AGILE: GIScience Series, 2, 43, 2021. https://doi.org/10.5194/agile-giss-2-43-2021
Proceedings of the 24th AGILE Conference on Geographic Information Science, 2021.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras.
This contribution underwent peer review based on a full paper submission.

Spatial Dataset Search: Building a dedicated Knowledge Graph

Mehdi Zrhal[a], Bénédicte Bucher[a], Marie-Dominique Van Damme[a] and Fayçal Hamdi[b]

[a]University Gustave Eiffel, LaSTIG, IGN, ENSG, F-94 160 Saint Mande, France
[b]Conservatoire National des Arts et Métiers, CEDRIC, 292 rue saint martin, Paris, France

Correspondence: Mehdi Zrhal (mehdi.zrhal@ign.fr)

Abstract. A growing number of spatial datasets are published every year. These can usually be found in dedicated web portals with different structures and specificities. However, finding the dataset that fits user needs is a real challenge as prior knowledge of these portals is needed to retrieve it efficiently. In this article, we present the problem of spatial dataset search and how the use of a geographic Knowledge Graph could improve it. A proposed direction for future work, extending these contributions, is then presented.

1 Introduction and Motivations

1.1 Introduction

With more and more government agencies and private entities adopting open data policies, the number of available datasets on the Web has grown exponentially, including spatial datasets. Retrieving the appropriate spatial dataset for an application is becoming an important issue, for example, appropriate topographic data to register GPS tracks, or the relevant training and running spatial data for a machine learning algorithm.

This search of spatial datasets starts with the discovery of appropriate catalogues. Indeed, the variety of technologies, funding programs, communities, and authorities has led to a variety of portals with too few links between them. Once datasets of potential interest have been identified from one or more catalogues, a follow-up step is to compare their benefits and costs for the application. It is so far left to the user to browse the description of each result and compare them manually.

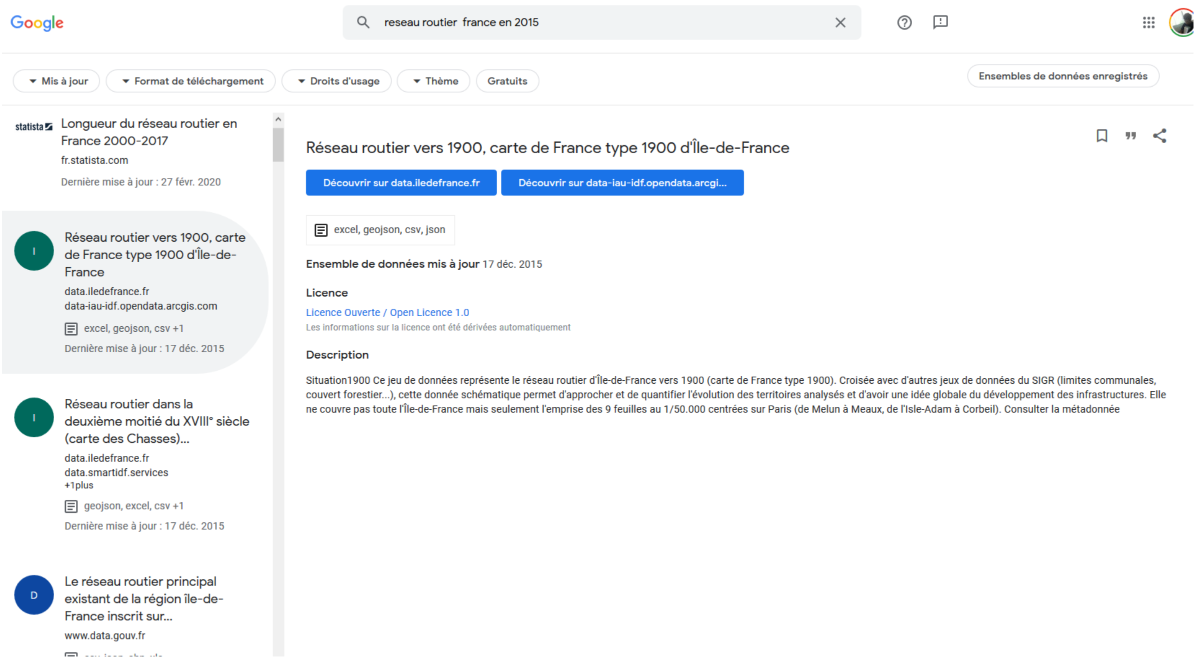
Spatial dataset search has received contributions from different domains. Since more than twenty years, the development of spatial data infrastructures relies on the creation of accurate metadata standards adapted to the complexity of geographical data and on the development of catalogue services. At the international level, the ISO 19115 international standard defines a set of geographical metadata covering: identification, extent, quality, spatial and temporal aspects, distribution, and other properties (ISO, 2014). Contributions also come from the development of the Web of data with more generic and widely adopted metadata standards for open data sets such as DCAT (W3C et al. (2014)) or its profile DCAT-AP[1]. Based on such standards, providers document their datasets, and catalogues process the corresponding metadata to support discovery, evaluation and reuse of datasets, like the European Data Portal[2] (EDP). EDP harvests more than 80 geographic catalogues and over a million different datasets.

