

# LA RECHERCHE DE JEUX DE DONNÉES GÉOGRAPHIQUES : UNE APPROCHE FONDÉE SUR LES GRAPHE DE CONNAISSANCES

par Mehdi Zrhal, Bénédicte Bucher, Marie-Dominique van Damme  
Université Gustave Eiffel, LASTIG-MEIG, ENSG-IGN

Alors que le nombre de données disponibles en ligne s'accroît avec les politiques d'ouverture de données, le problème de la recherche d'un jeu de données reste ouvert. Il consiste pour un utilisateur à connaître l'existence des données disponibles et à pouvoir les comparer pour faire un choix. Réciproquement, du point de vue des producteurs de données, le défi de la visibilité des données se pose encore. Ce problème est lié essentiellement aux silos existants encore entre les différents catalogues et portails selon des périmètres définis par : la couverture spatiale, temporelle, la technologie sous-jacente, le programme de financement ou encore la communauté. En France par exemple, une recherche de jeu de données disponibles sur la pollution des rivières françaises nécessitera au moins de consulter plusieurs catalogues comme le Sandre, le Géo-catalogue, data.gouv.fr et enfin le portail du Cerema, avec parfois des recoupements. Si l'utilisateur est en mesure d'identifier les catalogues qui répondent à ses besoins, il devra encore comparer les résultats obtenus. Cela s'avère généralement difficile. Il est rare qu'un jeu de données corresponde parfaitement à la demande et l'utilisateur devra évaluer différents avantages et coûts, en termes d'expertise requise et en termes d'incertitudes. Les métadonnées sont insuffisamment expressives pour déterminer si un jeu de données correspond à une tâche spécifique (Chapman *et al.*, 2020). Récemment, l'utilisation des graphes de connaissances (KG) s'est généralisée dans la recherche d'informations par des entreprises telles que Google, Microsoft, Amazon (Noy *et al.*, 2019). L'objectif du travail décrit ici est d'appliquer l'approche des graphes de connaissances au domaine des jeux de données géographiques. La section suivante présente les travaux existant sur ce sujet. Puis nous analysons les exigences fonctionnelles d'un graphe de connaissances dédié à la recherche de jeux de données géographiques et les composants à y inclure.

La recherche d'ensembles de données spatiales a reçu des contributions de différents domaines tels que la recherche d'informations, les métadonnées et

les catalogues de données, et le web sémantique. Des normes de métadonnées adaptées à la complexité des données géographiques sont développées pour faciliter la découverte et la réutilisation de données provenant de différentes sources dans les infrastructures d'information (Nebert, 2004). La directive européenne INSPIRE qui vise la création d'une infrastructure d'informations spatiales, exige des États membres qu'ils documentent leurs données au moyen de métadonnées conformes à un profil spécifique de la norme ISO 19115. Les moteurs de recherche utilisés dans les catalogues sont fondés sur une recherche verticale en texte intégral associée à des filtres sur certains champs de métadonnées (Hervey *et al.*, 2020). De nouveaux portails sont apparus ces 5 dernières années tels que Google Dataset Search (GDS) et European Data Portal (EDP). Ils ont en commun l'utilisation de normes de métadonnées Web basées sur les technologies du Web sémantique. DCAT est un schéma RDF développé et recommandé par le World Wide Web Consortium pour décrire les ensembles de données et les catalogues. DCAT-AP étend DCAT pour l'enrichir d'informations (lignage, provenance, etc.) nécessaires pour être conforme à la norme ISO 19115 et à la directive INSPIRE, et est utilisé pour développer EDP (Kirstein *et al.*, 2019). L'EDP comprend une fonction appelée «similar datasets» qui prend en charge la recherche par l'exemple. GDS utilise le Knowledge Graph de Google et l'applique au texte des métadonnées (Brickley *et al.*, 2019).

Pour élaborer un KG dédié à la recherche de jeux de données spatiales, nous identifions les étapes de ce processus en nous inspirant des étapes de la recherche d'informations (Purves *et al.*, 2007) afin d'identifier et organiser les connaissances nécessaires en réutilisant et alignant des ressources du Web (fig. 1).

Expression de la requête : l'utilisateur exprime une requête. Nous nous focalisons sur l'expression d'une requête *via* des concepts d'intérêt, *e.g.* « fleuve, berges ». Le KG doit contenir une ou plusieurs ontologies du domaine d'application ciblé, mais

aussi des vocabulaires de sens commun et permettre la phase de *disambiguation*. Parmi les KG de sens communs ouverts nous avons choisi Wikidata qui dispose d'un meilleur support, qui conserve le lien avec l'encyclopédie en ligne Wikipedia et supporte extrêmement bien la phase de *disambiguation* et de désignation d'un concept.

