Identifying the key resources and missing elements to build a knowledge graph dedicated to spatial dataset search

Mehdi Zrhal, Bénédicte Bucher, Fayçal Hamdi, Marie-Dominique van Damme

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 207 (2022) 2911–2920

Procedia Computer Science

www.elsevier.com/locate/procedia

26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022)

Identifying the Key Resources and Missing Elements to Build a Knowledge Graph Dedicated to Spatial Dataset Search

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

[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*

Abstract

The number of spatial datasets available online has increased exponentially in recent years. Therefore, the search for spatial datasets is becoming a flourishing research field. The use of knowledge graphs has become rampant in search engines and in information retrieval. In this article, we identify the main resources needed and those missing to allow a knowledge graph to support spatial dataset search. We then apply our approach to the water domain in France by building a dedicated knowledge graph and describe an evaluation method to measure its effectiveness.

Keywords: Knowledge Graph; Spatial Dataset Search; Spatial Metadata; Metadata clustering; Search engine; Metadata standards; Similarity measure.

1. Introduction and Objectives

The number of datasets available online has increased greatly in recent years. This is due, among other reasons, to the key role of datasets for different research and application domains and to the policies of many governments to open their data and make them available to the general public. In Europe alone, the number of datasets published by public or government entities has grown from about 877,000 in August 2019 to more than 1,400,000 in April 2022.

Spatial datasets make up a large part of data sets on-line. This inevitably leads to the problem of efficient discovery of these datasets. Spatial datasets can be found in dedicated portals, called catalogs, which collect metadata describing datasets provided by data producers.

* Mehdi Zrhal. Tel.: +33143988000
E-mail address: mehdi.zrhal@ign.fr

1877-0509
10.1016/j.procs.2022.09.349

In France, for example, there are a large number of catalogs, such as Sandre[1] which specializes in water-related datasets, the Géocatalogue[2] which collects geographic datasets, data.gouv.fr[3] which aggregates open governmental datasets, and Cerema[4] which contains environmental datasets. Therefore, if the user is looking for a dataset concerning the pollution of French rivers, he has to query every catalog. If the user is able to identify the catalogs that meet his or her needs, comparing all the results obtained can be difficult. In fact, in [8] the authors specify that metadata are sometimes not expressive enough to determine if a dataset fits a specific task. There rarely is one dataset that perfectly matches the request, but rather different datasets with different benefits and costs, in terms of required expertise and in terms of uncertainties.

However, from the point of view of data producers, the challenge of data visibility arises. The problem that arises in this case is how to make spatial datasets more visible to users interested in it, but whose needs and behaviors are not clearly defined.

All in all, the issue of discovering and reusing datasets is a retrieval issue; the user cannot specify exactly what he is looking for and needs assistance to select relevant resources among a number of heterogeneous assets.

Recently, the use of Knowledge Graphs (KG) has become rampant in information retrieval [27]. These include companies such as Google, Microsoft, Amazon, and others. Google, Bing, Amazon, and others [23]. In the case of Google and Microsoft, KGs are used to improve document search into an entity search. For its part, Amazon has created a KG that includes all products available in its market place to improve the recommendations made to its users. In [33], the authors present a framework to build a KG dedicated to spatial dataset search.

Our objective in this article is to identify the resources, query patterns, and assets required to create a KG dedicated to spatial dataset search. This article is organized as follows. Section 2 presents related work, then Section 3 specifies the functional requirements of the Knowledge Graph and identifies the components to include in it to support the spatial dataset search. Next, Section 4 describes our implementation of such a Knowledge Graph. Finally, we present our conclusions and future work.

2. Related Work

Spatial dataset search has received contributions from different domains such as information retrieval, metadata and data catalogs, and semantic web. For more than 20 years, the development of spatial data infrastructures has relied on the creation of accurate metadata standards adapted to the complexity of geographical data and to catalog services. This standard aims to facilitate the discovery and reuse of data from different sources in information infrastructures [13, 22] through different properties of the datasets such as identification, extent, quality, spatial and temporal aspects, distribution, and other properties. At the European level, the INSPIRE directive [11], which targets the creation of infrastructure for spatial information, requires member states to document their data through metadata compliant with a specific profile of the ISO 19115 [16] metadata standard.

