Sujet de thèse : **Lac de données et référentiels de métadonnées pour élaborer des indicateurs de développement durable de la ville à l'aide de l'open Big data. Application aux pratiques sportives en ville**

Laboratoires : LaSTIG (MEIG[1]), DVRC (Digital Group[2]), CEDRIC (ISID[3])
Institut national de l'information géographique et forestière (**IGN**) [a]
Conservatoire National des arts et Métiers (**CNAM**) [b]
Association Léonard de Vinci (**ALDV**) [c]

Discipline : Informatique
Spécialité : Système d'information géographique d'aide à la décision

Lieux de travail principaux : IGN (Saint Mandé, Marne La Vallée) / ALDV (La défense)
Lieu de travail secondaire : CNAM-Paris
Financement : Thèse co-financée par DVRC et l'IGN (36 mois)
École Doctorale : SMI

Contacts :
Malika Grim[a] Enseignant-Chercheur en informatique, LaSTIG-IGN, malika.grim@ensg.eu, co-encadrante
Faten Atigui [b] Maître de conférences en informatique, CEDRIC-CNAM, Faten.atigui@lecnam.net, co-encadrante
Bénédicte Bucher[a] Directeur de recherche, LaSTIG-IGN, benedicte.bucher@ign.fr, co-directrice
Nicolas Travers [b, c] Professeur en informatique, DVRC-ALDV & CEDRIC, nicolas.travers@devinci.fr, directeur de thèse

Mots clés : NoSQL, Lacs de données, Médiation des données, Métadonnées, Gestion de connaissance, Évènement sportif, Héritage sportif.

1. Contexte général

La disponibilité croissante de données couvrant des aspects variés de notre réalité est une opportunité pour mieux observer et comprendre cette réalité dans sa complexité en exploitant et croisant ces données. Des approches internationales se penchent alors sur la définition d'indicateurs suffisamment comparables dans l'espace et dans le temps, à l'échelle des pays ou des villes, pour évaluer et comparer des situations. Un indicateur des ODD[4] plus précis à considérer pour élaborer et valider notre proposition est l'indicateur ODD11.7 de l'ODD11 ; à savoir l'accès pour tous à des espaces publics sûrs, tels que les espaces verts, les espaces pour les pratiques sportives, etc. En particulier, un domaine qui nous intéresse ici est celui de la pratique sportive dans la ville durable et l'impact de l'organisation de (méga-) évènements sur ces villes et sur les pratiques sportives.

Un évènement sportif est un phénomène spatio-temporel qui affecte structurellement, économiquement et socialement un territoire (le lieu accueillant cet évènement), et générant ainsi un héritage (Harada, 2005 ; Preuss, 2019). L'étude de l'impact des évènements sportifs sur les territoires et sur les pratiques sportives, en comparant des situations par exemple avant et après un (méga-) évènement, ou encore entre deux villes différentes, nécessite l'exploitation de

[1] https://www.umr-lastig.fr/meig/
[2] https://www.devinci.fr/research-center/digital-group/
[3] http://cedric.cnam.fr/lab/equipes/isid/
[4] ODD : Objectifs de développement durable. ODD11 - Faire en sorte que les villes et les établissements humains soient ouverts à tous, sûrs, résilients et durables.

données massives. Il est également indispensable de pouvoir les croiser au-delà des domaines couverts, et de maîtriser suffisamment les biais possibles de comparaison. Cela peut s'avérer particulièrement complexe quand les données sont hétérogènes, de volume, de vélocité et de variété qui peuvent surpasser les capacités des systèmes traditionnels de stockage et de traitement des données. Par exemple, la région Île-de-France possède près de 2,4 millions de licenciés, 19 100 clubs et plus de 101 000 emplois dans le domaine sportif et plus de 7 millions de Franciliennes et Franciliens pratiquent une activité physique et sportive de manière régulière, sans compter les infrastructures et les équipements des pratiques sportives (Gautier et al., 2017).

Diverses solutions informatiques sont avancées dans la littérature pour améliorer le croisement de données hétérogènes et mettre en place des Systèmes d'Information (SI) plus ouverts. En géomatique, des référentiels de référencement direct ou indirect sont spécifiés et produits pour permettre que la caractéristique de localisation de sources diverses soit employée pour les croiser. Des modèles de métadonnées sont enfin proposés pour rendre compte de sources d'incertitudes et de biais.

