Intégration sémantique de données Raster pour l'observation de la Terre sur des unités territoriales

Ba-Huy Tran, Nathalie Aussenac-Gilles, Cassia Trojahn, Catherine Comparot

▶ **To cite this version:**

Ba-Huy Tran, Nathalie Aussenac-Gilles, Cassia Trojahn, Catherine Comparot. Intégration sémantique de données Raster pour l'observation de la Terre sur des unités territoriales. 33èmes Journées Francophones d'Ingénierie des Connaissances (IC 2022), Jun 2022, Saint-Etienne, France. pp.92-94. ⟨hal-03760543⟩

HAL Id: hal-03760543

https://hal.science/hal-03760543v1

Submitted on 25 Aug 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

Intégration sémantique de données Raster pour l'observation de la Terre sur des unités territoriales

Ba Huy Tran, Nathalie Aussenac-Gilles, Catherine Comparot, Cassia Trojahn

[1] IRIT, Université de Toulouse, CNRS, UT1, UT2, Toulouse, France

prenom.nom ou prenom.nom-composé@irit.fr

Résumé

*En Observation de la Terre, le format raster, standard de représentation de données et images, convient mal pour caractériser des zones d'intérêt par la seule valeur des pixels. Nous proposons d'intégrer sémantiquement les données raster à d'autres données sur la base de leurs propriétés spatio-temporelles. Ce processus s'appuie sur un modèle sémantique de données qualifiant une zone géographique grâce à des unités territoriales et sur un processus sémantique d'*extraction, transformation et chargement *(ETL) associant données agrégées et zones géographiques. Cet article est un résumé des travaux présentés dans [3].*

Mots-clés

Intégration sémantique de données spatio-temporelles ; Observation de la Terre ; détection de changement.

Abstract

In Earth Observation, the raster format, standard for data and image representation, is not well suited to characterize areas of interest by pixel values alone. We propose to semantically integrate raster data with other data according to their spatio-temporal properties. This process is based on a semantic data model qualifying a geographical area thanks to territorial units and on a semantic process of extraction, transformation and loading *(ETL) associating aggregated data and geographical areas. This paper presents a summary of the work in [3].*

Keywords

Semantic data integration ; spatial and temporal data ; Earth observation ; change detection.

1 Problématique

Depuis 2015, les satellites Sentinel-1 et Sentinel-2 du programme Copernicus [1] ont fourni un grand volume d'images de haute qualité de la Terre (environ 8-10 To de données par jour), offrant aux utilisateurs des données et des métadonnées d'Observation de la Terre (OT) gratuites, fiables et actualisées. Associées au développement d'algorithmes d'apprentissage automatique, ces sources de données ont stimulé le traitement des images et son application dans divers domaines. Elles ont ouvert la voie à de nouvelles applications, de l'agriculture à la gestion des forêts, ou la surveillance de catastrophes naturelles.

Un des formats de données les plus courants pour gérer des images satellite est le format raster. Un raster modélise les phénomènes géographiques comme une surface régulière dans laquelle chaque cellule (ou pixel) est associée à un indicateur (par exemple un indicateur de végétation) ou à une valeur de phénomène selon un codage ou une classification prédéfinie (comme le codage d'un niveau de changement). Ces représentations peuvent être construites automatiquement, y compris par apprentissage automatique. Plusieurs rasters sont disponibles pour la même zone géographique, ce qui permet de surveiller soit le même phénomène à différentes dates, soit des phénomènes différents ; ils peuvent être comparés ou combinés pour en générer un nouveau raster [4]. Cependant, dans une perspective de prise de décision, l'interprétation de leur contenu nécessite des données ou des représentations de connaissances de plus haut niveau associées à des caractéristiques qui donnent un sens à certaines zones d'intérêt sur Terre.

