Kauppinen, Tomi

Linked Science Enablement via Semantic Interoperability and Spatial Data Mining

Published in:
Proceedings of the 2nd Data Management Workshop

DOI:
10.5880/TR32DB.KGA96.6

Published: 01/01/2015

Document Version
Publisher's PDF, also known as Version of record

in: Constanze Curdt and Christian Willmes (eds.) (2016):
Proceedings of the 2nd Data Management Workshop. Kölner Geographische Arbeiten, **96**. Köln/Cologne: 31-37

LINKED SCIENCE ENABLEMENT VIA SEMANTIC INTEROPERABILITY AND SPATIAL DATA MINING

T. Kauppinen
Department of Computer Science, Aalto University School of Science
tomi.kauppinen@aalto.fi

Abstract

We are now witnessing a large-scale need for the use of spatial information. Examples range from monitoring of deforestation in the Amazon to everyday applications for navigation and map-based visualizations. However, the central theories for Geographic Information Science (GIScience) need to be developed further in order to support the range of useful applications of geographic information in the society. For this there is a need to understand whether the study of scientific assets and their spatial, temporal and thematic could help to reveal useful new theories. The task is to all of these assets like publications, scientific data, methods, tools or tutorials – and represent their links to each other and to space, time and themes. The core question thus is: can we interconnect all scientific assets? This calls for efficient methods to answer questions of where, when, what, who (and even why) about each asset. Linked Data provides means for both the representation and accessing of data about the scientific assets on the web. This way it becomes possible – likely for the first time – to study on a large scale what kind of stories the data about scientific assets has to tell. Spatial data mining together with ontological reasoning can help us make aggregations, visualizations, abstractions, and thus allow for exploration of massive collections of scientific data and related assets. If we achieve in interconnecting different assets then we can achieve Linked Science where not only different assets are connected but also different disciplines. In this paper we discuss the role spatial data mining, semantic interoperability, vocabularies and visualization to support enabling of Linked Science. We also provide examples from our different Linked Science projects to illustrate the ideas.

Keywords: Linked Science; Spatial Data Mining, Ontological Reasoning, Information Visualization

1. Introduction

We as Geographic Information Science (GIScience) researchers need to both study and show the benefits of geographic information for the society. This includes opening up, and linking of spatial data, and improving information usability at all levels, e.g. via establishing Future Spatial Data Infrastructures (SDI) (Díaz et al. 2012). The other challenge is whether and how can we support sense making and storytelling from data. The following questions emerge to be answered: What data should be linked together, or merely what can be linked i.e. where are the limits? How do we optimally aggregate data, given its semantic, spatial, and temporal dimensions for supporting decision making?

Scientists are requested to produce new knowledge. The society needs to have novel ways to interact with the scientific results in a shared manner online to make science more usable, reproducible, and transparent. This calls for link discovery at a large scale. For instance, how do we optimally support discovery of links between very different phenomena (e.g. ecological, economical and social) and the spatial and temporal characteristics of links? What are the roles of linked data and semantic interoperability for geospatial domains of the society? How can we maximally support reuse of scientific data for reproducible research? This calls for hybrid methods combining spatial data mining and semantic interoperability.

Weather sensors, for instance, provide massive amounts of raw data. This (often big) raw data has to both conceptualized (to get instances of heavy rain, low visibility and snow) and mined (to get co-occurrences of heavy rain, low visibility and snow) in order to discover weather storm instances (Devaraju & Kauppinen 2012). Finding these crucial weather instances is one of the need to timely serve massive amounts of people online. At the same time it is crucial to ensure the usability

Fig. 1: Central parts of Münster, Germany rendered with data from OpenStreetMap

of information served. Usability of information – be it thematic, spatial or temporal – is often neglected resulting inefficiency or even wrong decisions.

This challenges the current systems not just because of the Volume (of data) and Velocity (the data grows), but also because of Variety (of different sensor types)[1]. This calls for

- Linking of sensor data together by their spatial, temporal and thematic properties. Are space and time optimal integrators of data? What is the role of semantics in integrating data?
- Higher level conceptualization of data to make information usable. Which kind of conceptual structures support these higher level representations?
- Visualizations of information using maps, timelines and semantic nets

Data about representations of interconnections between scientific assets should allow for analysing trust (see Sztompka 1999) and reputation of actors involved (see Mezzetti 2004) and thus support to understand scientific quality underlying the variety of assets. The increased availability of Volunteered Geographic Information (VGI) (Goodchild 2007) allow for analysis of edits to reveal underlying semantic, qualitative and geometric changes to support computing of trustworthiness of VGI data and reputation of the contributors involved (d'Antonio et al. 2014). To exemplify the richness of VGI data let us introduce a representation of a tiny part of OpenStreetMap in Figure 1, representing the central parts of the city of Münster in Germany.