A complementary technology relevant to standards for datasets and catalogue services, is that of Knowledge Graph (KG). KGs are empowering our familiar technologies to search for information -search engines, marketplaces, or vocal assistants- and are now adapted to datasets (Noy et al. (2019)). KGs are used to encode domain knowledge that is relevant to interpret and extend a user query, as well as to capitalise on user queries and contexts to improve their capacity to provide relevant answers. The catalogue Google Dataset Search [3] (GDS) indexes a large number of datasets thanks to specific metadata and relies on Google Knowledge graph during the search process. (Brickley et al. (2019)). Bucher et al. (2020) propose

[1] https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/release/201-0

[2] https://data.europa.eu/en

[3] https://datasetsearch.research.google.com/


Google
reseau routier france en 2015
Mis à jour
Format de téléchargement
Droits d'usage
Thème
Gratuits
Ensembles de données enregistrés
statista
I
I
D
Découvrir sur data.iledefrance.fr
Découvrir sur data-iau-idf.opendata.arcgi...

Longueur du réseau routier en France 2000-2017
fr.statista.com
Dernière mise à jour : 27 févr. 2020

Réseau routier vers 1900, carte de France type 1900 d'Île-de-France
data.iledefrance.fr
data-iau-idf.opendata.arcgis.com
excel, geojson, csv +1
Dernière mise à jour : 17 déc. 2015

Réseau routier dans la deuxième moitié du XVIII° siècle (carte des Chasses)...
data.iledefrance.fr
data.smartidf.services
+1plus
geojson, excel, csv +1
Dernière mise à jour : 17 déc. 2015

Le réseau routier principal existant de la région île-de-France inscrit sur...
www.data.gouv.fr

Réseau routier vers 1900, carte de France type 1900 d'Île-de-France

excel, geojson, csv, json

Ensemble de données mis à jour 17 déc. 2015

Licence

Licence Ouverte / Open Licence 1.0
Les informations sur la licence ont été dérivées automatiquement

Description

Situation1900 Ce jeu de données représente le réseau routier d'Île-de-France vers 1900 (carte de France type 1900). Croisée avec d'autres jeux de données du SIGR (limites communales, couvert forestier...), cette donnée schématique permet d'approcher et de quantifier l'évolution des territoires analysés et d'avoir une idée globale du développement des infrastructures. Elle ne couvre pas toute l'Île-de-France mais seulement l'emprise des 9 feuilles au 1/50.000 centrées sur Paris (de Melun à Meaux, de l'Isle-Adam à Corbeil). Consulter la métadonnée

Figure 1. Example of a search in Google Dataset Search

to construct a KG about the geographical digital assets themselves, and not only about the reality, to meet unsolved issues in spatial data infrastructures with insights from semantic web communities.

1.2 Motivating Example

We use a short example to illustrate some issues of spatial dataset search. Figure 1 displays GDS' answer to the query "Road network France in 2015" in French. The first result (i.e., D1) is a dataset indicating the length of the whole road network of France for each year, between 2000 to 2017. The second result (i.e, D2) is a dataset about the road network of the Paris region (i.e, Ile-de-France) around the year 1900. The next result (i.e, D3) is a dataset of the road network of France in the XVIII century. The last result (i.e, D4) is a dataset of the road network of the Paris region in 2012. All four results have been published in 2015. This suggests that GDS does not consider the temporal extent of the datasets.

With the new query "roads France in 2015", D2, D3, and D4 are returned in the same relative order by GDS but are less well ranked overall. This may be related to the fact that none of these records contain the string "routes" in the description field. They all contain the term "road network", which accounts for the results obtained for Q1. In the metadata associated with the three datasets, the keyword "roads" is present in the keywords field of the metadata. This suggests that GDS does not recognise the concepts in the query and gives more importance to the descriptive text associated with a dataset compared to the other structured information that can be found in the metadata. This fits with the statement of Brickley et al. (2019) that GDS mostly relies on the classic Google relevance model adapted to some textual metadata fields, mainly description and publication.

