Vers une plateforme sémantique pour l'intégration des données massives appliquée à la surveillance environnementale

Maroua Masmoudi[1], Mohamed Hedi Karray[1], Sana Ben Abdallah Ben Lamine[2], Hajer Baazaoui Zghal[2], Chirine Ghedira-Guegan[3], Bernard Archimede[1]

[1] LGP-INPT-ENIT,Université de Toulouse, France
{prénom.nom}@enit.fr
[2] RIADI Laboratory, Université de Manouba, Tunisie
{prénom.nom}@riadi.rnu.tn
[3]Université de Lyon 3, France
{prénom.nom}@univ-lyon3.fr

Résumé : Les systèmes d'observation génèrent continuellement une multitude de données environnementales et hétérogènes. Cependant, ces données, provenant de différentes sources, sont généralement caractérisées par leur hétérogénéité syntaxique, structurelle et sémantique. La principale problématique soulevée concerne l'interopérabilité sémantique. Pour faire face à ce problème, le projet PREDICAT vise à fournir une plate-forme sémantique orientée-services pour l'intégration des données massives appliquée à la PREDIction des CATastrophes naturelles. Dans ce papier, nous abordons un des principaux objectifs de la plateforme qui consiste à assurer une interopérabilité sémantique des données par la proposition d'une ontologie que nous appelons pour le domaine de l'environnement. Un cas d'utilisation pour l'application de l'ontologie MEMOn dans l'intégration de données environnementales est aussi présenté.

Mots-clés : Intégration, Big data, ontologie, modularité, interopérabilité sémantique.

1 Introduction

De nombreux systèmes d'observation et de prédiction diffusant des données environnementales (météorologique, climatique, etc.) de façon continue, sont actuellement disponibles. Bien que cette augmentation exponentielle de ces données soit bénéfique pour l'analyse des catastrophes naturelles, elle fait encore face à certaines problématiques, telles que l'hétérogénéité des données, la diversité des sources (satellites et capteurs), la diversité des formats et la diversité sémantique. La principale problématique soulevée concerne l'interopérabilité sémantique de ces données.

Pour ce faire, nous proposons une plateforme sémantique, appelée PREDICAT (PREDIct natural CATastrophes), qui vise à assurer l'interopérabilité des données environnementales et la prédiction des catastrophes naturelles. PREDICAT vise à 1) assurer un accès uniforme à des données hétérogènes en fournissant des services adéquats, 2) intégrer les données environnementales provenant de plusieurs sources, y compris celles fournies par les citoyens, et 3) fournir une solution d'aide à la décision pour analyser en temps réel toutes les données afin de prévenir et réagir efficacement aux catastrophes naturelles.

La suite de cet article est organisée comme suit : la section 2 présente l'architecture globale de la plateforme PREDICAT. Dans la section 3, nous introduisons l'ontologie de domaine

modulaire proposée afin d'assurer un certain niveau d'interopérabilité sémantique entre les données hétérogènes. La section 4 présente un cas d'utilisation d'intégration en utilisant des données réelles fournies par différentes sources. Enfin, nous concluons l'article et proposons les perspectives de ce projet.

2 PREDICAT : Une plateforme pour l'intégration des données environnementales

De nombreux travaux ont été proposés visant à résoudre les problèmes de l'intégration des données massives et hétérogènes. Tous ces travaux visent également à intégrer sémantiquement des données hétérogènes provenant de multiples sources de données telles que les données satellitaires et y compris les données fournies par les citoyens via les médias. Cependant, le stockage et la gestion des données avec les plates-formes traditionnelles (tel que ETL) s'avèrent difficiles. Au fur et à mesure que le nombre et le type de sources de données augmentent, l'accès, le traitement et l'utilisation de ces données doit être facilité pour une meilleure prédiction en temps réel.

Dans ce contexte, nous la plateforme sémantique PREDICAT appliquée à la PREDIction des CATastrophes naturelles (Masmoudi et al, 2018a). La plate-forme PREDICAT a comme principaux objectifs d'assurer (1) un accès uniforme aux données basé sur les technologies des services Web, (2) Une interopérabilité sémantique des données par la proposition d'une ontologie pour le domaine de l'environnement ; (3) Une intégration sémantique des données, (4) et finalement une approche d'aide à la décision pour la prédiction des catastrophes naturelles.