[8] pointed out that the search for the data set is based mainly on keywords in the available metadata and that the search results are calculated based on filters and experiences that worked for web-based information. This is especially true when it comes to spatial dataset search. In [14]the authors confirm that the search engines used in the catalogs are based on a vertical full-text search associated with filters over some of the metadata fields. [3] introduced a relevance model dedicated to the search for spatial datasets based on three criteria: themes, spatial coverage, and temporal coverage. The relevance of each of the three criteria is computed independently and the global relevance is presented as a visual representation that is not convenient for automatic processing of the overall relevance.

Different portals allow to find spatial datasets. The most popular way to find datasets is through spatial catalogs [19] where data producers publish the metadata of their datasets. The large number of catalogs and the heterogeneities of the metadata make it difficult for the user to find a dataset. Therefore, new portals such as Google Dataset Search[5]

[1] https://www.sandre.eaufrance.fr/
[2] https://www.geocatalogue.fr/
[3] https://www.data.gouv.fr/fr/
[4] https://www.cerema.fr/fr
[5] https://datasetsearch.research.google.com/

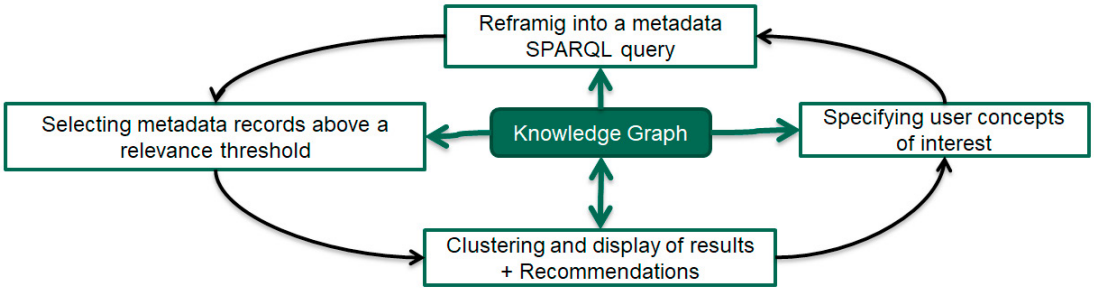

Fig. 1. Functional Requirements of the Knowledge Graph

(GDS) and European Data Portal[6] (EDP) have been developed with the objective of indexing data sets from different catalogs. Both search engines have in common the use of Web metadata standards based on semantic Web technologies.

DCAT [31] is a very popular RDF schema developed and recommended by the World Wide Web Consortium to describe datasets and catalogs. DCAT-AP is an application profile developed by the European Semantic Interoperability Community (SEMIC), which extends DCAT using preexisting vocabularies (locn, prov, etc.) to include the missing information (lineage, provenance, etc.) needed to be compliant with ISO 19115 and the INSPIRE directive. DCAT-AP was used to develop EDP [18], which allows it to index more than a million datasets collected from all over Europe. GDS was designed to discover all kinds of datasets that provide metadata using Schema.org or DCAT vocabularies. In [6], the authors pointed out the critical role of semantics and Google's Knowledge Graph. The relevance model of Google's search engine has been adapted for reuse in GDS, which is due to the lack of a suitable relevance model.

More recently, other methods have been developed to discover datasets based on their similarities. In [20], the authors developed a new method to group metadata records based on their abstract and titles. EDP includes a feature called "similar datasets" that supports query by example (that is, retrieving datasets similar to a specific dataset) based on the TLSH algorithm [25][7]. [2] propose a method to compare datasets based on articles citing them and the citation network between datasets. Finally, [4] described a method for the recommendation of RDF datasets using a concept similarity measure and TF-IDF cosine similarity.

3. Functional requirements of the Knowledge Graph

Before starting the construction of a KG dedicated to spatial dataset searching, it is important to analyze the functional requirements needed to support this task. To do so, we draw inspiration from the generic process of the information retrieval process, especially with regard to search engines as presented in [26]: the user starts by expressing his query, the engine interprets the user query and transforms into a query launched on the corpus of the resources (i.e. documents or datasets). The engine then evaluates a relevance score for each document to present them to the user. It may also extend the query and recommend additional results.

To support the search for spatial datasets, the KG must support these four steps and adjust them to spatial metadata as shown in Figure 1, providing the necessary knowledge, through SPARQL queries during the user session and through KG enrichments out of session. Let us consider the example of a user searching for spatial datasets on rivers.

- Step 1 : The KG should support the identification of user concepts of interest.
- Step 2 : Once the user query is provided, the KG should reframe it into an internal query on the metadata. As we are using RDF metadata, we will use SPARQL as the query language.