2. Sujet de thèse et verrou à lever

Ces recherches visent à faciliter l'étude comparée de phénomènes localisés grâce à l'open data et à des solutions avancées d'intermédiation, que ce soit pour étudier un même espace à deux dates (avant et après un évènement) ou pour étudier deux espaces (deux villes différentes). Plus précisément, nous ne visons pas la production automatique d'un diagnostic, mais plutôt d'accroître l'exploitabilité croisée des données ainsi que l'accès aux métadonnées nécessaires à l'adoption d'une perspective critique sur les résultats.

Le sujet de la thèse porte plus précisément sur la structuration de données et de métadonnées en vue de permettre des analyses critiques et comparées relatives à l'impact d'évènements et mégaévènements sur les pratiques sportives en ville. Ce sujet prend tout son intérêt pour les collectivités territoriales, pour des porteurs de projets numériques autour des pratiques sportives et pour les sponsors de grands évènements tels que les JO'2024 et d'autres Grands Évènements Sportifs Internationaux (GESI), tels que Roland Garros, et des Grands Évènements Sportifs Nationaux (GESN), tels que la *Parisienne*.

Le verrou principal est l'absence de cadre unificateur pour mobiliser des données pourvues d'hétérogénéités sémantiques. Celui-ci s'intéresse donc à réconcilier cette hétérogénéité, mais également à faciliter la manipulation et l'analyse de données avec une forte connectivité.

L'approche se positionne dans le domaine de la modélisation sémantique (extraction et transformation de schémas pour des bases de données graphes) et de la qualité en géomatique (description explicite des informations utiles à l'interprétation des données et à la détection de biais possibles).

Deux cas d'étude plus précis seront considérés pour élaborer et valider la proposition :

- La comparaison des parcours sportifs en ville, à vélo ou à pied, avant et après un mégaévènement comme les JO à l'aide d'open data : quels référentiels de données et métadonnées pour permettre le croisement et la comparaison ? Pour ce qui est du

référencement spatial, la thèse étudiera particulièrement les référentiels indirects adoptés par les communautés, c'est-à-dire la description d'une localisation dans une donnée par une référence vers un objet pourvu de coordonnées géographiques (ISO, 2003 ; Hill and Zheng, 1999 ; Chen et al., 2018). Pour ce qui est de la comparaison avant-après, une question concerne la valorisation durable d'un patrimoine, par exemple la promotion de lieux importants de l'histoire du sport cycliste lors de JO, comme l'INSEP ou le Vélodrome Jacques Anquetil, nouvelle dénomination de l'antique Cipale, ou encore le bâtiment où l'union cycliste internationale a été créée le 14 avril 1900. Ceci devrait permettre que ce patrimoine structure davantage de parcours après les JO qu'avant, soit en matière de tronçons parcourus ou de pauses. On étudiera dans cette thèse, a priori, quelles conditions de disponibilité de référentiels et de solutions de croisement de données permettront de conduire ces analyses a posteriori.

- La comparaison des parcours sportifs en ville, à vélo ou à pied, entre deux villes en se fondant sur des données produites par différentes administrations et participants et pourvues de biais différents.

3. Déroulement et environnement de la thèse/ Programme de travail

Le travail de thèse se déroulera selon les étapes suivantes :