La figure 1 présente l'architecture globale de la plate-forme PREDICAT et de ses différentes couches. L'architecture est composée de huit couches, à savoir :

1) La couche de collecte des données : cette couche englobe différentes sources de données pertinentes pour le domaine de l'environnement (tel que OSS[1], NOAA[2], etc.).
2) La couche de données massives : cette couche présente la diversité des formats dans lesquelles les données environnementales sont stockées (tel que les fichiers CSV, les bases de données, les images RASTER, etc.)
3) La couche de services d'accès aux données : Cette couche traite la mise en œuvre des services en utilisant le style architectural RESTful et en donnant à ces services une description sémantique du contenu des services et des données associées.
4) La couche de traitement des données ; cette couche traite le schéma d'exécution des services relatifs aux concepts demandés par l'utilisateur. Elle propose, en effet, un schéma d'orchestration pour les services d'accès aux données.
5) La couche sémantique : contient les différentes ontologies proposées dans cette plateforme telle que l'ontologie de domaine de l'environnement.
6) La couche d'intégration des données : L'objectif de cette couche est de combiner une grande quantité de données provenant de diverses sources hétérogènes en une seule vue cohérente et globale des données.
7) La couche d'application : La couche d'application consiste en deux composants, le composant d'apprentissage et le composant de prédiction. Le but du premier est d'exécuter des modèles d'apprentissage automatique et des modèles prédictifs qui tirent des leçons des données existantes pour prédire les comportements futurs. Quant au composant de prédiction, il traite les données en temps réel et prend en considération les connaissances déduites du composant précédent pour fournir des alertes précoces.
8) La couche de l'interface utilisateur : cette couche donne la main à l'utilisateur pour saisir des requêtes et visualiser les résultats et les alertes.

[1] http://www.oss-online.org/

[2] https://www.noaa.gov/

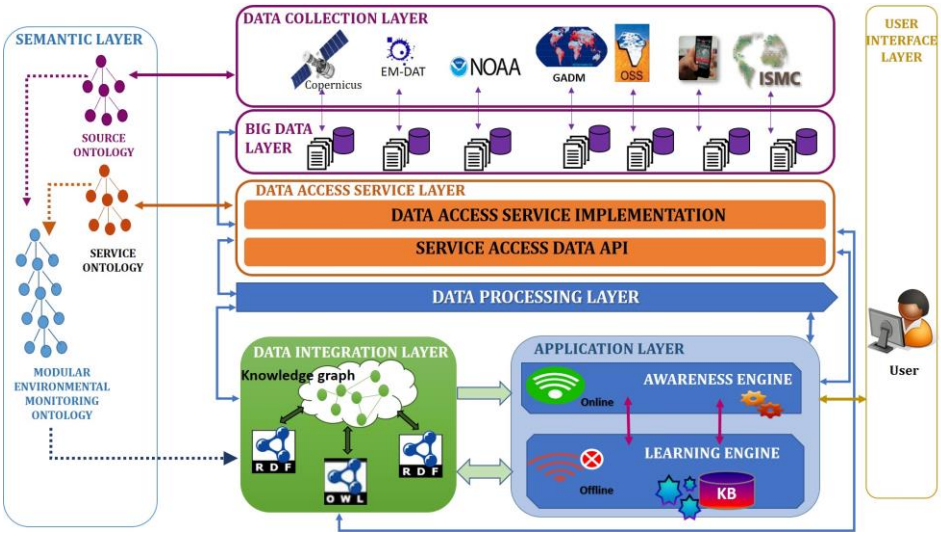

FIGURE 1 – L'architecture globale de la plateforme PREDICAT.

3 Une ontologie modulaire pour le domaine de l'environnement

Afin de garantir une sémantique commune qui délimite le domaine de l'environnement, nous avons proposé une ontologie modulaire basée sur l'ontologie fondamentale BFO, nommée MEMOn (Modular Environmental Monitoring Ontology). Pour ce fait, nous avons réutilisé des ontologies existantes telles que *Common Core Ontologies* (une ontologie de niveau intermédiaire), *SSN* (Haller et al., 2018) *et ENVO* (Buttigieg et al., 2016) (des ontologies de domaine). L'ontologie MEMOn comporte 8 modules, à savoir le module des catastrophes naturelles, le module des matériaux environnementaux, le module des infrastructures, le module temporel, le module géospatial, le module des artefacts, le module des processus environnementaux et le module des régions géographiques.