[6] https://data.europa.eu/en
[7] https://gitlab.com/european-data-portal/metrics/edp-metrics-dataset-similarities/-/tree/master/src/main/java/io/piveau/metrics/similarities

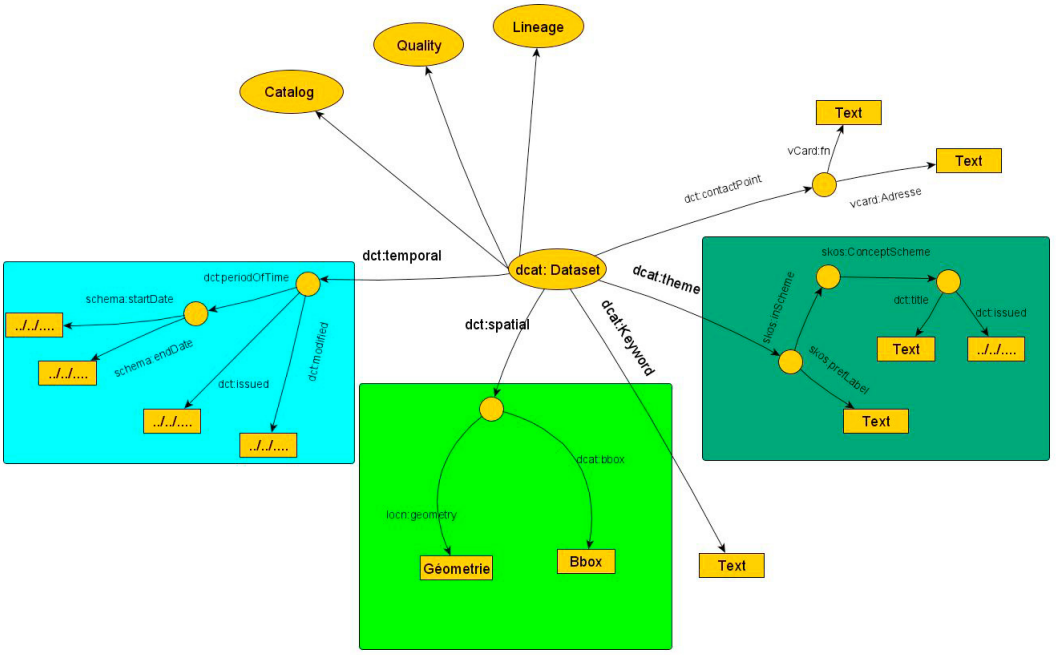

Fig. 2. DCAT-AP Metadata Diagram

- Step 3 : The relevance score must be calculated for each metadata record. Then, we select records above a relevance threshold.
- Step 4 : The results are then displayed to the user. To make it easy for them to find the dataset they need, we propose clustering search results that are similar. The KG can also give user recommendations.

3.1. Identifiying user concepts and creating the corresponding query

To support the identification of user concepts, the KG must contain one or more ontologies of the targeted application domain, but also common sens vocabularies. Using the example of river datasets, the user must be able to express a query containing related concepts (river, hydrography, watershed, etc.) or entities (Seine, Rhône, Loire, etc.). KG then must integrate the metadata that describe the spatial datasets. In the following, they will be referred to as records. The records are in the form of structured data containing various information organized in fields such as title, description, spatial coverage, themes, keywords, provenance, etc. The fields that can be found in a record and the way they are filled depend mainly on the standard used and the producers of spatial datasets[33].

The first hurdle to the exploitation of records through the KG is to identify the relevant fields of metadata to consider for the spatial dataset search. A partial answer to this question can be found in [29, 17], in which the authors identified that the three main criteria for geographic relevance are the topics of the datasets in addition to their spatial and temporal coverage. Focusing on existing spatial metadata, we can see that these criteria are present in fields in metadata standards such as ISO19115 and DCAT-AP. The INSPIRE Directive [11] has made certain metadata fields mandatory for all public and government entities in Europe, including topics and spatiotemporal coverage. As shown in 2, topics can be found in two different fields : ”dcat:theme” and ”dcat:keyword”. The spatial coverage is represented in the field ”dct:spatial” as a bounding box using the property ”dcat:bbox”. Sometimes, the spatial coverage can be expressed as a keyword. There are three properties that capture temporality in metadata: issue date, modification date, and period of time.

