Open Government Data Knowledge Graph: ESGREEN Case

Supervisor: Carlos Utrilla Guerrero (data scientist) , Institute of Data Science (IDS)

Description of the topic:

Many governments have mandated the publication of Open Government Data (OGD) to the public via the Web that enables the tracking of a government. The intention is to facilitate the maintenance of open societies and support governmental accountability and transparency initiatives. The current OGD sources however has number of drawbacks and bottlenecks, which have been subject of discussion lately: among others, the rising number of published OGD makes it nearly impossible to get links and connections to other datasets, does not conform to standards and requirements, and in some cases even shortage of usefulness or reusability. Several initiatives have proposed Knowledge Graphs (KG) [1] and FAIR [2] approaches as a set of data-driven solutions to many of these current data challenges. The focus of these approaches is modeling them in standards and optimized fashion in order to realise the goal of government efficiency, transparency, and reusability of OGD.

In the context of Open City Project, which aims at developing smart analytic tools for policy makers using 40 ODGs and different ontologies, we are currently investigating the usefulness of implementing one unique SIO ontology [3] to enhance interoperability between OGDs. One important domain within the Open City Project is the Green Infrastructure and Biodiversity topic. Using SIO ontology, the IDS (together with colleagues from Complutense University of Madrid) is working on the harmonization and integration of biodiversity-ODG for tree monitoring (ESGREEN KG) that will serve as inputs for better measures on biodiversity indexes [4]. The ultimate goal is to map the diversity of street inventories across EU cities using OGD and create visualization tool (http://senseable.mit.edu/diversitree/ or https://mgds.oeg.fi.upm.es/dashboard.html) that is relevant to capture the complexity of biodiversity change and indispensable to provide basis for evaluating city's ability to support healthy urban forest. With a special focus to biodiversity-OGD from other EU countries, the following research questions would merit undertaking Master's thesis (one or two of the following):

1. What are the benefits and disadvantages of using SIO ontology for the creation of the KG [5]? To what extent does the SIO ontology correspond with the data needs of building biodiversity indicators?
2. How adding semantics to link datasets can help infer and answer complex questions e.g. compute Shannon Index?
3. Which data driven methods and tools are most suitable for mapping external datasets and conversion to RDF?
4. Which tools and technologies are the most suitable to create a dashboard based on RDF data?

Prerequisites: Working knowledge of Python and machine learning is desirable.

Suggested readings:

[1] Paola Espinosa-Arias et al. (2020). The Zaragoza's Knowledge Graph: Open Data to Harness the City Knowledge, *MDPI Information 11(3), 129,* https://doi.org/10.3390/info11030129
[2] Wilkinson, M.,Dumontier, M., et al. The FAIR Guiding Principles for scientific and stewardship *Sci Data 3*, *160018*(2016), https://www.nature.com/articles/sdata201618 .
[3] Michel Dumontier et al. (2014). The Sematicscience Integrated Ontology (SIO) for biomedical research and knowledge discovery, *Journal of Biomedical Semantics* https://jbiomedsem.biomedcentral.com/articles/10.1186/2041-1480-5-14
[4] Nadina J. Galle et al. (2021). Mapping the diversity of street tree inventories across eight cities internationally using open data, Urban Forestry & Urban Greening, 61, https://doi.org/10.1016/j.ufug.2021.127099
[5] [2] P.J Stephenson, Carrie Stengel. (2020). An inventory of biodiversity data sources for conversation monitoring. *PLOS ONE*, *44*(1), https://doi.org/10.1371/journal.pone.0242923