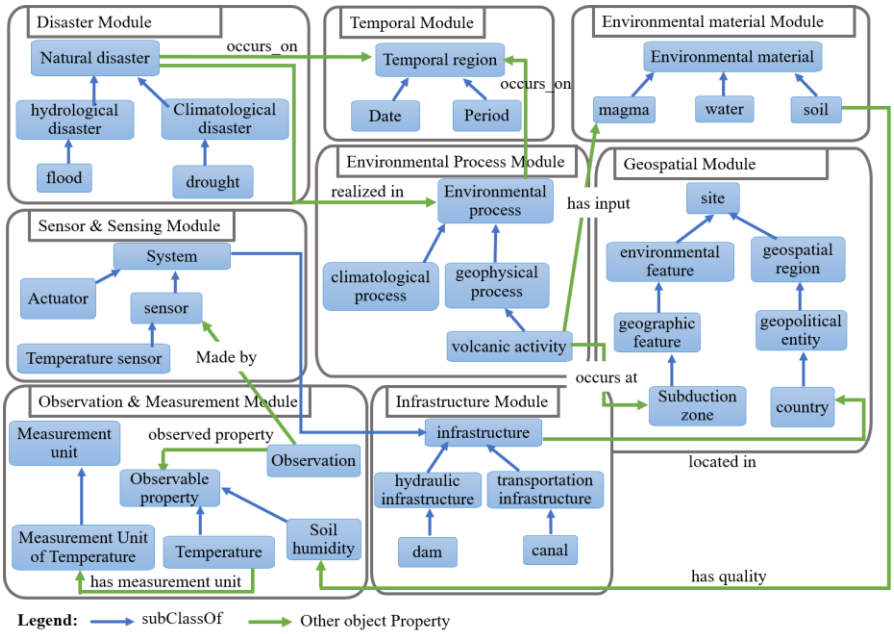

FIGURE 2 – Une vue partielle de l'ontologie MEMOn.

La figure 2 présente une vue partielle des différents modules. Cette ontologie fournira non seulement un vocabulaire commun du domaine, mais elle facilitera également la liaison sémantique des données provenant de différentes sources par le biais d'un graphe de connaissances virtuel (knowledge graph). Plus de détails sur le contenu de l'ontologie est disponible dans (Masmoudi et al., 2018b) et le site Github https://github.com/MEMOntology/memon

4 Un cas d'utilisation pour l'intégration de données hétérogènes

Dans cette section, nous présentons un exemple pour démontrer la capacité de MEMOn à garantir l'interopérabilité sémantique, intégrer et relier les données observées provenant de différentes sources. Ce cas d'utilisation est basé sur un exemple réel et traite des données de précipitations provenant de l'OSS sous forme d'images RASTER, des données sur les tempêtes de la NOAA et des données sur les inondations de EMDAT[3] en format CSV. Parmi les problèmes auxquels nous avons été confrontés lors de l'intégration de ces données multi-source : les schémas de données des trois sources de données sont hétérogènes. De plus, les termes sont présentés différemment dans chaque source. MEMOn est utilisée alors pour résoudre ce problème et assurer l'interopérabilité sémantique entre les schémas de données hétérogènes. Puisque l'ontologie MEMOn est développée pour garantir l'intégration sémantique et la liaison des données environnementales, nous devons générer le fichier RDF (Resource Description Framework) qui décrit la vue globale des données. Ainsi, nous avons utilisé Karma Web system (Gupta et al., 2012), un framework de modélisation et d'intégration de données, qui peut intégrer des données provenant d'une variété de sources incluant des bases de données relationnelles, des tableurs, des fichiers XML, des fichiers JSON et des ontologies pour la modélisation des données. La figure 3 représente une vue partielle du graphe de connaissances construit suite au processus d'intégration. Les données peuvent être vues comme un graphe de connaissances grâce aux relations hiérarchiques et sémantiques de l'ontologie MEMOn qui permettent la liaison sémantique des données. En fait, avec notre approche, nous n'avons pas seulement intégré sémantiquement les données, mais nous avons aussi les corrélations entre elles. Avec ce graphe de connaissances, nous pourrions ainsi transformer l'information en connaissances pratiques et en extraire des connaissances implicites avec l'utilisation d'un raisonneur. Le but de cette vue globale des données (graphe des connaissances) est de récupérer l'information corrélée et de l'exploiter pour en tirer des leçons et prévenir des événements similaires dans le futur.