The other obstacle to the exploitation of the KG is the high heterogeneity of the metadata records themselves- and the difference in their structure. In fact, one of the fundamental principles of KGs is their ability to contain, generate, and infer knowledge by interlinking different entities [10]. In the current state, metadata is created to optimize the

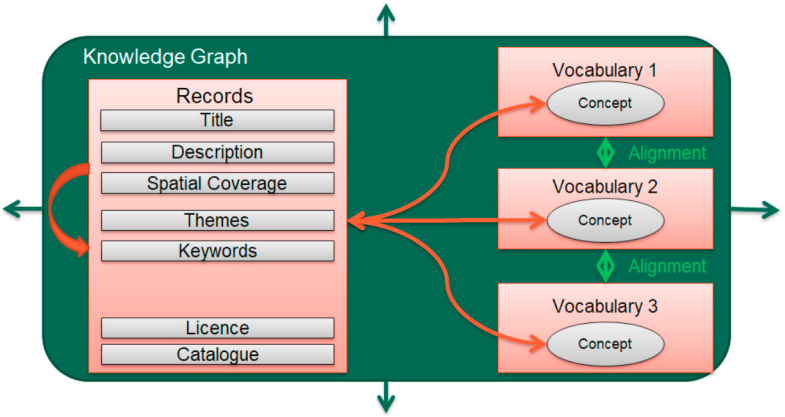

Fig. 3. Knowledge Graph Components

visibility of the datasets they describe on the portals where they are available. The search engines used in these portals are essentially based on vertical full-text search, as presented in [14]. The most used ranking functions are TF-IDF and BM-25 and are based on the frequency of keywords in the user's query in the metadata, hence the importance of textual fields (description, title, keywords, themes, etc.). We have noticed that it is not uncommon for themes or keywords to be included in the description field of records.

The use of a knowledge graph makes it possible to overcome these problems by linking themes and keywords with common sense, thematic, or domain-specific vocabularies. The main difference between themes and keywords is that themes necessarily belong to a controlled vocabulary, whereas keywords are simple strings. The vocabularies used in the records are usually available as linked data, and therefore their themes can be identified through Unique Resource Indentifiers (i.e. URIs). This indicates that, in addition to the records, the KG will have to contain different vocabularies that will allow one to better identify the records and link them with web entities. It is also crucial to link the different vocabularies using alignments to enhance the KG. Figure 3 illustrates the key data to include in the KG.

From this point on, the issue that arises concerns the choice of vocabularies to be included in the KG. To answer that, it is necessary to look deeper into the content of the records and the vocabularies that are mainly used in them. As we focus mainly on INSPIRE-compliant metadata, the GEMET thesaurus (i.e. GEneral multilingual Environmental Thesaurus) should naturally be included in the KG. Indeed, within the INSPIRE directive [11], the use of at least one GEMET theme is mandatory for the metadata to be compliant. GEMET hierarchically categorizes more than 5,500 concepts related to the environment organized into 32 groups and four supergroups. It is available in 27 languages and provides a large number of alignments with other vocabularies and KGs such as DBpedia, Agrovoc, EuroVoc, etc. GEMET contains only concepts, but does not contain any instance of these concepts. A user can find concepts like "river", "hydrography", "city", or "country", but will not find instances such as "Paris", "Germany", or "Garonne". On its own, GEMET does not allow fully accomplishing the first step of our framework; therefore, it is crucial to include another vocabulary to meet this requirement.

To allow the user to fully express his query and achieve Step 1, we must include a vocabulary that contains both conceptual data and real-world instances. A large number of open KGs exist and can be used for our KG, including DBpedia, Wikidata, Yago, etc. We have chosen to use Wikidata in our KG based on the knowledge graph recommendation framework in [12]. In fact, Wikidata is continuously queryable and more reliable than its peers and has better support for non-English labels [28]. Being a central element in LOD, a large number of vocabularies can easily be linked to it.

Completing Step 2 involves creating SPARQL query patterns for the KG to search for a specific record. Having created links between records and Wikidata concepts, it is easier to transform the user query into a query on records. The goal is to retrieve candidate records. If the number of candidate records is null or insufficient, the query will have to be extended to provide the user with records that do not entirely match his query, but that could be of interest to him.

