Position Paper Heritage & Location

Towards a geotemporal semantic infrastructure for Dutch cultural historical data

 A position paper for the Linking Geospatial Data conference, 5-6 March 2014, London. Mailto: erfgoedenlocatie@den.nl

Introduction

Historical databases contain an abundance of cultural information. This wealth becomes visible when the information is interconnected, to tell a larger story. The Heritage & Location project enables heritage institutions in the Netherlands to exchange information and develop new services to reach a new public. Heritage & Location is a nationwide project developed by Digital Heritage Netherlands and the results are expected to be up and running in June 2015.[1]

In most historical information, the time frame is of as much importance as the location. Nearly all heritage information is directly or indirectly tied to a place, as well as a time. However, the geographical terms used in these sources often do not match the current names in use. An approach that uses semantic as well as spatial and temporal techniques is required here, in building a solution for a geotemporal semantic web. The potential of this approach is huge. Linking data in every way imaginable, events that hitherto would be isolated can now be shown to be related. Historians, collection managers and the broad public in general can benefit from analysis, bundling and visualisation of information from a wide range of cultural sources.

The geosemantic issue

It seems that a combination of semantic and geospatial is of great importance to disclose the historical information from the databases and documents in wide use - often described as 'information silos'. Inferencing or semantic reasoning is needed to enable meta-searches over different collections, interconnecting them thematically. With heritage data that is not thematically related, geospatial and temporal links are needed to infer their shared spatial and temporal characteristics.

However, this scenario is still mostly theoretical. In practice, an integrated system that can both reason semantically and spatially search is far from complete. The OGC GeoSPARQL and W3C Core Location Vocabulary standards help describe data and helper functions, but as

[1] A blog is kept on http://www.erfgoedenlocatie.nl and experimental services running on http://erfgeo.nl.

Athanasiou et al. have shown[2], the current implementations look experimental at best. None of the current applications scale very well, nor are they particularly stable or particularly well-documented. This is a problem, for the larger the collection of datasets grows, the harder good semantic and geographic filtering is needed. Things are probably in development at the larger semantic software companies, but that leaves a big gap to be filled towards smaller institutions unable to buy into expensive infrastructure schemes.

Apart from the unified input, the geosemantic issue has an output problem as well. The current web mapping frameworks and Geographical Information Systems client software are unable to work with the standardised linked data output formats of most triple stores. The output can be bent into GeoJSON, but that The widely used Web Feature Service protocol in use in GIS currently on the other hand has no software candidate for interoperability with semantic web formats.[3] In short: the semantic web and geospatial world are only sniffing at each other at the moment, which leaves Heritage & Location looking for a solution that is well scalable and stable, provides geospatial, temporal and semantic filtering combined with great interoperability standards for both the semantic web and GIS users.

Geotemporal semantic use cases

Having GeoSPARQL as a standard is a great asset in itself. Questions and responses have been defined and can be referred to. It is very elaborate as well - which leaves GeoSPARQL triple store developers a lot of work to do, work that is perhaps not always as relevant to further the interests of end users. The GeoKnow study by Athanasiou et al. referred to earlier defines 21 queries with a spatial operation.[4] Specifically, these operations can be subdivided in:

1. Find an entity based on a given location - “reverse geocoding”
2. Retrieve geospatial objects within a bounding box or other simple, singular geometry - “range queries”
3. Find all instances (or parts thereof) of one type of spatial object intersecting any instance of another type of spatial object - “spatial join”
4. Find a nearest neighbour on a given location - “nearest neighbour”
5. Combine or buffer separate spatial objects - “spatial aggregation”
6. Not researched by GeoKnow: spatial reprojection, scaling or rotation - “transformations”

A lot of functionality seems tailored to professional GIS users that may not need complex spatial operations on large linked data sets. Linked data is not the data source of choice for GIS pros - they usually make use of desktop clients connected to spatial databases. The ‘classic’ end user of linked data (though not necessarily the Heritage & Location end user) is the public in the broad sense. When someone needs to perform complex spatial operations on data, it is most often on a subset of data and more importantly, it often needs to be visualized in ways that the

---

[2] Athanasiou et al., 2013.
[3] Heritage & Location is working on a Javascript-based adapter on https://github.com/erfgoed-en-locatie/sparql-geojson to provide basic translation from JSON SPARQL results to GeoJSON, but it is far from complete.
[4] Athanasiou et al., 2013: 76 and onwards.

web-oriented semantic systems are hard pressed to perform. So what are the basic needs and user goals?

Heritage & Location tries to match the interests of collection owners and their intended public by supplying services in order to connect both parties. As the specific interest of the people accessing these collections is still somewhat unclear, we make an educated guess of their goals. We imagine tourists visiting a place that spurns their interest in finding out the past uses of the place, house-owners wanting to find out more about the history of their property and surroundings, students writing papers on cultural historical subjects, historians researching artwork and archives enriching their collection data with crowd sourced content. The means by which they try to achieve these goals are myriad, but as far as we are able to deduce from these goals, there are no complex spatial operations involved - only simple ones. There is neither spatial aggregation involved, nor transformation or nearest neighbour processing. For the moment, there isn't even a use case requiring a spatial join.[5] So, it could be argued that the advanced user of cultural heritage datasets would benefit more from a good download service than elaborate spatial search capabilities.

The basic necessities for the public in a broad sense boil down to reverse geocoding and simple range queries, operations that do not need the full range of GeoSPARQL filter functions[6] but can rely on the basic pre-2.0 Web Feature Service filter capabilities[7]. This opens some interesting opportunities.

