

Linked Data and geoinformatics

a gap analysis

Position Paper for the workshop 'Linking Geospatial Data'

Frans Knibbe, Geodan Research

frans.knibbe@geodan.nl

<http://www.geodan.com>

Table of Contents

1 Introduction.....	1
2 Why Linked Data and Geoinformatics need each other.....	1
3 Use case.....	3
4 Current issues.....	3
4.1 Semantics.....	3
4.2 Software support.....	3
4.3 Data volume.....	4
4.4 Dataset metadata.....	4
4.5 Application development.....	5
5 The way forward.....	5

1 Introduction

Recent developments in the domains of Linked Data (or the Semantic Web) and Geoinformatics have been largely independent, but mutual interest is clearly growing. Both worlds have a lot to offer to each other, and can make each other stronger. But there are a few gaps between the two that have yet to be bridged. In this document some of those gaps are described, together with possible ways of narrowing or even bridging the gaps.

2 Why Linked Data and Geoinformatics need each other

The field of geoinformatics has always been a somewhat isolated area of information technology, but it did manage to improve thanks to common technological progress. A notable development was the move from isolated file based data storage to storage in relational databases, allowing data to be communally maintained and shared. This was made possible by extensions to existing databases, consisting of definitions of data types, spatial indexes and topological functions. Although this development allowed geographical data to coexist with other types of data in the same storage medium, geoinformatics remained an isolated area for specialists, and applications based on geographical data for a large part remained applications with a pure focus on geography.

Standards for the exchange of geographical data have not managed to break out of isolation, despite being inspired by general standards such as XML and web services. Illustrative of the continued domain constraints is the way the official standard to express geographical data, GML, was developed: geography is not just a data type, it is the complete framework of a data set. In order to exchange data of which geography *could* be a part, all of the data need to be modelled as geography.

Even within the domain of geoinformatics the exchange of data is not without barriers.

Traditionally, the significance of metadata of datasets has been acknowledged, but in practice datasets and their metadata are coupled only very loosely, giving them ample opportunity for being inconsistent. Geoinformatics did come up with the concept of the Spatial Data Infrastructure (SDI), a way of improving integration of datasets and services and application doing something with those data. In practice, different SDI's do not interoperate well, so they too can be considered isolated silos.

So data in geoinformatics in their current state can be described as silos within a silo. Interoperability between different geographical information systems is low, as is interoperability with the outside world. The usefulness of available geographical data would be much improved if geoinformatics could burst out of its enclosing barriers.

So in comes that tireless combatant of data silos: Linked Data. With its universal data model and its promise of having one global database for all data, it seems like the perfect way out of isolation for geoinformatics. Linked Data allows metadata and data to be tightly integrated, semantics to be shared and geography to be freely mingled with other kinds of data. It would allow geographical data to be used not only by services and applications that specialize in geographical data, but by any kind of software. Linked Data could be the means to have much higher returns on any investment in the publication of geographical data.

There is a lot that Linked Data could do for geoinformatics, but a tighter embrace would certainly be beneficial for the Semantic Web too. For one thing, the Semantic Web needs more data. Lots more. Sure, there are many interconnected triples on the web already, and the number is growing, but a large majority of data has not reached the five star level. And that means that Linked Data is not yet where it should be, it has not been able to reach critical mass, the stage where further expansion will be a natural result of that what is already there. For the web of five star data to fulfil its potential, it needs a lot more data. And that means that it needs more things that do something with the data, like turning them into information, knowledge and understanding. Supply and demand will keep each other in balance and will stimulate each other. Geoinformatics can help growth in both areas.

For one thing, there is a massive amount of data that is geographical at its core, or has one or more geographical aspects. A lot of data are already readily available, at many governmental agencies for example. And a lot more is yet to come, thanks to ubiquitous devices equipped with GNSS receivers or sensors with a known location. If it were possible to publish all these data in the semantic web, in a way that preserves the usefulness of locational data, it would mean an enormous increase in the size of the web of data, perhaps even enough to send it snowballing.

On the demand side geoinformatics has something to contribute too. There is a wealth of knowledge available on the art of data visualisation, mostly in the form of maps. When developers and designers of graphical user interfaces for geographical data can depend on a large web of high quality raw data being available, they will surely come up with great ways of demonstrating the added value of linkage and semantics, and increase demand for more good data.

Another important thing that geoinformatics can bring to Linked Data is connectivity. One of the advantages of Linked Data is that data are not only freely available on the web, they are also linked, which greatly improves usefulness. This linkage may be brought about by direct linking: triples that provide direct connections between data sets. Linkage may also come from shared semantics: data sets using the same vocabulary to publish facts are naturally linked. Geography can provide yet another type of linkage: topology. Features that have some kind of topological relationship are linked, whether this link is explicitly coded in the data or not. This means that geoinformatics can add a new dimension to the connectivity of the data web, thereby increasing its awesomeness to new heights.

3 Use case

When the subject of geographical Linked Data is discussed, the term 'use case' is often used. It describes particular way of interacting with data, a way that dictates which features are desirable and which not.

Let us consider just one use case: using data from the global data web in a general web browser. Perhaps the mother of all use cases, perhaps not, but if we succeed in getting geodata to work well for the general case of someone looking for data, finding it and displaying it using a web browser, then a lot of more particular use cases will be covered as well.

