Linking Geospatial Data Workshop: 5th - 6th March 2014

Proposed discussion item:
GeoSPARQL – where should it go, and when should it stop?

Authors:
Steve Peters, Department for Communities and Local Government
Bill Roberts, Swirrl IT Ltd

Intro and summary

The purpose of this paper is to stimulate a discussion on:

- Potential improvements to the GeoSPARQL standards, to support more sophisticated management and re-use of geographic LinkedData.
- Future relationships between and the respective roles of triple-stores and spatially-enabled databases as technologies to power more open, integrated geo-spatial data and intelligence.

Our main observations and conclusions are that:

- LinkedData and SPARQL technologies are increasingly being adopted by Government organisations as channels for routinely disseminating significant data assets – examples include: DEFRA's sets on Bathing Water quality; Land Registry's Price Paid series; geographic data from Ordnance Survey and ONS, and ; DCLG's "Open Data Communities" service.
- Geography and "place" are key to unlocking and binding together datasets within and across these services.
- GeoSPARQL therefore has an important enabling role: it could be the foundation for triple-stores as products akin to the spatially-enabled Relational Databases powering the GeoWeb.

However….

- Current GeoSPARQL standards appear to be limited to straightforward spatial querying – e.g. to identify "points in polygons", or touching/overlapping geometries. There are emerging requirements to support more sophisticated use cases, for instance:
    - Areas covered by or overlapping other areas – e.g. LSOAs covered by or overlapping Parish Councils.
    - Transforming Coordinate Reference Systems (e.g. British National Grid to WGS 84).
    - Generalising (or simplifying) lines and polygons.

- Serialising geometry in a wider range of formats, in particular GeoJSON and TopoJSON which arguably are now the “de-facto” standard amongst the software developer and innovator community.

- Take-up of GeoSPARQL in commercial and open source triple-stores appears to be very limited. Further work is needed to understand why, and identify how to improve take-up across the vendor community.

- In our experience, it can be difficult for users to locate and consume geospatial LinkedData in their application. We need better tooling to support geo-LinkedData reuse in a range of scenarios – including desk-top GIS, and web-based mapping services such as OpenStreetMap or CartoDB.

- Alternative geo-spatially enabled databases are well established as both commercial - e.g. Oracle - and open-source - e.g. PostGIS - products, and have a strong track-record in delivering geographic data using open standards.

- A key question is: **should triple-stores, enabled by GeoSPARQL, evolve to provide features and functionality akin to a geo-spatial database?** Our view is yes – not least because it helps to standardise and streamline management of LinkedData assets, plus creates opportunities for innovative new analysis using datasets in the Semantic-Web. However, the reality is that geographic data publishers will continue to operate hybrid triple-stores/ geospatial database solutions for some time. Further work is therefore needed to strengthen integration between these two technologies.

Background: requirements and use-case overview

In the call for participation, we are asked to consider this fundamental question:

> **How can we discover that different facts in different data sets relate to the same place, especially when 'place' can be expressed in different ways and at different levels of granularity?**

This matters to DCLG. Many of our policies and programmes depend on improving intelligence about place at various inter-related spatial scales. For instance:

- The Department's **Neighbourhood Planning** policy gives more power to local communities, allowing them to shape their area and its future. There are currently circa 1,000 local communities actively engaged in developing Neighbourhood Plans.

- The Department is leading policy work on **housing markets**, focussed on delivering more homes to buy and to rent at prices people can afford. This is complemented by work to put local councils and businesses in charge of **economic growth** and bring new business and jobs to their areas

- DCLG is overseeing delivery of more **efficient, cost-effective public services** – including achieving better outcomes for vulnerable people through, for example, DCLG's Troubled Families programme.

Within and across these policy areas, location and spatial intelligence is essentially the glue to bind smarter, more effective data linking and – therefore - enable innovative new, data-driven insights **about different places**.

Developing a better, shared understanding of policy problems commonly involves using data from multiple sources at the local and national levels. For example, we know that Neighbourhood Planning communities are using a diverse set of sources[1], including affordable housing needs assessments, transport impact analysis, landscape, infrastructure, and local facilities.

The range and types of sources with a geographic element is also growing rapidly. For example, Land Registry's Price Paid data on individual property transactions is a useful new resource to improve understanding of housing markets. Police.uk's Street-level crime data is helpful for work on community safety, particularly when linked to related sources – such as postcode or neighbourhood data on demography, from the 2011

[1] See http://mycommunityrights.org.uk/wp-content/uploads/2012/04/Neighbourhood-Planning-Evidence-Base-updated.pdf

Census. New initiatives[2] will lead to publication of significant volumes of new data, particularly at the local level.

Crucially, linking contextual data about place will also involve processing and interrogating data on complex, overlapping geographies. In the mix here are administrative - e.g. Local Authorities and Wards - and statistical geographies - e.g. Super Output Areas, and Postcodes. User-defined areas will be increasingly important too, ranging from the very local – say, a particular housing estate or retail park - to aggregations of pre-defined areas: e.g. Local Authorities participating in Local Economic Partnerships, or LSOAs forming a locally defined region of a town or city.

A role for GeoSPARQL?

We believe that there is a clear and strong argument for triple-stores to evolve and incorporate features and functions available in commercial and open-source geospatial database software products.

From a purely practical perspective, the absence of these features is creating additional overheads for geo-spatial data publishers. For example, DCLG's triple-store contains no geometry data, and is instead limited to holding attributes of individual features – names, codes and so on. The geometry is held in separate multiple external files, in essence to provide a choice of serialisation formats including KML, GeoJSON and SHP.

It appears the GeoSPARQL can help us solve this problem, by enabling geometry to be held inside the triple-store, alongside related attribute data. This would reduce the complexity and overhead of managing geospatial data, plus improve usability – e.g. by creating opportunities for more efficient and innovative re-use.

The additional features and functionality to be supported by GeoSPARQL could include:

- Management– e.g. to add and set spatial data types, or set and transform Coordinate Reference Systems

- Relationships – building on current GeoSPARQL standards, to include new functionality – such as identifying areas covered or overlapped by other areas (e.g. Lower Super Output areas covered or overlapped by Parish Councils).

- Processing– methods to calculate a polygon centroid or area, and generate buffers - with an ability to write results back to the triple-store.

[2] Examples of new iniatives to unlock additional datasets are DCLG's "Code of Practice for Transparency" in Local Government, and the Open Data User Group's work to encourage user requests for data.

- Accessors and measurement – methods to query and retrieve bounding boxes; to test whether polygon coordinate rings are properly closed; or to calculate the number of coordinates in individual polygons.

- Geometry constructors and outputs – to create point, line and polygon geometries from text, and output them in different serialisations (incl. GeoJSON and TopoJSON).

- Editors – to add or remove bounding boxes and points; or to merge/simplify polygons – again with an ability to write results back to the triple-store.