Transport and Mobility Datathon

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

Data allowing comparison of public transport across European cities are hard to come by and typically follow local formats and standards. In this task we propose using volunteered geographic information (VGI) to explore the availability of public transport, and its relation to other data in a range of European cities.

In order to do that, use the data sets and data sources provided to assess the public transport availability of (at least) 10 European cities pre-selected by the organizers. Combine volunteered and authoritative data sources as much as possible.

As a starting point, refer to the 2015 regional working paper titled „Measuring access to public transport in European cities".

http://ec.europa.eu/regional_policy/sources/docgener/work/2015_01_publ_tran…

Perform the analysis for (at least) the following cities:

- Budapest, Hungary
- Stockholm, Sweden
- Berlin, Germany
- Enschede, Netherlands
- Barcelona, Spain
- Florence, Italy

- Zurich, Switzerland
- Sofia, Bulgaria
- Novi Sad, Serbia
- Skopje, FYRM

Available datasets and additional potential data sources

Use the following *obligatory* data sources:

- OpenStreetMap (OSM) – public transport stops and networks (if available), as well as city boundaries from OSM
- European Environment Agency - Population Density for EU27 V5

http://www.eea.europa.eu/data-and-maps/data/population-density-disaggregate…

- DBPedia – population and other general information about cities – if the data within the above EU27 report is insufficient or does not exist
- Flickr – social media data points created within the city boundary

Other *optional* data sources:

- Google' General Transit Feed Specification (GTFS) data from public transport agencies
- Additional available social media data sources, e.g. Foursquare, Instagram, etc.

Potential questions to be asked of the data

Assess the level of access to public transport in the above listed urban areas by analyzing the data sources specified above.

There are two overarching objectives:

- do the same public transport access analysis as in the above report, but based on VGI (only)
- create a similar report, but instead population density, use social media density, i.e. the density of social media data points

In order to meet the overarching objectives, do some or all of the below listed detailed objectives:

- determine the size of the urban area
- extract population density and its distribution in the urban area
- find the number of bus/tram/subway stops and lines (if available)
- find the number of social media data points and their density within the urban area
- calculate the ratio of population density to the number of public transport stops/lines
- calculate the ratio of social media point density to the number of stops/lines

- categorize the public transport access 'levels' using the same measures as in the above mentioned report (i.e. No, Low, Medium, High, Very high)
- categorize the public transport access level to the points of interest (POI) identified in the social media sources
- compare all the results across the range of the listed cities
- propose a method of improving the datasets where there is place for improvement – explain which additional data sources would be used and which method would be applied (e.g. manual annotation, building a custom application for replacing/assisting manual labor). Suggested improvement examples: add missing stops, apply structured naming to stops, build a network of stops, etc.

Possible methods and tools

Use the following methods and tools:

- Apply the methodology used in the above mentioned report as much as possible during your work
- If there is no detailed population density information, then use an average population density, i.e. divide population size by the urban area size
- OSM and Flickr data will be made available by the organizers
- Additional OSM data might be queried via the Overpass API, e.g. public transport network information
- Basic spatial analysis (density, point-in-polygon)
- Linked data querying (geosparql, ...)

Reporting your results

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

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 at 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.

Contact information

For this particular datathon:

Imre Lendák lendak@uns.ac.rs

Karoly Farkas farkask@hit.bme.hu

For the Energic Datathon challenge:

ross.purves@geo.uzh.ch

f.o.ostermann@utwente.nl

r.l.g.lemmens@utwente.nl