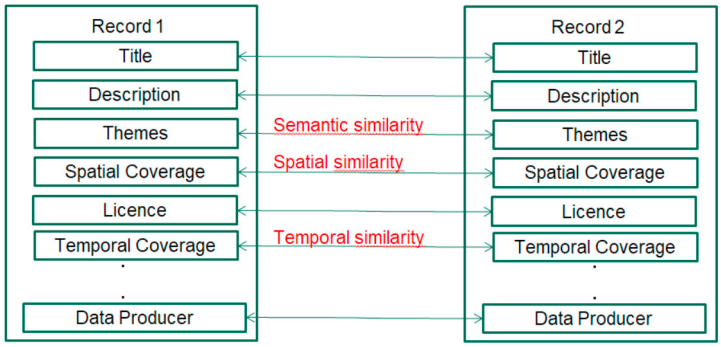

Fig. 4. Comparing Records

3.2. *Identifying record candidates and creating groups of records*

Once the candidates have been found, in order to accomplish Step 3, the relevance of each record to the user's query must be evaluated.

Multicriteria Decision Analysis (MCDA) methods [30] are designed to combine independent similarity criteria to find the instance that best satisfies the chosen criteria. For example, TOPSIS [21] is a method that, based on a set of criteria of interest associated with evaluation or similarity methods, as well as weights for each criterion, can find the best solution that satisfies the criteria best. To use TOPSIS, it is necessary not only to identify the relevant criteria, but also to associate them with a weight that represents their weight and greatly affects the results obtained [24].

There are several similarity measures that can be used to calculate the similarity for each criterion. Thus, for spatial and temporal similarities, there are measures based on the topological relations between two bounding boxes and the overlap between them [3], and equivalently, there exist measures based on the topological relations between time intervals [1].

There are different similarity measures for themes, including semantic similarity measures that compare not the similarity between two strings (i.e., Levenshtein [32], Jaro [9], Hamming [5], etc.) but the proximity of meaning between two concepts based on an ontology, a thesaurus, a vocabulary or a KG (i.e., Wu-Palmer, JC, Tversky) [7]. The results obtained by the latter strongly depend on the vocabulary used. Since we mainly use Wikidata concepts for user query and for identifying the themes and keywords of the records, it is important to implement these measures on Wikidata. The Knowledge Graph Toolkit [15] (KGTK[8]) provides an API to compute semantic similarity between a large number of concepts in Wikidata, which can be easily used by our KG.

To complete Step 4, the KG must be able to create clusters of records. This allows the user to have aggregated search results. In fact, existing portals present search results as an ordered list of records from which the user must choose the most suitable one. When this list contains a limited number of results, the task is easy, but this remains an exception. It is not uncommon for the search engine to find hundreds, if not thousands, of results in the case of GDS and EDP. The presentation of an ordered list of clusters allows the user to find the dataset he is looking for more quickly and easily.

Clustering records implies having a method to compare them. Yet again, it is crucial to identify what criteria (i.e. properties) to take into account when measuring the similarity of two records. The criteria selected for Step 3 and Step 4 do not necessarily have to be the same. In [20] the authors decided to consider only the title and description of the records to group the records. To better leverage the relevant fields of metadata, we believe that the use of a multi-criteria similarity measure dedicated to the comparison of records would allow for a better comparison, and thus better clusters. Thus, Figure 4 shows an example of fields that could be used to create such a multi-criteria similarity measure.

