This is a position paper for W3C's "Linking Geospatial Data" Workshop.

Geospatial Data and Linked Data

A few informal observations on Linked Data and GeoData.

Linked Data (considered as a slogan for the practice of publishing RDF-based graph data in the public Web) has a few points of common agenda with various forms of geospatial data.

1. Approaches to the scoping challenge.

A dataset is rarely a pure "GIS" dataset. Rather, geographical aspects pervade almost all datasets. A collection of scholarly articles might be associated with a conference or research institution that has geographical information. A school might appear on a catchment area map. A description of a business might include information about the location of its branches, while a description of an architect or inventor might include information about the location of their works. Geospatial data serves to bind together other datasets, and consequently shares the emphasis on cross-domain information mixing that is also central to the concerns of various RDF, Linked Data and Semantic Web efforts.

Geospatial, by being literally grounded in the real world, provides a mental model and existence proof for the kind of data integration that is attempted across all domains via RDF graph merges. General graph-based frameworks (e.g. Knowledge Graph / Freebase from Google and products from Bing, Facebook, Yahoo and others) are often described as a "map" of real-world relationships.

The RDF "entity graph" approach to the problem of scoping schemas for heterogeneous datasets is to acknowledge that, in practice, all problem domains tend to overlap awkwardly. By adopting a common cross-domain data model, RDF and similar graph-based approaches allow for one or more graph schemas to be combined within a broader framework.

2. Approaches to authority and reference "background" data.

Historically, a lot of geo-data has been closed. In recent years however, there has been a move towards opening of datasets - across various fields including cultural, scientific, government. W3C's RDF standards and the Linked Data community have been central in many of these efforts. A common deployment pattern has been for the creation of machine-readable "reference" pages for various sets of entities (people, places, things), alongside the use of "sameAs" links to help disambiguate descriptions by pointing to well-known pages, including general services such as Wikipedia (Wikidata/DBpedia) and Freebase, or domain-specific sites. So for example in the UK, per-postcode reference pages are published by the Ordance Survey <http://www.ordnancesurvey.co.uk/blog/2010/11/linked-data-at-ordnance-survey/> as Linked Data.

3. Possible shared goals

Both GIS and Linked Data practitioners share a concern for cross-domain data merging, and for having common public reference datasets available in standard formats.

There are various levels of possible collaboration. Starting with the simplest:

W3C RDF's basic graph data model provides a simple, pragmatic formalism that can be used to link GIS data with other datasets. W3C itself also provides a community environment where various groups can come together to collaborate on schemas <http://www.w3.org/blog/data/2014/01/06/vocabularies-at-w3c/>, including but not limited to the Web Schemas effort which guides the evolution of schema.org. Since schema.org (an RDF vocabulary supported by Google, Bing, Yahoo, Yandex) is now appearing on 5+ million domains, these technologies are increasingly mainstream and it is the right time to reconsider their relationship to more traditional geospatial approaches.

Beyond basic maps, it is increasingly important to explore techniques for spatial analysis that integrate standards-based datasets. This is another point of shared concern between communities, in particular W3C's new CSV on the Web group <www.w3.org/2013/csvw/wiki/Main_Page> emphasises the importance of pragmatism regarding data formats. Rather than insisting that everyone always use RDF for everything, the W3C community is exploring mechanisms for using RDF to integrate other existing formats (such as CSV). We expect this to be important for example when combining spatial data with other open datasets. It is not clear at this point whether additional standards are needed, but W3C seems an appropriate environment to have these conversations.

Beyond formal standards and vocabularies, the W3C can also be an environment where practitioners can get together to discuss and develop tools for exploring data graphs within a standards based environment. Tooling, whether in the form of open libraries or application based approaches can help drive adoption of existing and new standards.

Mano Marks <mmarks@google.com>
Dan Brickley <danbri@google.com>