To summarize this introduction, it remains an open issue to discover, compare, and select relevant spatial datasets. Knowledge graphs today are a promising domain to assist human users in accessing resources. Our work is positioned in the field of knowledge graphs for spatial dataset search. We target the design of an open KG visible to all stakeholders, including the ranking mechanism. In the remaining of the paper, we analyse what categories of information and associated operations must be represented in a Knowledge Graph dedicated to spatial dataset search and present the first results of building such a knowledge graph reusing existing resources on the web like metadata and ontologies.

2 Approach and First Results

2.1 Approach

To analyse and prototype our Knowledge Graph, we follow a step-by-step approach. We decompose the process of spatial dataset search using the structure of information retrieval in general as presented by Purves et al. (2007), cf Figure 2, and for each step analyse what categories of knowledge and operations are required, and what existing resources already exist.

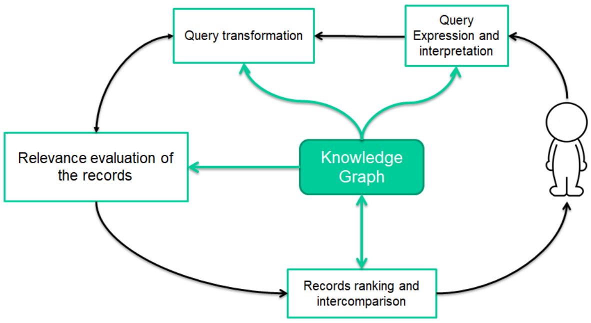

Query transformation
Query Expression and interpretation
Relevance evaluation of the records
Knowledge Graph
Records ranking and intercomparison

Figure 2. Key Steps for Spatial Dataset Search

The generic process of information retrieval is composed of the following steps: the user expresses his query, the engine interprets the query, transforms it into a query that can be confronted to the corpus of metadata and indexes used by the search engine to index resources, i.e. a query for records. The search engine then evaluates a relevance score for each record in the result, and possibly cluster records. It ranks the records, or clusters of records, to present them to the user. It also may extend the query and recommend additional records. The user can then interact with the clusters and the recommended record for further exploration.

To support query expression and interpretation, we consider in our work that the user can express his query with concepts belonging to formal vocabularies which can be integrated into the KG. To identify these formal vocabularies that need to be integrated into the KG, it is important to understand the way in which a user expresses his query. (Kacprzak et al., 2019) studied the behavior of users searching for datasets in the UK based on search logs on four government open data portals. They found out that datasets queries are generally described by using boundaries and restrictions about location, temporality, specific data type and/or specific granularity. They suggest that the most important criteria are temporal and spatial extent with varying granularity. Even though the study was not about spatial datasets, the results are close to the criteria identified by (Sabbata and Reichenbacher, 2012) for spatial data. Indeed, in this study, the authors identified that the primary criteria for geographic relevance are spatio-temporal proximity and topicality. Although the above criteria are not specific to spatial datasets, some conclusions can be drawn. One can assume that in the context of spatial datasets, spatial extent is an even more important criterium. Kacprzak et al. (2019) point out that spatial extent is usually expressed in the form of named places (i.e., names of cities, countries, regions, etc.). Similar results can be observed by looking at the logs of the data.gouv.fr portal (the French portal dedicated to open datasets). Thus, we need to integrate in the KG a gazetteer such as Geonames that is available in RDF format. Last, to support the expression of topicality from a user perspective, there exist universal commonsense ontologies, like DBpedia or Wikidata (Vrandečić and Krötzsch (2014)).

A second step is to transform the user query into an internal query to retrieve records about potentially relevant datasets. For this step, the KG needs to describe the available datasets through records. As mentioned in the introduction, there already exist such models : these are metadata standards, adopted by the GI or by the open data community. The KG also needs to construct a query pattern to express spatial extent criteria, temporal extent criteria, and topic criteria on the different metadata records. Alignments between the user vocabularies to describe a concept of interest, a location and time period of interest, like for example, Wikidata or Geonames, and the metadata vocabularies are also required. In our work, we firstly consider ISO 19115 metadata. In Europe, the INSPIRE Directive (European Parliament (2007)), that targets the creation of an infrastructure for spatial information, requires member states to document their data through metadata compliant with a specific profile of the ISO 19115 metadata standard. DCAT-AP is an application profile that extends DCAT reusing preexisting ontologies to include missing information to be compliant with the ISO 19115.

In ISO 19115 and DCAT-AP, the spatial extent is documented in a dedicated element, called geographic extent. The geographic extent can be a bounding polygon, a geographic bounding box, and a geographic description, like for example "France". Values can be documented in different formats, like for example GeoSPARQL, WKT, GML, or GeoJson for the bounding box.