- Etape 1 (M1-M6)
  - Effectuer une étude bibliographique détaillée des différents aspects du sujet : modélisation de l'héritage sportif, en se concentrant sur la pratique sportive en ville (des lieux, des aspects spatio-temporels) ; la gestion de problèmes de qualité dans les big open data, etc.
  - Collecter les différentes sources de données et étudier les différents problèmes et lacunes et faire un état de l'art sur les solutions techniques existantes qui puissent être utilisées pour collecter et analyser ces sources.
- Etape 2 (M6-M22) :
  - Proposer une méthode de conception de lacs de données permettant la manipulation et l'analyse des données en se basant sur le cas d'usage. Grâce aux résultats de la phase 1, la caractérisation des besoins utilisateurs servira pour définir un schéma optimal de bases de données graphes facilitant l'accès et le croisement de données. Différentes approches de modélisation de schémas sont adaptées aux bases orientées documents ou colonnes (Mali 2022), mais peu sur des données graphes (Chen 2018 ; Djebali 2020). Les approches sont peu flexibles et ne s'intéressent pas au cadre spécifique des données géomatiques. Cette seconde phase s'intéressera donc à concilier le cas d'usage avec le schéma de données, pour faciliter la conception de ce lac de données tout en reposant sur la sémantique.
  - Publication
- Etape 3 (M14-M30)
  - Développer une méthodologie de conception de BD graphe optimale visant à implémenter l'approche sémantique, guidée par les cas d'usage.
  - Un prototype permettant de valider l'approche proposée ainsi que son évaluation
  - Publication + Démonstration
- Etape 4 (M30-M36) : Les 6 derniers mois seront consacrés à la rédaction de la thèse.

4. Profil attendu

Le candidat doit répondre aux exigences suivantes :

- Possède de solides compétences en informatique, en science des données ou en mathématiques (Master 2 ou équivalent en Informatique ou en Sciences de l'Information Géographique),
- Possède de bonnes connaissances en modélisation des systèmes d'information,
- A un intérêt marqué pour la recherche en science des données et les applications réelles de l'analyse avec un goût pour la pluridisciplinarité orientée vers les sciences de la ville,
- Possède de solides compétences en développement de logiciels pour pouvoir réaliser des idées de recherche en matière de prototypes de logiciel,
- Possède d'excellentes compétences en communication en anglais.

5. Toute candidature doit inclure :

- Un CV,
- Une lettre de motivation adaptée au sujet proposé,
- Les relevés de notes des dernières années d'étude,
- L'avis du directeur de master (ou de la personne responsable du diplôme donnant l'équivalence du master), le cas échéant des lettres de recommandation.

6. Références bibliographiques

Chen, H., Vasardani, M., Winter, and Martin Tomko, M. (2018). A Graph Database Model for Knowledge Extracted from Place Descriptions, International Journal of Geo-Information. ISPRS Int. J. Geo-Inf. 2018, 7, 221; doi:10.3390/ijgi7060221

Djebali, S., Loas, N., T*ravers, N.* (2020). Indicators for Measuring Tourist Mobility. *WISE (1) 2020*: 398-413

Grim-Yefsah, M. and Bucher, B. (2019). Towards Improving Knowledge Capitalization System for Sport Events Legacy. In *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 3: KMIS,* ISBN 978-989-758-382-7, pages 264-270. DOI:10.5220/0008348102640270

Gautier, C., Peuvergne, C., Thévenot, L., Chardon, B., Corne-Viney, N. (2017). Vers un schéma régional de développement des activités physiques et sportives en Ile de France. Phase1 : un diagnostic problématisé. IRDS, département dédié de l'IAU-IdF ISBN : 978-2-11-152475-0

Harada, M. (2005). Hosting a mega-sports event and its impact upon city development challenges of the city of Osaka. Comparing Sports Policy, Sports Investment and Regional Development Initiatives in the Hosting of Sports Events in East Asia and Europe. Université d'Edimbourg, Royaume-Uni, 3 au 11 Mars 2005.

Hill, L., Zheng, Q., 1999. Indirect Geospatial Referencing through Place Names in the Digital Library: Alexandria Digital Library Experience with Developing and Implementing Gazetteers, in: Proceedings of the 62nd Annual Meeting of the American Society for Information Science. Washington, DC,pp.57–69.

ISO, 2003. Geographic information -- Spatial referencing by geographic identifiers (International Standard No. ISO 19112:2003). International Organization for Standardization (TC 211).

Mali, J., Shohreh, A., Atigui, F., Azough, A., Travers, N. (2022). A Global Model-Driven Denormalization Approach for Schema Migration, In International Conference on Research Challenges in Information Science, Mar 2022, Barcelona, Spain

Preuss, H. (2019) Event legacy framework and measurement, International Journal of Sport Policy and Politics, 11:1, 103-118, DOI: 10.1080/19406940.2018.1490336