[8] https://github.com/usc-isi-i2/kgtk-similarity

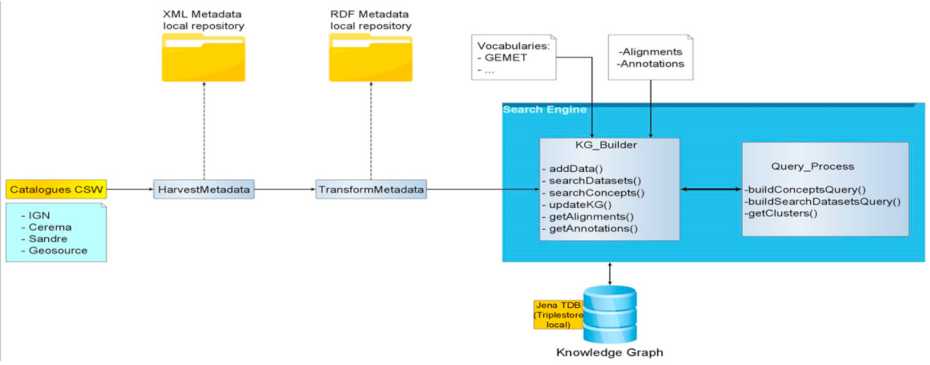

Fig. 5. Production Process of the Knowledge Graph

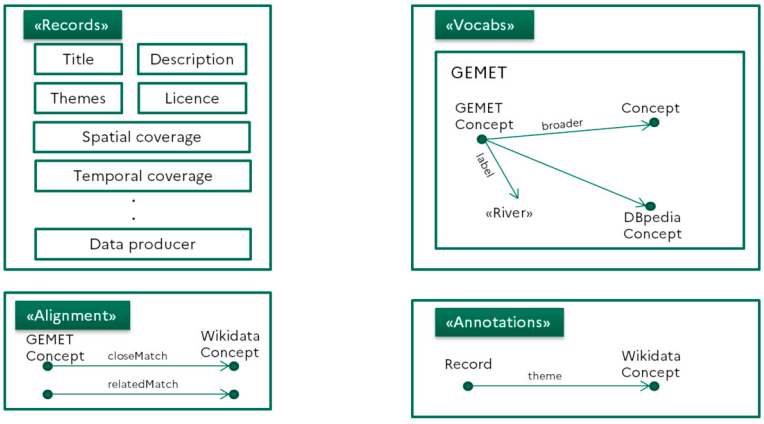

Fig. 6. KG Implementation Components

4. Building the Knowledge Graph

To build a KG and assess our approach, we foster on an application domain, that of water. We produce a first version of the KG and search components and organize its evaluation by a pool of experts who are familiar with the application and with the data. This first evaluation will lead to revising the KG and components before organizing an evaluation by experts who do not necessarily know the domain of metadata.

We identified two French catalogs that contain datasets related to the water domain (Sandre) and the environment domain (Cerema). Figure 5 summarizes the production process of our KG[9]. We also included a few datasets from two other catalogs, IGN (Cartography) and Géosource (Geology).

In total, 209 ISO19115-compliant records files were recovered from the four and transformed into DCAT-AP format using a dedicated file provided by SEMICeu[10]. The RDF conversion step is often mandatory when metadata is not available in RDF. It can also be useful to transform RDF records and reduce heterogeneities, such as the use of equivalent properties to define a metadata field (i.e. dcat:bbox and locn:geometry). However, it is important to note that some errors occurred while transforming a few metadata files specifically concerning three fields (licenseDocument, rightStatment, label) that were corrected in post-processing. The components of our KG implementation can be found in Figure 6.

[9] https://github.com/MehdiZrhal/SpatialDatasetSearch/tree/master/data/sandBox_data/NewKG
[10] https://github.com/SEMICeu/iso-19139-to-dcat-ap

Then, the RDF records were integrated into a Jena TDB triplstore for a total of 51,741 triples. Next, GEMET was added to the KG adding 51,844 more triples. GEMET contains 5,569 different concepts, of which 4,385 are aligned with external vocabularies.

We used the links between GEMET and DBpedia (that is, 3004 concepts) to align the GEMET concepts with Wikidata entities using the owl:sameAs properties that exist between Wikidata and DBpedia entities. Two types of links are available in GEMET to link its concepts with DBpedia, skos:closeMatch and skos:relatedMatch. As the links extracted between Wikidata and DBepdia are owl:sameAs, the links between GEMET and Wikidata will be skos:closeMatch or skos:relatedMatch. For example, starting from (gemet:concept/7244, skos:closeMatch, https://dbpedia.org/ontology/River), which is provided in GEMET and (https://dbpedia.org/ontology/River, owl:sameAs, wd:Q4022), which can be computed through a simple SPARQL query on DBpedia, we can identify the new triple (gemet:concept/7244, skos:closeMatch, wd:Q4022) using transitivity. Thus, 2765 connections between GEMET and Wikidata have been added to the KG.

The last step is to connect the themes found in each record with the corresponding GEMET concepts. To do so, we tried to match the theme labels with the GEMET concepts labels, but noticed that a significant number of themes were not identified. Therefore, we released the constraint on the GEMET labels and manually validated the identified concepts. We then used the alignment of GEMET and Wikidata to directly connect the records to Wikidata concepts. Thus, for the 845 themes present in the records, we created 1,318 links between the records and Wikidata. It is normal to have more links created than existing themes, since a GEMET concept can be linked to several Wikidata concepts.

The KG has been divided into four named graphs. ”Records” gather the triples found in the records, ”Vocabs” for GEMET and if needed other vocabularies, ”Alignments” for all the GEMET-Wikidata links that have been created, and finally ”Annotations” which are reserved for the other links that we have created such as the links between the themes of the records and the Wikidata concepts.

To better enable the user to interact with the KG, we have also developed a user interface available on Github [11]. This allows the user to search for one or more Wikidata concepts using a keyword of his choice. A SPARQL query is sent to Wikidata and a set of candidate concepts is proposed. The user selects the concepts of interest. Then, he can choose a semantic similarity measure among the six available in the KGTK. The relevance is then computed for each record in the KG using the similarity API of KGKT.