This standard allows to specify different dates related to a dataset: the creation date, publication date, modification date, and the temporal extent of the dataset. At least one of them is required to comply with the standards, and it is not uncommon that only one of these is provided. This leads to a situation where the temporal extent metadata element is not documented.

Last, topicality refers to the various concepts covered by spatial datasets. In the metadata, this information can essentially be found in two dedicated fields: themes and keywords. The difference between them is that a theme is necessarily associated with a thesaurus or controlled vocabulary, whereas keywords are free text. The INSPIRE metadata profile specifies mandatory metadata, defined as crucial for discovery and selection: the “Topic Category” that should be documented using the code list defined in ISO 19115 and the General Multilingual Environmental Thesaurus (GEMET) and the “keywords” that are free text. The use of another thesaurus is also allowed. It is not uncommon for metadata providers to use an ad hoc specific thesaurus like for example, HydrOntology (Sinha et al. (2014)) for hydrographic data.

A third step is to evaluate the relevance of each retrieved record. This may be done by measuring the similarity between each record and the user query so that the most similar records get the highest relevance score. A minimal relevance model can then be achieved combining measures of similarity for each criterium. Semantic similarity can be achieved using available measures (Elavarasi et al. (2014)). Geographical similarity can be computed using the geometries present in the metadata. Other criteria should be considered, like for example, the licence, or the platform, or the availability of a user forum. Our strategy to investigate relevance model is to focus on a type of application and experiment the feasibility of encoding in a KG an ad hoc relevance model for this type of application through interviewing experts and users. And before interviewing experts, a first prototype is needed to illustrate our objective.

A last step is to prepare the presentation of records to users to facilitate his task of comparing the different results, assessing their similarities and dissimilarities. For example, going back to our motivating example, it could be the mention that all results relate to France and to the road network, that the first result is not a spatial dataset, and that the second and third results have very dissimilar temporal extents. To achieve this, a basic strategy is to evaluate similarity and dissimilarity of metadata records by measuring the similarity of comparable components of the metadata, like for example, comparing elements within records that correspond to the spatial extent.

2.2 Experiment

A small-scale experiment was conducted to test some of our first assumptions and to yield a prototype that is needed to engage with expert users in a second experiment. The selected application is that of environment and water management as there are strong communities already identified in this application domain in France. For this first experiment, we do not integrate the user and its vocabularies. Metadata were retrieved from two French catalogues serving datasets relevant to this application domain: Sandre [4] (188 records) and Cerema [5] (128 records). These were served through CSW interface, usually generated by the Geonetwork software, in the XML implementation of ISO 19115. These XML sometimes are not valid an need some manual editing. Then, these metadata are transformed into DCAT-AP using an XSLT file [6] developed by the Semantic Interoperability Community. Afterwards, the metadata was integrated into a Triplestore Database (TDB) using the Apache Jena [7] framework. GEMET was also integrated to yield a simplified version of the KG containing 397709 triples.

Queries formed by a set of GEMET concepts - "Hydrography" (Q1), "Environmental policy" (Q2), "Hydrography" and "Environmental policy" (Q3) - were submitted to the KG to return the metadata containing the topics in question. One main obstacle during this step is the heterogeneity of metadata. Indeed, even though metadata standards should ensure interoperability, a closer inspection of spatial metadata published on different portals shows that there are still many heterogeneities, which makes it difficult for a search engine to exploit metadata as a graph. Date description is a good example of this situation. Hence, having metadata about the metadata themselves is necessary to improve the processing of metadata during query transformation and during the clustering of records. From our experience, metadata are likely to have the same characteristics -in terms of documentation- when the provider is the same, or also within some portals. 58 results were obtained with Q1, 160 results with Q2, and 57 results with Q3. So far, only an exact match is made between the concepts in the query and those in the metadata. During the intercomparison step, we evaluated the availability of similarity measures that could be applied to metadata elements. Metadata heterogeneity, mentioned above, is an obstacle to this step. These heterogeneities exist in other metadata elements than the spatial and temporal extent, for example, the documentation of coordinate system encoding for which Cerama provides structured metadata whereas Sandre provides textual metadata. Provided, the engine can transform the metadata into an homogeneous metadata graph, it is feasible to implement similarity measures thanks to the capacities of the Silk framework [8]. Various similarity measures and comparison methods are already implemented within Silk such as Levenshtein distance, Jaro distance etc. for character based comparison, centroid distance, overlaps, etc. for spatial comparison. It is also possible to implement customized similarity measures. During our experiment, we applied Levenshtein distance to keywords.