Spatial data mining supports finding of interesting relationships and characteristics, e.g. via clustering techniques (Ng & Han 1994). For instance, by using clustering combined with qualitative models we now know that classification of VGI features have local characteristics (Ali et al. 2014). By using intra-user agreement analysis we also know that VGI contributors have a slight agreement about classification of geospatial features (Ali et al. 2014). These both findings can be taken into consideration when creating platforms for using VGI data (e.g. for navigation or urban planning). Crucial in analysing data from crowdsensing projects is to combine spatial data mining techniques with qualitative representations of relationships between geospatial features. Linked Data (see e.g. Bizer et al. 2009) allows for creating network structures representing relationships between different resources. These relationships (links) can be a result of data mining or reasoning procedures or be as an input for them.

This paper is structured as follows. Section 2 discusses space and time as integrators of data. Section 3 follows by discussing the role of vocabularies and their terms for aligning spatial data for data mining and comparisons. Section 4 illustrates the Linked Science community support and results of activities so far. Section 5 finishes the paper with concluding remarks.

1 For evidently first mention and discussion about these "3 Vs" see LANEY (2001)

*Fig. 2: Browsing of LODUM data by buildings (*Kessler & Kauppinen *2012)*

2. Space and Time as Integrators of Data

Space and time support to organize knowledge (Janowicz 2010) by enabling efficient integration. Two resources sharing references to the same space (or even time) can be analysed together to find whether they have some other interesting mutual linkages. Technical sensors produce amazing amounts of data, and in an increasing speed. For humans trying to interpret, and interact with these sensor observations there is a need for new ways of abstracting and conceptualizing information. The goal is to make information usable, and via that make the systems communicating that information also usable, thus calling for integration at different levels and ways.

Figure 2 gives an example of using space and time as an integrator. Here the amounts of publications done over the time in different university buildings of the University of Münster are visualized as the building heights (see Kessler & Kauppinen 2012). Space serves as an integrator since the publication amounst are collected (and thus integrated) by buildings. Time serves as the integrator when we create slots (e.g. one year or one month) and collect information (e.g. publication amounts) together by using that time slot.

In order to study how and if space and time serve as integrators of data the following types of questions arise:

- How resources are related to space and time?
- What events happen and where?
- Which processes can we evidence?
- What data do we need to study the environment?
- How to link ecological, economical and social data?
- Which links would be useful?
- Which links we aim to discover and how?
- How can we cluster nearby phenomena via spatio-temporal data mining?
- When are space and time efficient integrators of data and when not?

There are several challenges in answering these questions. We need intelligent systems that realize spatio-temporal mining and reasoning about distributed sensor data in order to get a higher-level representation of the spatial aboutness. For instance, explicit representation of spatial aboutness of scientific publications allow for visualizing the overview of which places are covered by a certain discipline (see Atanassova et al. 2015).

Moreover, many organizations make the use of place names as the reference system. However, place names are disambiguous. How do we establish a reference system for place names which would solve the ambiguity problem? Moreover, temporal references are tricky to establish since temporal information is often uncertain, imprecise or it does not exist. How do we relate such temporal information together and provide valid referencing mechanisms?

Further on, the following tasks for supporting information usability arise:

- What kind of spatio-temporal links there are?
- What are the methods needed (data mining, identity resolution, semantic reasoning) for discovering and representing those links?
- Essentially: what different phenomena are correlating with each other?
- At which spatial, temporal or conceptual scales this happens?

3. Supporting Data Mining by Integrating Data via Shared Terms

Vocabularies can also serve as thematic integrators. TEACH vocabulary (Kauppinen et al. 2012), for instance, allows for describing courses and their respective resources (like teacher, student group, room, semester) as Linked Data. TEACH is currently already in use in the Aalto University in Finland and in the University of Muenster in Germany to encode course information. These allow for efficient analysis of course offerings via data mining and other statistical tools. For instance, one can aggregate[2] different lectures of a given university department to months of a year and thus get an overview of amount of lectures over time. Figure 3 shows this idea applied to retrieve lecture amounts of the Department of Real Estate, Planning and Geoinformatics at the Aalto University. Mining data for creating this aggregation of lecture amounts to months benefits from following three links types:

- link between a lecture and its timestamp,
- link between a course and its lectures
- link between a department and a courses it offers