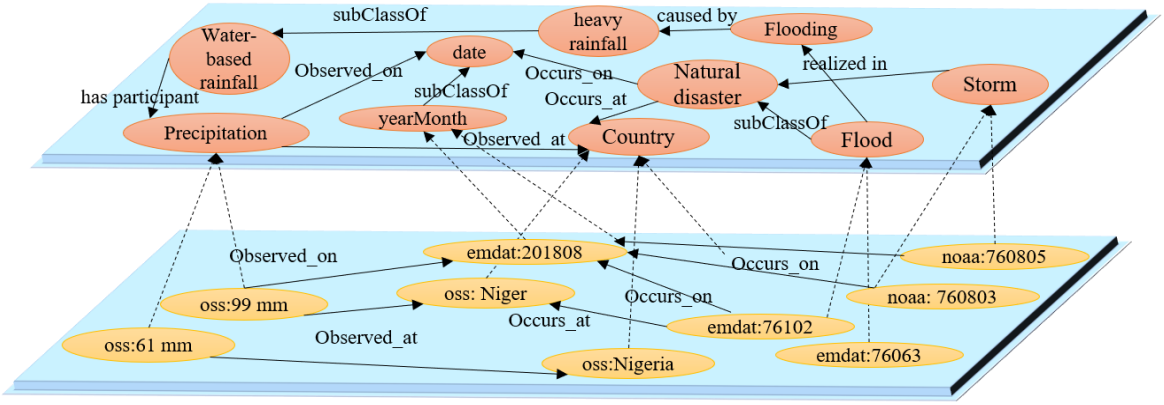

FIGURE 3 – Un cas d'utilisation de l'application de MEMOn dans l'intégration de données environnementales.

[3] https://www.emdat.be/

5 Conclusion

Dans cet article, nous avons proposé une plateforme d'intégration sémantique des données hétérogènes et multi-source appliquée à la prédiction des catastrophes naturelles. La plateforme PREDICAT vise à intégrer et traiter des données hétérogènes à grande échelle provenant de sources multiples, y compris celles fournies par les citoyens, afin de prévenir efficacement des catastrophes naturelles. Les contributions de nos travaux portent sur 1) la construction d'une ontologie modulaire pour garantir l'interopérabilité sémantique entre les données, 2) la proposition d'une approche d'intégration des données qui assure une vision globale des données environnementales et 3) la proposition d'un système d'aide à la décision qui permet la prédiction des catastrophes naturelles. Dans nos futurs travaux, nous prévoyons utiliser l'ontologie MEMOn comme un support pour une nouvelle approche d'intégration de données multi-source. Un suivi passionnant consiste à appliquer MEMOn dans des applications réelles telles que les applications de prédictions de catastrophes naturelles ou de prévisions climatiques.

Remerciement :

Ce travail a été mené dans le cadre du projet PHC-Utique,17G-1122, du Ministère français des affaires étrangères et le Ministère tunisien de l'enseignement supérieur et de la recherche scientifique.

Les auteurs tiennent à remercier les experts de l'Observatoire du Sahara et du Sahel (OSS) pour leur coopération en apportant leur soutien, leur connaissance du domaine et les données environnementales.

Références

BUTTIGIEG, P. L., PAFILIS, E., LEWIS, S., SCHILDHAUER, M., WALLS, R., MUNGALL, C. (2016). The environment ontology in 2016: Bridging domains with increased scope, semantic density, and interoperation, Journal of biomedical semantics, 7-57.

GUPTA, S., SZEKELY, P., KNOBLOCK, C. A., GOEL, A., TAHERIYAN, M., MUSLEA, M. (2012). Karma: A system for mapping structured sources into the Semantic Web. In Extended Semantic Web Conference (pp. 430-434). Springer, Berlin, Heidelberg.

HALLER, A., JANOWICZ, K., COX, S. J., LEFRANÇOIS, M., TAYLOR, K., LE PHUOC, D., LIEBERMAN, J., GARCÍA-CASTRO, R., ATKINSON, R., STADLER, C. (2018). The SOSA/SSN ontology: a joint WeC and OGC standard specifying the semantics of sensors observations actuation and sampling. In Semantic Web (1), 1-19.

MASMOUDI, M., TAKTAK, H., LAMINE, S. B. A. B., BOUKADI, K., KARRAY, M. H., ZGHAL, H. B., ARCHIMEDE, B., MRISSA, M. & GUEGAN, C. G. (2018). PREDICAT: A Semantic Service-Oriented Platform for Data Interoperability and Linking in Earth Observation and Disaster Prediction. In 2018 IEEE 11th Conference on Service-Oriented Computing and Applications (SOCA) pp. 194-201.

MASMOUDI, M., LAMINE, S. B. A. B., ZGHAL, H. B., KARRAY, M. H., & ARCHIMEDE, B. (2018). An ontology-based monitoring system for multi-source environmental observations. In 22nd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems. Procedia Computer Science, 126, p. 1865-1874.