3 Discussion and Future Work

This paper focuses on the design of an open Knowledge Graph dedicated to search for spatial datasets. In a context where data will become increasingly open, along with the European Directive on open data (European Parliament (2019)), the identification of the most relevant datasets, through a transparent ranking mechanism is an important societal stake.

[4] https://www.sandre.eaufrance.fr/

[5] https://www.cdata.cerema.fr/

[6] https://github.com/SEMICeu/iso-19139-to-dcat-ap

[7] https://jena.apache.org/index.html

[8] http://silkframework.org/

We analyse what are the essential components of such a KG by decomposing the spatial dataset search process into different steps. Many of the required components already are present on the Web, like metadata or RDF vocabularies, as well as tools to reuse them like RDF loaders. Metadata about spatial datasets tend to become more standardized even outside the specific field of geographic information and spatial data infrastructures. As Google Dataset Search has been designed to discover only DCAT and Schema.org metadata, the chances are that these vocabularies will be even more predominant in the next few years.

Yet, some categories of knowledge are still lacking. A first category is the links between application domain vocabularies and vocabularies used to document metadata. These are necessary during query transformation and during similarity computation for the simple relevance measure. In an open KG, these links should be shared and adopted by the communities who design the corresponding vocabularies. The second category is the description of how metadata are documented, so to say metadata quality assessment. This information is necessary for the engine to derive a homogeneous metadata graph before measuring the similarities between metadata elements.

Future work will concentrate firstly on the automatic computation of similarity between records to assist the user's comparison of results. This will necessitate a homogenisation of metadata records. Secondly, we will encode in a second experiment an application-oriented relevance model. Our strategy is to interview expert users from our application domain, water management and environment, to acquire important criteria to consider during the whole retrieval process, and to acquire their feedback on different multicriteria similarity measures between metadata records used during the intercomparison step.

References

Brickley, D., Burgess, M., and Noy, N. F.: Google Dataset Search: Building a search engine for datasets in an open Web ecosystem, in: WWW, pp. 1365–1375, ACM, 2019.

Bucher, B., Tiainen, E., Brasch, T. E. v., Janssen, P., Kotzinos, D., Čeh, M., Rijsdijk, M., Folmer, E., Damme, M.-D. V., and Zhral, M.: Conciliating Perspectives from Mapping Agencies and Web of Data on Successful European SDIs: Toward a European Geographic Knowledge Graph, ISPRS International Journal of Geo-Information, 9, 62, 2020.

Elavarasi, S. A., Akilandeswari, J., and Menaga, K.: A survey on semantic similarity measure, International Journal of Research in Advent Technology, 2, 389–398, 2014.

European Parliament: Directive 2007/2/EC establishing an Infrastructure for Spatial Information in the European Community (Inspire) (OJ L 108, 25.4.2007, pp. 1-14), 2007.

European Parliament: Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information. Official Journal of the European Union. 26 June 2019. Retrieved October 2020, 2019.

ISO: ISO 19115:2014, 2014.

Kacprzak, E., Koesten, L., Ibáñez, L. D., Blount, T., Tennison, J., and Simperl, E.: Characterising dataset search - An analysis of search logs and data requests, J. Web Semant., 55, 37–55, https://doi.org/10.1016/j.websem.2018.11.003, https://doi.org/10.1016/j.websem.2018.11.003, 2019.

Noy, N. F., Gao, Y., Jain, A., Narayanan, A., Patterson, A., and Taylor, J.: Industry-scale knowledge graphs: lessons and challenges, Commun. ACM, 62, 36–43, 2019.

Purves, R. S., Clough, P., Jones, C., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Syed, A. K., Vaid, S., and Yang, B.: The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet, 2007.

Sabbata, S. D. and Reichenbacher, T.: Criteria of geographic relevance: an experimental study, International Journal of Geographical Information Science, 26, 1495–1520, https://doi.org/10.1080/13658816.2011.639303, https://doi.org/10.1080/13658816.2011.639303, 2012.

Sinha, G., Mark, D., Kolas, D., Varanka, D., Romero, B., Feng, C.-C., Usery, E., Liebermann, J., and Sorokine, A.: An Ontology Design Pattern for Surface Water Features, pp. 187–203, https://doi.org/10.1007/978-3-319-11593-1_13, 2014.

Vrandečić, D. and Krötzsch, M.: Wikidata: a free collaborative knowledgebase, Communications of the ACM, 57, 78–85, 2014.

W3C et al.: Data catalog vocabulary (DCAT), 2014.