We can retrieve the aggregation using this link structure with a SPARQL query as follows:

```
SELECT (xsd:string(?month) as ?month) (count(?course) as ?lectures)
WHERE {
    aaltodata:dept_T2050 aiiso:teaches ?course.
    course teach:arrangedAt [ical:dtstart ?start] . }
GROUP BY (substr(str(?start),6,2) as ?month)
ORDER BY ?month
```

where we make use of the following prefixes (i.e. short forms) to avoid the need for using the full URIs in the query:

```
PREFIX aaltodata: <http://data.aalto.fi/id/courses/noppa/>
PREFIX aiiso: <http://purl.org/vocab/aiiso/schema#>
PREFIX teach: <http://linkedscience.org/teach/ns#>
PREFIX ical: <http://www.w3.org/2002/12/cal/icaltzd#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
```

2 See online tutorial available at http://linkedscience.org/tutorials/analyzing-and-visualizing-linked-data-from-the-aalto-university-with-r/

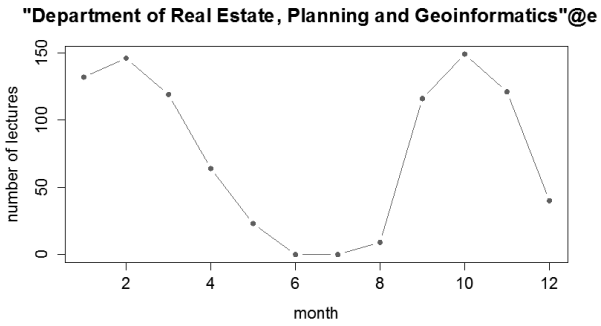

Fig. 3: Aggregation of lectures to months (1-12) of one university department

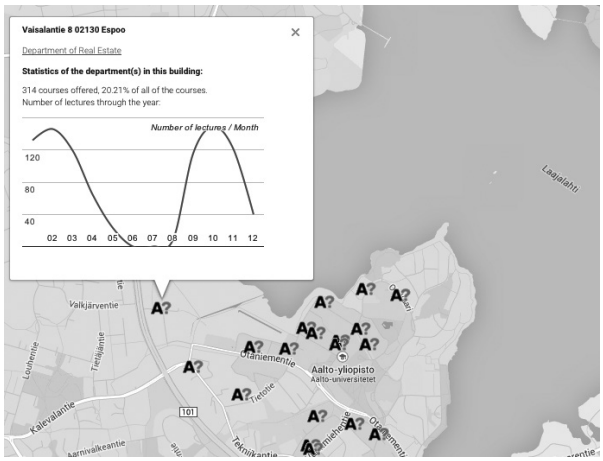

Fig. 4: Information about lecture amounts linked to a building

Mining and querying of link structures allows for further following of links. This becomes very interesting when we follow the link from department to the building it occupies. This allows for visualizing the amounts of lectures by buildings (see Fig. 4), thus supporting to understand the use of spaces in different university buildings. In this visualization one can click any of the building symbols (marked with Aalto University logos) to get an overview of lectures in that particular building over a year. Like we can evidence here the curve for the Spring semester (months 1-5) is quite different from the Autumn semester (months 9-12).

By analyzing pair-wise correlations (e.g. in R) it is clear to be the case for most of the university buildings, thus revealing interesting phenomenon about the generic pattern of use. Once revealed, the university management can make decisions accordingly, e.g. to use the free auditorium spaces for other purposes.

Comparing different things calls for understanding what exactly can be compared. Linked Data approach supports this well by allowing direct queries of data via online SPARQL endpoints. For instance, Figure 5

