

Talking about Place – Annual Progress Workshop 2012

Break-out session report

The first group discussed **challenges with address data**. Questions that were identified are: 1) what location is characterized by an address, 2) if, and for what purposes, sub-addresses are needed, e.g., building points, 3) whether stronger semantics of a geocode are necessary e.g., entry points as default (?), and 4) could we gather addresses as they are used by people, in contrast (or in addition) to addresses in GNAF data. Some of the challenges for addressing these questions are the following:

PSMA masters already merging independent address data sets (e.g., state governments, Post and Electoral Commission). Each of them is made for a purpose, and accordingly they differ. Adding, for example, a crowd-sourced address data set adds only gradually to the complexity. However, crowd-sourced address data (see, e.g., openaddresses.org) could be filtered by geometric and semantic constraints, trust measures (take use of data into account as a measure of trust, or test number of edits as a trust measure), and other knowledge (e.g., from mining the web for support).

In the same way as producers maintain address data for their purposes, there are users expecting address data suiting their (different) purposes. One challenge might be to provide contextualized addresses.

Assigned addresses appear to be inflexible to changes, and sub-numbering seems to be applied in an unsystematic manner (e.g., subdivisions). Semantic metadata appears to be desirable (e.g., ‘residential’ vs. ‘commercial’). What is in an address beyond a location code: physically, socially, culturally, and emotionally? What’s the address of the future?

Can place databases (gazetteers), point of interest databases, address databases, cadastral databases, property databases be linked by semantically meaningful pointers and if so how? Who will do the linking and who will maintain them?

Update cycles challenge real-time needs (e.g., of 000). If flow through a system is so slow, how can a change in an authoritative data source immediately trigger updates with user copies?

The second group discussed the **challenges with interpreting place names and descriptions**. The group discussed a broad range of issues with the place names themselves. Namely:

There are multiple representations, identities and names for each real world location. Interpretation tools need to have more sophisticated place name data inputs to search against for supporting disambiguation, the ability to leverage multiple gazetteers for name matching, etc.

A specific challenge relates to the existence and uses of multiple languages for both place names themselves (the endonym/exonym challenge) and the prepositions. Added to this, there are official and unofficial formal/vernacular names for the same real world location. Research needs to be driven by clearly articulated uses cases for processing. As Paul Box indicated in the group’s discussion, much of the research appears to be driven by the classical and fairly limited use case for gazetteers – as

map indexes supporting a human machine interaction, i.e. relating a human label for a place to a location on a digital map. But this perspective represents a small proportion of the potential value of a gazetteer – the other 70% being as a means to reliably geocode and integrates information – this is the driver for the UNSDI project.

A possible use case offered by CSIRO (and potential target for collaborative project) is the Emergency Services Alert tool – a live twitter feed analyzing anomalous frequencies of key search terms to identify disaster events – mapping events to locations, has yet to be addressed.

The group also identified the need to deliver information to users in multiple representations or at least make users aware that they exist. For example when searching for Mount X – a user might be offered a range of representations, the polygon extent of the mountain, or the mountain range or location of the peak. And also, information needs to be delivered in formats that are useful to the human and machine users – spatial formats such as GML but increasingly as linked Data, RDF Triples, etc.

The third group discussed the topic of **volunteered (crowd-sourced) and vs. authoritative data**; or how, when and why VGI becomes usable for ‘authoritative’ agencies. The initial statement in the group was that in the end users (also) look for ease of use, the possibility of aligning different data sets, and having single (or simple) APIs and processes around the data. In the end, the discussion was culminated in the following issues:

- Trust: What are ways to handle / assess / understand completeness of the data? What are quality attributes of such data? There is risk involved with using volunteered data, but if someone else is using the data already, or some other known data set is referencing the data, that would make it more trustable.
- Processes: What are registration processes of the data and where does the data come from (within and across data sets)?

Some additional questions that the group identified were whether a dedicated (volunteered) collection process would be better / more trustworthy than a mining / exploitation of existing data sets, and how could one account for the demographics of different communities that may hold knowledge about needed data.