Le moteur transforme ensuite cette requête de l'utilisateur en une requête de ressources, ici les enregistrements de métadonnées des jeux. Pour cette phase, il convient d'inclure les vocabulaires des métadonnées, comme le thesaurus GEMET et d'identifier les champs de métadonnées à prendre en compte et de préparer les alignements entre le concept d'intérêt et ces champs. Nous choisissons comme champs le thème, la couverture spatiale et temporelle, le titre, les mots-clés et la description. Les enregistrements sont repris de catalogues français qui contiennent des jeux de données liés au domaine de l'eau (Sandre) et au domaine de l'environnement (Cerema), ainsi que l'IGN et Géosource. Au total, 209 fichiers d'enregistrements conformes à la norme ISO19115 ont été récupérés des quatre catalogues et transformés en format DCAT-AP. Une étape de post-traitement est nécessaire pour harmoniser le KG car les métadonnées sont très hétérogènes. A titre d'exemple, la couverture spatiale peut être représentée dans le champ «dct:spatial» sous la forme d'une boîte de délimitation à l'aide de la propriété «dcat:bbox» ou être exprimée sous forme de mot-clé.

Le moteur évalue ensuite un score de pertinence pour chaque ressource afin de classer les réponses.

Cette pertinence peut être une mesure de similarité entre un enregistrement et le concept d'intérêt. Nous choisissons TOPSIS (Zrhal *et al.*, 2021 et 2022) comme mesure multi-critères pour combiner les résultats de mesures de similarité champ par champ appliquées aux champs « thèmes », « couverture spatiale », « couverture temporelle ».

Enfin, le KG est utilisé pour regrouper les enregistrements candidats et faciliter l'exploration des réponses. Il peut également étendre la requête et recommander des résultats supplémentaires, en se fondant sur une mesure de distance entre les enregistrements.

Pour évaluer notre approche, nous conduisons une première évaluation des fonctionnalités prises séparément par un groupe d'experts qui connaissent bien l'application et les données. Celle-ci met en évidence les limites de Wikidata, dans son état, à supporter l'expression des requêtes. Une autre évaluation a porté sur la capacité du KG à identifier des enregistrements pertinents. Cette évaluation est restée qualitative car la mise en place d'une campagne d'évaluation quantitative de fonctionnalités de recherche de jeux de données s'est heurtée à la difficulté d'avoir des experts pouvant réaliser de façon homogène les jeux témoins. Il existe des différences trop importantes d'un expert à l'autre. Ces premières évaluations ouvrent des perspectives pour la suite : contribuer à améliorer la présence dans Wikidata de concepts applicatifs, mettre en place une campagne d'évaluation avec la communauté, améliorer la présentation visuelle des groupes de jeux.

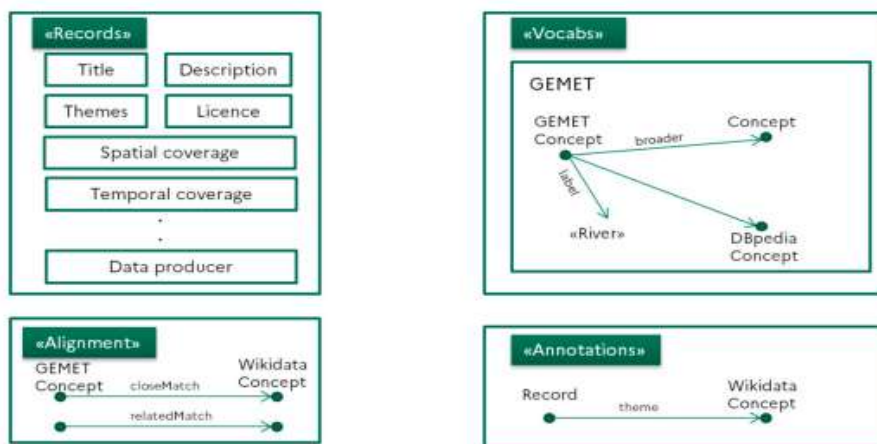


Figure 1 : Composants du Graphe de Connaissances

## Bibliographie

- Chapman, A., Simperl, E., Koesten, L., *et al.*, 2020. Dataset search: a survey. *VLDB J.* 29, 251–272.
- Noy, N.F., Gao, Y., Jain, A., *et al.*, 2019. Industry-scale knowledge graphs: lessons and challenges. *Commun. ACM* 62, 36–43.
- Nebert, D.D., 2004. Developing spatial data infrastructures: the SDI cookbook .
- Hervey, T., Lafia, S., Kuhn, W., 2020. “Search facets and ranking in geospatial dataset search”, *11th International Conference on Geographic Information Science (GIScience 2021)*
- Kirstein, F., Dittwald, B., Dutkowski, S., *et al.*, 2019. “Linked data in the European Data Portal: A comprehensive platform for applying dcat-ap”, *International Conference on Electronic Government, Springer*. pp. 192–204.
- Brickley, D., Burgess, M., Noy, N.F., 2019. “Google dataset search: Building a search engine for datasets in an open web ecosystem”, *WWW, ACM*, pp. 1365–1375.
- Purves, R.S., Clough, P., Jones, C., *et al.*, 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.
- 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.
- Zrhal, M., Bucher, B., Hamdi, F., van Damme, M.-D., 2022, “Identifying the Key Resources and Missing Elements to Build a Knowledge Graph Dedicated to Spatial Dataset Search”, *Procedia Computer Science*, 207, Elsevier, pp.2911-2920.