To evaluate our approach, we will proceed in two phases. The first will be done in close collaboration with experts who will allow us to improve our KG based on their feedback. The second phase will involve a few less experienced users who will give us their opinion on our approach as a whole.

In phase 1, we will focus on assessing three important points. First, the we will investigate the expressiveness of Wikidata and to what extent it allows the user to formulate a request related to the water domain. Then, we need to evaluate the ability of our approach to identify relevant records and the relevance of precomputed clusters. Indeed, it is necessary to evaluate the ability of the search engine to identify relevant records to the user's query. There are many metrics dedicated to the evaluation of the effectiveness of search engines; the most widely used are precision, recall, and F-measure. Precision can be defined as the proportion of the results that are relevant within the results retrieved by the search engine, recall refers to the proportion of relevant results retrieved by the search engine within all relevant results that are indexed by the search engine, while the F-measure is a combination of precision and recall. To use these metrics, it is essential to have a benchmark of records and queries, as well as the relevant records for each of these queries. To the best of our knowledge, such a benchmark does not exist and must be built. We intend to contact experts to define the datasets to be included in the KG and the set of user queries, as well as the relevant records associated with each query. This will allow us to set a threshold at which a dataset is relevant and should be returned to the user as a result, and to compare the impact on the search results when the similarity measures change.

Next, to evaluate the clusters of records, we first need to identify the criteria for the comparisons of records. Then, we will ask the experts to create reference clusters from the records included in the KG. This will allow us to compare the clusters obtained with our approach to the reference clusters. Furthermore, we will be able to compare the impact of the different similarity measures on the final results.

[11] https://github.com/MehdiZrhal/SpatialDatasetSearch/tree/master/src/main/java/fr/ign/lastig/application

In phase 2, we will be interested in the feedback from non-expert users about the use of the KG as a whole. This concerns the relevance of the identified records and the interest of recommending, through the precomputed clusters, other datasets.

5. Conclusions and Future Work

This article focuses on the construction of a KG dedicated to the search for spatial datasets. Our goal has been to identify the resources, query patterns, and assets that our KG needs to support this task. The elements we have identified are the metadata records in RDF format, the GEMET thesaurus, which is widely used in spatial records, and Wikidata to allow the user to best express his query. Then we discuss the importance of creating links between all these components. We also point out the missing and necessary elements for the KG.

Next, we describe the construction of a first prototype of such a KG and some of the problems encountered during its creation, notably, in transforming records into RDF and in aligning the themes and keywords of records with the vocabularies. Then we discussed an evaluation method for our KG that involves experts and non-expert users which will allow us to enhance our prototype.

In the near future, we will focus on phase 1 of our evaluation method. We have already contacted some experts and started discussions with them. We will make the necessary modifications to our prototype based on their feedback. Finally, based on a questionnaire, we will validate our approach with non-expert users.

References

[1] Allen, J.F., 1983. Maintaining knowledge about temporal intervals. Communications of the ACM 26, 832–843.

[2] Altaf, B., Akujuobi, U., Yu, L., Zhang, X., 2019. Dataset recommendation via variational graph autoencoder, in: 2019 IEEE International Conference on Data Mining (ICDM), IEEE. pp. 11–20.

[3] Beard, K., Sharma, V., 1997. Multidimensional ranking for data in digital spatial libraries. International Journal on Digital Libraries 1, 153–160.

[4] Ben Ellefi, M., Bellahsene, Z., Dietze, S., Todorov, K., 2016. Dataset recommendation for data linking: An intensional approach, in: European Semantic Web Conference, Springer. pp. 36–51.

[5] Bookstein, A., Kulyukin, V.A., Raita, T., 2002. Generalized hamming distance. Information Retrieval 5, 353–375.

[6] Brickley, D., Burgess, M., Noy, N.F., 2019. Google dataset search: Building a search engine for datasets in an open web ecosystem, in: WWW, ACM. pp. 1365–1375.

[7] Chandrasekaran, D., Mago, V., 2021. Evolution of semantic similarity—a survey. ACM Computing Surveys (CSUR) 54, 1–37.

[8] Chapman, A., Simperl, E., Koesten, L., Konstantinidis, G., Ibáñez, L., Kacprzak, E., Groth, P., 2020. Dataset search: a survey. VLDB J. 29, 251–272.