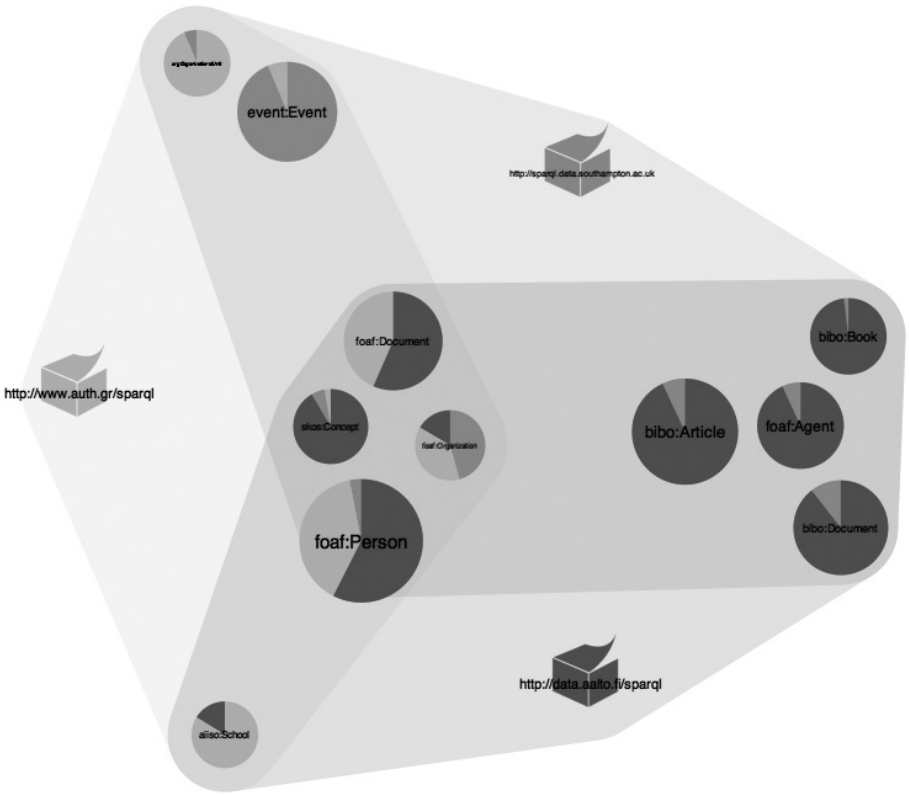

*Fig. 5: Comparing term use in three different universities with V2 (*Alonen et al. *2012)*

shows a comparison of the use of joint classes from three different university datasets with V2 visualizer[3] (Alonen et al. 2012) for understanding the similarities and differences. As we can evidence, only three classes (foaf:Person, skos:Concept and foaf:Organization) are used to type instances in all three datasets. Thus one can compare all three universities only on information about these classes but not about other things unless similar concepts are aligned to each other.

4. Community as the Key Actor for Linked Science

We established the LinkedScience.org community in 2011 for sharing scientific assets like data and tutorial materials[4]. Thus the idea is to show what Linked Science can mean in practice. In other words, it is a community-driven effort to show how data and other scientific assets can be efficiently annotated, distributed, reused, mined, computed, explored, and visualized. The tutorial materials have been tested already in six courses in three different universities, and in six international tutorial events. There is also evidence that the materials are in active use in other teaching occasions. Indeed, it seems that online tools greatly support the running of teaching occasions.

Tools support students to sketch and initiate different representations as Linked Data, validate the syntactic correctness of them, and then forward the data to visualization platforms to see and feel the first results. In this it is important to use real data and provide real information usability and visualization examples to evaluate the theoretical methods. Online tools together efficient face to face meetings support students to produce interesting results, even in form of new spatial data mining methods, interaction tools, visualizations, and vocabularies. These results become assets as themselves, and thus be used as learning materials and building blocks for the follow-up courses and tutorials to serve the community at large.

5. Conclusions

In this paper we argue that hybrid methods combining support for semantic interoperability and spatial data are crucial in making sense of spatial data and to serve as an enablement for Linked Science. When applied to scientific assets this supports storytelling and sensemaking of interconnections. The idea of Linked Science is to inter-

3 http://data.aalto.fi/V2
4 See http://LinkedScience.org/tutorials for a growing number of online tutorial materials

connect scientific assets. By doing that it becomes a new type of science in itself to enable studying of many disciplines and assets together rather than separately. The opportunity here is that the analysis of links radically supports creation of insight about underlying processes. Promisingly we now evidence an emerging research community around the Linked Science idea.

The questions regarding space and time get increasingly highly important. There is a growing need to efficiently be able to mine patterns from spatial data. Similar challenge remains for handling time: for example when certain processes and events took place. For this reason the research agenda for the years to come should include development of mining and reasoning techniques to deduce high-level descriptions of the processes and events, and evaluation of interaction mechanisms to make ideal use of them. This calls for developing hybrid methods for Geocomputation, capable of

1) supporting the usability of information by semantic interoperability, reusable vocabularies, and spatio-temporal and semantic reasoning,
2) spatial data mining, aggregation, classification, identity resolution and link discovery, and
3) sensemaking and information visualization given the cognitive abilities of people.

We argue that tutorials and workshops organized provide essential stepping stones for growing the community and in general to support learning and collaboration at LinkedScience.org. Furthermore, different online data quality assessment tools using Linked Data are helpful for automatic data assessment and visual analysis. This brings a set of questions for the future work to tackle. Which data about scientific assets should be opened? How would you like to use visual analytics for that data? What new spatial data mining methods should be developed for analysing networked data? Overall, what data needs sensemaking? How can science be more open? How can we build scientific environments that truly support transdisciplinary problem solving?