Alternative approaches

It may be that we are still away from a fully scalable, fully interoperable GeoSPARQL implementation, or maybe it is just around the corner. As it is hard to keep up with the developments, be it planned, intended or currently testing, we have to consider approaches that rely on other combinations of server components. There may be other rewards here: a fully production-ready GeoSPARQL store could be a great business asset, possibly rendering such a solution a costly buy. An approach that combines current applications can perhaps make use of open source components: there are both several semantic systems and geospatial systems that scale very well. But how can these components be combined into a working geotemporal semantic system?

The challenge lies in several aspects. Firstly, the components would have to interface with each other with shared standards and interchange formats. One could consider a combination of semantic and geographic components controlled by an API that separates the query semantics from the queried geographical filter and assembles the output in the desired format - one legible for linked data purposes, but interoperable with GIS web components. The second challenge lies in maintaining a single data store for several components. The last thing

[5] However, this could change: it is not hard to imagine a historian wanting to find photograph objects taken in or around current archaeological monument sites, for example, without having to rely on semantic searches.

[6] Supplied on http://www.opengis.net/def/function/geosparql/.

[7] See http://dev.openlayers.org/releases/OpenLayers-2.13.1/examples/wfs-spatial-filter.html for example. WFS 2.0 filter capabilities also offer spatial joins, see Vretanos 2010: 81.

we need is having to synchronize a semantic/temporal data store and a geographical datastore as separate points or sources of truth. One needs to avoid data duplication.

Several paths can be explored here as a thought experiment. The first requires a well-scalable triple store to handle the semantic and temporal query part, with drivers for integration in a WFS server for the spatial search[8]. Of course, the triple store would be needing a solidly performing spatial index, which is often a problem, particularly with full (multi-)polygon support. It should support semantic reasoning, as we want to offer search results based on thesaurus-related concepts. The main challenge would be in getting a WFS server to consider the triple store a viable data source. For example: the OpenLink Virtuoso triple store comes with an ODBC driver[9], possibly providing an API for a WFS server data store. This API could in theory be tied to MapServer as "Virtual Spatial Data", so that MapServer could expose the data (or at least entity URI's) as WFS interface layer.[10] Since there are no known tests having tried this solution, the performance, stability and scalability is completely unknown.

A second path can be found in adapting a geodatabase with scalable SQL triplifier.[11] There are already several applications developed in this direction, by live mapping of databases to RDF, including one employing geospatial support[12]. However, there doesn't seem to be any support for semantic reasoning, which is to be expected as it would be hard to edit or maintain ontologies in standard SQL database tables using the same mapping strategies. A separate reasoning provider would be needed, using probably a different data source for ontologies, but this would implicate moving away from the Single Source of Truth. However, in this way a combination of the database mapper Sparqlify, including geospatial support, and Jena as reasoning engine could provide the needed functionality. We cannot foresee yet whether this will amount to a scalable and stable solution.

A third approach could be sought in developing or adapting a NoSQL data store to a geotemporal semantic triple store. As a thought experiment, Apache Accumulo[13] has recently seen the development of a geospatial implementation called GeoMesa[14], while Accumulo is also undergoing work on adapting it as a RDF triple store under the name Rya[15]. As a cloud-based storage system, Accumulo is expected to scale very well, but no semantic reasoner is expected in the foreseeable future, so this would have to be developed. Similarly, as a Graph Database Neo4j[16] already has both spatial, RDF and reasoning support although at the moment it is unclear to me whether all these features interact to produce the desired combined functionality and scalability.[17]

---

[8] Such as GeoServer or Mapserver.
[9] See http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VOSClient
[10] http://mapserver.org/input/vector/VirtualSpatialData.html
[11] cf. D2RQ or Sparqlify.
[12] See http://sparqlify.org/
[13] http://accumulo.apache.org/
[14] https://github.com/geomesa/geomesa
[15] Punnoose, Crainiceanu, Rapp (no date).
[16] http://www.neo4j.org/
[17] Some apparently companies seem to move away, see http://blog.genealogysystems.com/2014/01/developer-day-19-boundary-queries-and.html

Conclusion

The Dutch Cultural Heritage is building on the Heritage & Location project to bring its content to the semantic web. Heritage & Location is looking for strategies to make this happen, not only to make cross-domain searches, but also to spatially and temporally enable the cultural heritage data. The current state of affairs, however, leaves the project in a situation where the desired search capabilities using combined semantic, temporal and spatial functionality has no clear-cut server stack candidate. In theory, the GeoSPARQL functions provide all the necessary spatial functions, but current implementations seem to lack either stability, scalability, documentation, semantic reasoning or a combination of these. Perhaps a solution can be found in combining several components that specialise in any (combination) of these requirements, or maybe even a fully developed single package is waiting around the corner.

Literature

All URL links listed in the footnotes are accessed on January 14th unless stated otherwise. These links are not repeated below.

Athanasiou, S./L. Bezati/G. Giannopoulos/et al. 2013: *Deliverable 2.1.1. Market and Research Overview* (accessed January 10th 2014 on http://svn.aksw.org/projects/GeoKnow/Public/D2.1.1_Market_and_Research_Overview.pdf)

Vretanos, P. (ed.) 2010: *OpenGIS Web Feature Service 2.0 Interface Standard* (OpenGIS Project Document OGC 09-025r1, ISO/DIS 19142, accessed January 14th 2014 on http://portal.opengeospatial.org/files/?artifact_id=39967)

Punnoose, R./A. Crainiceanu/ D. Rapp (no date): *Rya. A Scalable RDF Triple Store for the Clouds* (accessed on 14-1-2014 on http://www.usna.edu/Users/cs/adina/research/Rya_CloudI2012.pdf)