[9] Cohen, W.W., Ravikumar, P., Fienberg, S.E., et al., 2003. A comparison of string distance metrics for name-matching tasks., in: IIWeb, Citeseer. pp. 73–78.

[10] Ehrlinger, L., Wöß, W., 2016. Towards a definition of knowledge graphs. SEMANTiCS (Posters, Demos, SuCCESS) 48, 2.

[11] European Parliament, 2007. Directive 2007/2/ec establishing an infrastructure for spatial information in the european community (inspire) (oj l 108, 25.4.2007, pp. 1-14).

[12] Färber, M., Rettinger, A., 2018. Which knowledge graph is best for me? arXiv preprint arXiv:1809.11099 .

[13] Guptill, S.C., 1999. Metadata and data catalogues. Geographical information systems 2, 677–692.

[14] Hervey, T., Lafia, S., Kuhn, W., 2020. Search facets and ranking in geospatial dataset search, in: 11th International Conference on Geographic Information Science (GIScience 2021)-Part I, Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[15] Ilievski, F., Garijo, D., Chalupsky, H., Divvala, N.T., Yao, Y., Rogers, C., Li, R., Liu, J., Singh, A., Schwabe, D., et al., 2020. Kgtk: a toolkit for large knowledge graph manipulation and analysis, in: International Semantic Web Conference, Springer. pp. 278–293.

[16] ISO, 2014. Iso 19115:2014.

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

[18] Kirstein, F., Dittwald, B., Dutkowski, S., Glikman, Y., Schimmler, S., Hauswirth, M., 2019. Linked data in the european data portal: A comprehensive platform for applying dcat-ap, in: International Conference on Electronic Government, Springer. pp. 192–204.

[19] Kunze, S.R., Auer, S., 2013. Dataset retrieval, in: 2013 IEEE Seventh International Conference on Semantic Computing, Irvine, CA, USA, September 16-18, 2013, IEEE Computer Society. pp. 1–8. URL: https://doi.org/10.1109/ICSC.2013.12, doi:10.1109/ICSC.2013.12.

[20] Lacasta, J., Lopez-Pellicer, F.J., Zarazaga-Soria, J., Béjar, R., Nogueras-Iso, J., 2022. Approaches for the clustering of geographic metadata and the automatic detection of quasi-spatial dataset series. ISPRS International Journal of Geo-Information 11, 87.

[21] Lai, Y.J., Liu, T.Y., Hwang, C.L., 1994. Topsis for modm. European journal of operational research 76, 486–500.

[22] Nebert, D.D., 2004. Developing spatial data infrastructures: the sdi cookbook .

[23] Noy, N.F., Gao, Y., Jain, A., Narayanan, A., Patterson, A., Taylor, J., 2019. Industry-scale knowledge graphs: lessons and challenges. Commun. ACM 62, 36–43.
[24] Odu, G., 2019. Weighting methods for multi-criteria decision making technique. Journal of Applied Sciences and Environmental Management 23, 1449–1457.
[25] Oliver, J., Cheng, C., Chen, Y., 2013. Tlsh–a locality sensitive hash, in: 2013 Fourth Cybercrime and Trustworthy Computing Workshop, IEEE. pp. 7–13.
[26] Purves, R.S., Clough, P., Jones, C., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Syed, A.K., Vaid, S., Yang, B., 2007. The design and implementation of spirit: a spatially aware search engine for information retrieval on the internet. International Journal of Geographical Information Science 21, 717–745.
[27] Reinanda, R., Meij, E., de Rijke, M., et al., 2020. Knowledge graphs: An information retrieval perspective. Now Publishers.
[28] Ringler, D., Paulheim, H., 2017. One knowledge graph to rule them all? analyzing the differences between dbpedia, yago, wikidata & co., in: Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz), Springer. pp. 366–372.
[29] Sabbata, S.D., Reichenbacher, T., 2012. Criteria of geographic relevance: an experimental study. International Journal of Geographical Information Science 26.
[30] Triantaphyllou, E., 2000. Multi-criteria decision making methods, in: Multi-criteria decision making methods: A comparative study. Springer, pp. 5–21.
[31] W3C, et al., 2014. Data catalog vocabulary (dcat) .
[32] Yujian, L., Bo, L., 2007. A normalized levenshtein distance metric. IEEE transactions on pattern analysis and machine intelligence 29, 1091–1095.
[33] Zrhal, M., Bucher, B., Van Damme, M.D., Hamdi, F., 2021. Spatial dataset search: Building a dedicated knowledge graph. AGILE: GIScience Series 2, 1–5.