contributors reputation. Any *active* participation in editing map features should initially increase a user's reputation, as discussed below.

Corrections. While users can gain reputation by adding information to the OpenStreetMap database, a mechanism is required that *decreases* a user's reputation if wrong or low-quality information is added. Such faulty information can only be identified if it is subsequently corrected by another user.[6] The following rule formalizes the reclassification of an edit as a correction in the case where an added tag is corrected subsequently:

$$\texttt{FeatureState(?f}_1\texttt{)} \wedge \texttt{FeatureState(?f}_2\texttt{)} \wedge \texttt{precededBy(?f}_2\texttt{,?f}_1\texttt{)} \wedge \texttt{Edit(?e}_1\texttt{)} \wedge \texttt{Edit(?e}_2\texttt{)} \wedge \texttt{createdBy(?f}_1\texttt{,?e}_1\texttt{)} \wedge \texttt{createdBy(?f}_2\texttt{,?e}_2\texttt{)} \wedge \texttt{changesValueOfKey(e}_2\texttt{,?t)} \rightarrow \texttt{Correction(?e}_2\texttt{)}$$

[6] See, e.g., http://giv-heatmap.uni-muenster.de:4434/history/node/369332524

Rollbacks. Corrections should intuitively decrease a user's reputation if they revert a feature to a state before. Such rollbacks[7] point to the fact that an update was faulty from the point of view of the user who made the rollback. A rollback is hence defined by three subsequent versions of a feature, where the first and last of the three subsequent versions of a feature are identical:

$$\texttt{FeatureState(?f}_1\texttt{)} \wedge \texttt{FeatureState(?f}_2\texttt{)} \wedge \texttt{FeatureState(?f}_3\texttt{)} \wedge \texttt{precededBy(?f}_3\texttt{,?f}_2\texttt{)} \wedge \texttt{precededBy(?f}_2\texttt{,?f}_1\texttt{)} \wedge \texttt{Edit(?e}_2\texttt{)} \wedge \texttt{createdBy(?f}_2\texttt{,?e}_2\texttt{)} \wedge \texttt{equalState(?f}_1\texttt{,?f}_3\texttt{)} \rightarrow \texttt{Rollback(?e}_2\texttt{)}$$

If corrections or rollbacks are performed by the same user who made the initial edit, they need special handling. Particularly, the user's reputation should not decrease for fixing her own errors. This preliminary set of patterns can serve as input to trust and reputation measures in future work.

5 Conclusions

In this paper, we have introduced a provenance vocabulary for OpenStreetMap that commits to an existing, generic provenance vocabulary. This vocabulary allows us to make implicit provenance information about the lineage of features in OpenStreetMap explicit and classify them according to recurring editing- and co-editing patterns. We have shown how these patterns can be inferred from the provenance information using Horn rules that reclassify any instances of the `Edit` class to the corresponding pattern classes. The next step for future work will be the development of trust and reputation functions based on these reclassified instances. Trust values for features can be computed based on a feature's editing history. Likewise, a user's contributions and the subsequent edits on the affected features by other users can be used to measure user reputation. In order to come to trust and reputation measures that reflect the actual reputation of users within the community, empirical research is required that evaluates how informational trust and user reputation are judged. This may include analyzing environments where the community interacts that are not part of our model yet, such as the OpenStreetMap wiki.

Acknowledgments

Funded by the Int. Research Training Group on Semantic Integration of Geospatial Information (DFG GRK 1498) and the SimCat II project (DFG JA1709/2-2). Special thanks to Olaf Hartig for his feedback on the provenance ontology.

References

1. B. Adler and L. De Alfaro. A content-driven reputation system for the Wikipedia. In *Proceedings of the 16th international conference on World Wide Web*, pages 261–270. ACM, 2007.

[7] See, e.g., http://giv-heatmap.uni-muenster.de:4434/history/node/88875206

2. D. Artz and Y. Gil. A survey of trust in computer science and the semantic web. *Web Semantics*, 5:58–71, June 2007.
3. S. Auer, J. Lehmann, and S. Hellmann. LinkedGeoData – Adding a Spatial Dimension to the Web of Data. In *Proceedings of 7th International Semantic Web Conference (ISWC)*, 2009.
4. M. Bishr and K. Janowicz. Can we Trust Information? – The Case of Volunteered Geographic Information. In A. Devaraju, A. Llaves, P. Maué, and C. Keßler, editors, *Towards Digital Earth: Search, Discover and Share Geospatial Data 2010. Workshop at Future Internet Symposium*, September 2010.
5. M. Bishr and W. Kuhn. Geospatial Information Bottom-Up: A Matter of Trust and Semantics. In S. I. Fabrikant and M. Wachowicz, editors, *The European Information Society – Leading the Way with Geo-information*, Lecture Notes in Geoinformation and Cartography, pages 365–387. Springer-Verlag Berlin Heidelberg, 2007.
6. N. Chrisman. The error component in spatial data. *Geographical information systems*, 1:165–174, 1991.
7. M. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, August 2007.
8. O. Hartig. Provenance Information in the Web of Data. In *Proceedings of the Linked Data on the Web (LDOW) Workshop at the World Wide Web Conference (WWW), Madrid, Spain*, April 2009.
9. O. Hartig. Querying Trust in RDF Data with tSPARQL. In *Proceedings of the 6th European Semantic Web Conference (ESWC), Heraklion, Greece*, June 2009.
10. O. Hartig and J. Zhao. Publishing and consuming provenance metadata on the web of linked data. In *Proceedings of The third International Provenance and Annotation Workshop*, Troy, NY, U.S.A, 2010.
11. S. Javanmardi and C. Lopes. Modeling trust in collaborative information systems. In *International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2007)*, pages 299–302. IEEE, 2008.
12. T. Kauppinen, K. Puputti, P. Paakkarinen, H. Kuittinen, J. Väätäinen, and E. Hyvönen. Learning and visualizing cultural heritage connections between places on the semantic web. In *Proceedings of the Workshop on Inductive Reasoning and Machine Learning on the Semantic Web (IRMLeS2009), The 6th Annual European Semantic Web Conference (ESWC2009)*, May 31 - June 4 2009.
13. T. Kauppinen, J. Väätäinen, and E. Hyvönen. Creating and using geospatial ontology time series in a semantic cultural heritage portal. In *S. Bechhofer et al.(Eds.): Proceedings of the 5th European Semantic Web Conference 2008 (ESWC 2008), LNCS 5021, Tenerife, Spain*, pages 110–123, 2008.
14. C. Keßler, K. Janowicz, and M. Bishr. An Agenda for the Next Generation Gazetteer: Geographic Information Contribution and Retrieval. In *GIS '09: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Seattle, Washington*, pages 91–100, New York, NY, USA, 2009. ACM.
15. Z. A. Schmitz, S. and P. Neis. New Applications based on Collaborative Geodata – the Case of Routing. In *XXVIII INCA International Congress on Collaborative Mapping and SpaceTechnology, Gandhinagar, Gujarat, India*, 2008.
16. Y. L. Simmhan, B. Plale, and D. Gannon. A survey of data provenance in e-science. *SIGMOD Rec.*, 34:31–36, September 2005.
17. P. Sztompka. *Trust: A sociological theory*. Cabridge University Press, Cambridge, 1999.
18. J. Trame and C. Keßler. Exploring the Lineage of Volunteered Geographic Information with Heat Maps. In *GeoViz 2011, Hamburg, Germany*, 2011.