4 Current issues

Following is a description of issues that somehow hamper the full incorporation of geographical data in the semantic data web. The list is not meant to be comprehensive.

4.1 Semantics

There is a need for a standard way of expressing things geographical in RDF. At this moment, there are several vocabularies that offer different solutions. While competition between different solutions is a good thing, because it helps in identifying strengths and weaknesses, in the end we need unambiguous ways of denoting geography: a single vocabulary, or a set of closely related vocabularies.

First and foremost, we need consensus on the best way to code a geometry. Just like numbers, texts or dates, geometry is a data type that needs its own generally known way of notation. A good candidate seems to be Well Known Text (WKT), an OGC/ISO standard that is well established, is relatively simple, allows many (all?) kinds of geometry and has a high level of software support.

Next to coding geometry, we need to be able to specify topological relationships. Just as it is clear how two numbers or two text strings can interact, it needs to be clear how two geometries can spatially interact.

Other specifications that would be nice to have are geographical functions that are common in GIS software, but have not been described in vocabularies like GeoSPARQL yet. Functions that perform coordinate transformation or calculate a bounding box can be very helpful because they could be effectively put to use in SPARQL queries.

The publication of the GeoSPARQL standard was a major milestone. Not only did it specify how geometry can be coded and how topological relationships can be expressed, it also came with the stamp of approval of the main authority of geoinformatics, the OGC. That mark can do a lot for acceptance of a vocabulary and for the broader technology of Linked Data.

Still, the process of establishing OGC standards is not the most open, and if there are any perceived flaws in its current version it is not easy to resolve them. Besides that, the OGC is a platform for the geoinformatics industry, with a low participation of experts on Linked Data. The OGC does work together with ISO in publishing standards. Would it be possible for the OGC to have a similar relationship with the W3C, making GeoSPARQL a joint OGC/W3C venture?

4.2 Software support

After looking at specifications for coding and processing geography it comes natural to look at the level of software support for those specifications. The state of affairs in 2013 is very nicely written

up in GeoKnow document D2.1.1 'Market and Research Overview'¹. It shows that support for geographical data in Linked Data software is emerging, but that it is still far from being ready for full scale deployment.

It would be nice to make the research done in the GeoKnow project an annual affair, or maybe to develop it into a more automatic benchmark of available software. And perhaps the tests could be extended, for example including testing federated SPARQL queries using topological relationships. This would help awareness of the status quo, and be an incentive for software developers to keep improving their products.

4.3 Data volume

Geographical data tend to be high in volume, because geometries are encoded as series of coordinates that can be quite long. This means that poor performance of applications and services making use of geographical data is a risk. Fortunately, there are some options to mitigate the risk:

- Use significant digits in coordinates, so there are no superfluous digits in the coordinates. Note that this is good for data quality too.
- Publish multiple geometries with multiple levels of detail, so data consumers can select the appropriate generalisation of geometries, avoiding receiving coordinates with an amount of detail that won't be used.
- Use on or more techniques to compress a response from a server. One can think of compressing the literals, or compressing the entire response.

Further research and experimentation will be welcome in this area, especially when this is a coordinated effort.

4.4 Dataset metadata

Linked Data can improve the provisioning of geographical data by tight integration of metadata. Metadata can play a vital role in searching for data on the web and finding appropriate data sets. Publishing metadata is well established in Linked Data, making use of vocabularies like Dublin Core², the Vocabulary of Interlinked Datasets (VoID)³ and the Data Catalog Vocabulary (DCAT)⁴. But geographical data need certain special metadata, like

- Spatial extent of the data set (bounding box),
- Coordinate reference system (if all geometries use the same),
- Number of dimensions (usually 2, but 3 if height is included in geometries),
- Level of detail (if it is the same for all geometries),
- Positional accuracy (if it is the same for all geometries).

It would be beneficial for exchange of geographical data on the web if metadata elements like these were standardised.

1 http://svn.aksw.org/projects/GeoKnow/Public/D2.1.1_Market_and_Research_Overview.pdf

2 <http://dublincore.org/documents/dcmi-terms/>

3 <http://www.w3.org/TR/void/>

4 <http://www.w3.org/TR/vocab-dcat/>

4.5 Application development

Increased availability of high quality geographical data in high quantities is a matter of supply and demand. On the demand side, there is a need for more applications and services making use of spatial data on the web. Application developers can be enticed to make use of the data web just by the data that is on offer, but currently the level of complexity in the RDF family of standards is an obstacle for many. What would be very nice to have, for development of Linked Data in general, and for giving low threshold access to developers of geospatial applications and services, are ways of interacting with the web that are simpler than SPARQL. A certain loss of richness or functionality is probably acceptable for most cases, especially when it would always be possible to turn to SPARQL for extra functionality. Several endeavours in the Linked Data community have identified the need for simpler access next to SPARQL and are offering solutions. It would be a good thing for developers of software used in geoinformatics to critically follow these endeavours, to try them out and to share experiences.

5 The way forward

This document brings up a few areas in which geoinformatics and Linked Data need to grow towards each other, to break geospatial data out of their silos and to strengthen the web of data. For many issues described the best solution will only be found after communal thought and experimentation. What is needed is for the people and organisations involved to break out of their silos too. It is time to look past organizational and occupational boundaries, to freely discuss problems and possibilities and to share insights. After all, we are all in the same web.