VGI Analysis with Linked Data

General introduction

The Energic COST action aims to demonstrate the potential of Volunteered Geographic Information such that data, generated by a wide range of participants ranging from authoritative bodies across scientists to individual citizens, can be used to provide information relevant to scientific, societal and policy in a European context. The objective of activities within the Energic Datathon is to allow anyone to participate in our activities, and demonstrate the potential of transformations of data to knowledge. To lower the barrier to entry for participants each Datathon task provides:

- a description of the underlying motivation for the task
- sets out an initial set of questions that might be explored through the data
- gives access to some prepared data and suggests potential additional sources, and
- suggests potential tools and methods which might be used in the task.

However, these guidelines are only intended to give a starting point to the activity, and we encourage you to be as creative as possible. Entries to the datathon will be judged by a panel of Energic members, and the best will be invited to present their results at the Energic closing meeting in London.

Specific introduction including overarching objectives

A smart citizen needs to be able to make sense of bewildering array of sensory inputs, which constantly change (i.e. some data sources dwindle or disappear, while others emerge). Large and rich ontologies can be overkill for semantically poor data, which produces semantic richness only through large numbers. Hence, this datathon is looking for ways to quickly discover all available datasets, retrieve them and integrate them in a way that results in usable data.
This datathon aims to produce a rich information picture of a smaller area in a medium to large city. The emphasis is on speed, and not on completeness or accuracy. The objective is to use as many data sources as possible, i.e. geosocial media, VGI, authoritative and other third party sources, and test lightweight methods to integrate them. Here, anything goes as long as at the end of the day, the various data sources are either integrated or linked in such a manner that a comprehensive analysis (e.g. smart urban planning or social sciences) is possible. The approach (workflow) and the results should be documented exhaustively (with the aim of enabling reproducibility).
The combination of different VGI sources and authoritative information can be supported by Linked Data to disclose semantic relationships between concepts used across multiple domains. This can ease the discovery of relevant data sources and the automation of mash-ups.

Available datasets and additional potential data sources

OSM RDF dataset on http://linkedgeodata.org/
http://downloads.linkedgeodata.org/releases/2015-11-02/
Data made available from the other challenges (‘Tourist mobility in smart cities’, ‘Transport’, ‘Telecommunication’)

Authoritative data

Linked Data sources at https://datahub.io/
Also refer to the data made available from the other challenges ('Tourist mobility in smart cities', 'Transport', 'Telecommunication')

Questions to be asked of the data

What role can Linked Data play in analyzing VGI?

1. What are relevant linked data sources (triple stores in Linked Open Data Cloud) for the challenges below?

a. Single domain:
i. Tourist mobility in smart cities
ii. Transport
iii. Telecommunication

b. Multiple domains:
i. Complete digital representation of a geographical area (richest information possible)
2. What is the added value of Linked Data compared to the original VGI?
3. What are potential (web) application that can make use of queries on the Linked Data (SPARQL queries)?

Possible methods and tools

1. Triplify VGI and authoritative data available from the other challenges ('Tourist mobility in smart cities', 'Transport', 'Telecommunication') or in case of multi-domain, as many data sources as possible. Tool: Triple store creation/editing tool, e.g., Parliament, Virtuoso, AllegroGraph.
2. For the chosen challenge, create (Geo)SPARQL queries across the abovementioned Linked Data. Tool: Linked Data / Ontology explorer (e.g., SPEX) and/or SPARQL query tool
3. Extra challenge: Visualise the queries of 3. in a web application. Tool: e.g., JavaScript

Linked Data / Ontology explorer (e.g., SPEX) and/or SPARQL query tool PLUS triple store creation/editing tool, e.g., Parliament, Virtuoso, AllegroGraph.

More details on data sources, methods, tools, and tips & tricks can be found here:
https://docs.google.com/document/d/1C3B5d7SXjXDQuS6SvBbMCZv6lyZPKE7ywCHSiYbYJcQ/edit?usp=sharing

Reporting your results

The expected outputs include (Geo)SPARQL queries and results (within SPARQL editor, such as Parliament or SPEX) (for the extra challenge: (web)application that uses (Geo)SPARQL queries), and an evaluation of the available integration and/or linking methods.
You should prepare a report of your results which explains briefly:

- The data and methods you used (and provides links to these such that your work can be reproduced)

- Interprets your results, concentrating on what you learnt through the datathon and linking to the questions set out above
- Emphasises challenges in carrying out the datathon
- Illustrates the originality and novelty of your approach
- References any external sources you used to help you complete the task
- A 2 minute video pitch presenting your report

Your report should be prepared as a self-contained set of HTML pages which can be accessed by the judges and uploaded to the Energic website after the challenge. All content on the website should be licensed CC-BY-SA (where you use data sources covered by other licenses you should provide tools and access to these and make clear any limitations in their use).

Judging criteria

A panel of Energic members will judge the quality of entries to the Datathon and select the best examples for presentation at the final Energic meeting in London. The following criteria will be used in judging entries:

- Overall quality of the entry to the datathon
- Originality and novelty of the approach taken
- Quality of the description of the data and tools used, especially with respect to reproducibility
- Soundness of the approach taken
- Potential scientific, societal and policy impacts of the results
- Quality and engagement in the video pitch

Information for organisers

Target group: Linked Data professionals and advanced BSc, MSc and PhD students.
The Energic Datathon is open to anyone. However, it will be most fun, and probably also most productive for small groups (typically 3-4 people). The tasks have been designed such that they can be carried out by groups with different levels of skills, ranging from basic spatial analysis using standard GIS to creation of more complex workflows using programming skills. We estimate that typical time investment for a Datathon task should be of the order of 12 hours - however, it is of course up to participants how much or how little time you invest. The only hard rule is our deadline for submissions of **31.07.2016**.
There is no need to register for the Datathon, just submit your report to ross.purves@geo.uzh.ch by the deadline. However, we'd like to know that you're taking part, so feel free to drop us a mail telling us who you are, how many of you are participating in which challenges, and whether or not others are welcome to join you. Please Tweet about the event using the HashTags #Energic and #Datathon.
Some useful information about running datathons events can be found at:

- https://hackathon.guide/
- http://guide.mlh.io/

Contact information

For further information please send your queries to the following email address:
r.l.g.lemmens@utwente.nl
f.o.ostermann@utwente.nl

For the Energic Datathon challenge:

ross.purves@geo.uzh.ch
f.o.ostermann@utwente.nl
r.l.g.lemmens@utwente.nl