Acknowledgements

I would like to acknowledge my colleagues in different Linked Science related projects both in Germany and in Finland for fruitful collaboration and discussions that has led to many examples presented and referred to in this paper. I would like to especially mention Werner Kuhn, Alkyoni Baglatzi, Anusuriya Devaraju, Carsten Keßler, Krzysztof Janowicz, Johannes Trame, Umut Tas, Miika Alonen, Salli Hukkinen, Willem Robert van Hage, Osma Suominen, Iana Atanassova and Marc Bertin. My research during 2010-2012 was partially funded by the International Research Training Group on Semantic Integration of Geospatial Information (DFG GRK 1498, see http://irtg-sigi.uni-muenster.de). My research from 2012 onwards has been partially funded by a Aalto University postdoctoral research grant.

References

Ali, A.L., Schmid, F., Al-Salman. R., Kauppinen, T. (2014): Ambiguity and Plausibility: Managing Classification Quality in Volunteered Geographic Information. Proceedings of the ACM SIGSPATIAL GIS 2014, USA, Dallas, Texas.

Alonen, M., Kauppinen, T., Suominen, O., Hyvönen, E. (2013): Exploring the Linked University Data with Visualization Tools. The Semantic Web: ESWC 2013 Satellite Events. Lecture Notes in Computer Science, 7955: 204-208.

Atanassova, I., Bertin, M., Kauppinen, T. (2015): Exploitation de données spatiales provenant d'articles scientifiques pour le suivi des maladies tropicales. Gestion et Analyse des données Spatiales et Temporelles (GAST'2015), 15ème conférence internationale sur l'extraction et la gestion des connaissances (EGC-2015). Luxembourg.

Bizer, C., Health, T., Berners-Lee, T. (2009): Linked-Data – The Story So Far. International Journal On Semantic Web and Information Systems, Special issue on Linked Data, 5 (3): 1-22.

d'Antonio, F., Fogliaroni, P., Kauppinen, T. (2014): VGI Edit History Reveals Data Trustworthiness and User Reputation. Proceedings of the 17th AGILE Conference on Geographic Information Science, Connecting a Digital Europe through Location and Place. Castellon, Spain.

Díaz, L., Remke, A., Kauppinen, T., Degbelo, A., Foerster, T., Stasch, C., Rieke, M., Schaeffer, B., Baranski, B., Broering, A., Wytzisk, A. (2012): Future SDI – Impulses from Geoinformatics Research and IT Trends. International Journal of Spatial Data Information Research (IJSDIR), 7: 378-410.

Devaraju, A., Kauppinen, T. (2012): Sensors Tell More than They Sense: Modeling and Reasoning about Sensor Observations for Understanding Weather Events. International Journal of Sensors, Wireless Communications and Control, Special Issue on Semantic Sensor Networks, 2 (1).

Goodchild, M.F. (2007): Citizens as sensors: Web 2.0 and the volunteering of geographic information. Geofocus, 7: 8-10.

Janowicz, K. (2010): The role of space and time for

knowledge organization on the semantic web. Semantic Web, 1 (1): 25-32.

Kauppinen, T., Trame, J., Westermann, A. (2012): Teaching Core Vocabulary Specification. http://linkedscience.org/teach/ns. 2014-01-17.

Kessler, C. & Kauppinen, T. (2012): Linked Open Data University of Muenster – Infrastructure and Applications. Demos of the 9th Extended Semantic Web Conference (ESWC2012), May 2012, Heraklion, Greece.

Mezzetti, M. (2004): A Socially Inspired Reputation Model. In: Katsikas, S.K., Gritzalis, S., López, J. (eds.): Public Key Infrastructure: First European PKI Workshop: Research and Applications, EuroPKI 2004, Samos Island, Greece, June 25-26, 2004. Lecture Notes in Computer Science, 3093: 191-204. doi: 10.1007/978-3-540-25980-0_16.

Laney, D. (2001): 3D Data Management: Controlling Data Volume, Velocity and Variety. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf. 2014-01-17.

Ng, R. & Han, J. (1994): Efficient and Effective Clustering Methods for Spatial Data Mining. In: Bocca, J. B., Jarke, M., Zaniolo, C. (eds.): Proceedings of the 20th International Conference on Very Large Data Bases (VLDB ,94). Morgan Kaufmann Publishers Inc. San Francisco, CA, 144-155.

Sztompka, P. (1999): Trust: A sociological theory. Cambridge University Press.

Contact information
Tomi Kauppinen
Aalto University School of Science
Department of Computer Science
FI-00076 Aalto
Finland
tomi.kauppinen@aalto.fi
+358 50 431 5789