Google Dataset Search workshop 2018-09-19

Location: INSPIRE Conference 2018, Antwerp

Moderators: Marcin Grudzien (Head Office of Geodesy and Cartography, Poland), Michael Lutz (Joint Research Centre)

Panellists: Simon Ilyushchenko (Google), Paul van Genuchten (Geocat), Gianfranco Cecconi (Capgemini, European Data Portal), Clemens Portele (interactive instruments), Martin Tuchyna (Ministry of Environment, Slovakia), Daniele Rizzi (DG Connect, European Data Portal)

Agenda

1. Introduction to the workshop (Marcin, presentation)
2. Google Dataset Search introduction (Simon, presentation)
3. Panel discussion (chair: Michael)
    a. Google Dataset Search as a game changer / paradigm shift
    b. indexing of individual objects / direct data access
    c. testbed organisation
4. Wrap-up & next steps (Marcin & Michael)

Google Dataset Search as a game changer / paradigm shift

- Metadata should be created as closely as possible to the actual data providers, but should also be made available in as many places as possible (e.g. in regional, national, European data portals and/or search engines). Duplicate records should be filtered out by the harvesting portals/search engines.
- Most geo or open data portals have problems with a good ranking of a larger number of search results. The web search engines are very good at ranking such results. However, dedicated thematic data portals will still have a role also in the future (when/if dataset search through search engine is widespread), but they should focus on value-added aspects for the community (e.g. curated data content, forums, engagement, integration in platforms providing processing capacity, …) rather than search/discovery.
- We need to see portals as solutions to different problems in different timeframes. Today, portals help addressing issues like data literacy, creating easily accessible “human readable” resources for anybody to find data. In the future, as the ecosystem matures, discoverability of data will be less of a problem. Then, portals will still be suitable venues for some content that does not naturally stay with the publisher (see point above).
- Google is a private company and can always change its business model related to their services. Therefore, INSPIRE community should build parallel solutions.
- It is not only about Google - we should keep in mind that there are other search engines or portals that we should support. Any metadata annotations should therefore be based on open standards.

- [ ] In schema.org a dataset distribution is only a downloadable file while in DCAT a dataset distribution may also be an APIs, a feed, etc. Clemens to raise the issue directly with the schema.org community.

- Collaborative approach – INSPIRE stakeholders should develop tools they find useful and at the same time support other developers even if they build similar applications.

Indexing of individual objects and data access

- Most people (especially non-GI experts) are interested to get information about one specific object (e.g. a specific building or lake).
- It is possible to annotate websites which are representation of a specific object using schema.org (e.g. https://schema.org/Place). Such annotations make this information indexable (and they are indexed), but search engines today do not use them in a visible way, e.g., for rich snippets (only curated content like from Wikipedia or the CIA World Factbook for places). In general, like there is a Dataset search for schema: Dataset annotations, there could also be a Place search for schema:Place annotations.
- We need more experience to see individual objects visible in search engines.
- Open point for discussion: Should search engines also support direct access to data sets, or only through the data portals from which the metadata has been harvested? Are users actually interested in downloading the data (or rather just having answers to questions)?

Testbed organisation

- A testbed will be organised with all interested data or solution providers, aiming initially at experimenting with DCAT/schema.org annotations, observing results in search engines and sharing experiences and good practices with the other participants.

- [ ] Michael and Marcin to set up a communication space for the testbed, including information provided by Simon what data providers should do to annotate their websites describing data sets and issue a call for participation.

- [ ] JRC to share consolidated issues, questions or suggestions for improvement with the Dataset Search team at Google.

Other relevant information/comments

- The development of new INSPIRE geoportal goes in the right direction, focusing on pre-defined views that allow to browse through the metadata and not a general search function. Full-text and/or faceted search may still be useful for expert users (in particular data providers who want to improve their implementations), but this will be implemented in a dedicated area (“INSPIRE analytics”) and/or through an API.
- It is not always easy to integrate data in the Google knowledge graph, e.g. because it is difficult to judge which data Google should trust.
- Interactive Instruments has started work on using sitemaps for pages with schema.org Datasets and Places.
- Relevant work was already done under the Geonovum testbed “Spatial Data on the Web” (topic 4) – this should be important background reading before starting experiments in the new testbed.