Cet article traite de l'intégration de données calculées à partir de rasters et de données ouvertes sur la base de leurs propriétés spatiales et temporelles afin de qualifier des zones géographiques d'intérêt. Nous utilisons la notion d'*unité territoriale*, définie comme une division d'un territoire plus vaste, selon un critère lié aux activités humaines (administration, droit, agriculture, ...) et normalisée dans des nomenclatures légalement définies. Les *zones d'intérêt* correspondent alors à certaines unités territoriales sélectionnées pour leur pertinence pour une tâche donnée. Elles sont généralement représentées sous forme d'entités géospatiales dans un format vectoriel. Le croisement de données raster (au format matriciel) avec ces zones d'intérêt et même avec d'autres données au format vectoriel n'est donc pas trivial.

Dans la lignée des travaux de [1, 2], nous proposons de représenter ces données au sein de graphes de connaissances, afin d'en faciliter l'intégration, mais aussi l'exploitation, l'interrogation ainsi qu'une restitution intelligible auprès des utilisateurs. Nous sommes intéressés à étudier (i) quel type d'ontologie est nécessaire pour soutenir l'extraction de connaissances à partir de données de raster d'OT et pour décrire de manière homogène les résultats d'analyse de ces données, également fournis au format raster ; (ii) comment rendre accessibles et utilisables des données d'OT riches

1. http://www.copernicus.eu/en

collectes grâce au traitement d'images et d'autres types d'OT ; et (iii) comment améliorer la traçabilité des données au fil de leur traitement (sources de données, calcul des rasters, processus sémantique) pour améliorer la confiance des utilisateurs et l'exploitation des données.
Le projet européen CANDELA [2] vise à créer une plateforme fournissant des modules et des services permettant aux utilisateurs de manipuler, d'explorer et de traiter rapidement les données Copernicus ainsi que de grands ensembles de données ouvertes. Nous contribuons à ce projet en proposant une intégration sémantique des données tirées des images raster et de données ouvertes, et par un module de recherche sémantique sur les données intégrées.

2 Contributions

Les principales contributions de cette étude concernent les composants suivants :
Un modèle sémantique générique qui permet la description sémantique et homogène de données spatio-temporelles pour qualifier des zones prédéfinies et garde la trace de leur provenance. Ce modèle est extensible pour traiter tout type de propriété d'OT observée et a été appliqué à plusieurs cas d'utilisation. Il est composé de plusieurs sous-modèles interconnectés décrivant les différents types de données : *tom*, un modèle d'observation territoriale permettant de représenter les unités territoriales et les observations associées (tirées des fichiers raster) ; *eom*, un modèle d'OT permettant de représenter les métadonnées des images Sentinel ; et *eoam*, un modèle d'analyse d'OT permettant de représenter les rasters produits par le traitement des images ou les vecteurs.
Un processus sémantique configurable et reproductible de type *Extraction, transformation et chargement* (ETL) [3] basé sur le modèle proposé. Nous avons défini un ensemble de fonctions de transformation pour peupler le modèle sémantique avec des données et obtenir une représentation sémantique homogène des données. Ce processus extrait les données des rasters et les agrège avec des données provenant d'autres sources. L'agrégation a lieu sur les zones des unités territoriales. Ensuite, ce processus relie les données extraites aux concepts du modèle sémantique et assigne les données à une unité territoriale.
Un éco-système EO Sentinel qui permet d'exploiter des données matricielles tirées d'images Sentinel (disponibles au format raster), de représenter et de calculer différentes propriétés à partir de ces données puis d'importer d'autres ensembles de données géolocalisées, matricielles ou vectorielles, à partir de sources externes (par exemple, des données sur la couverture du sol).

3 Evaluation

Nous avons évalué notre approche en termes d'adaptabilité du modèle proposé pour répondre à différents cas d'utilisation (surveillance des vignobles et de l'expansion urbaine), l'adaptabilité de la chaîne de traitement, et la valeur ajoutée des ensembles de données générés pour aider à la prise de décision. Nous discutons également de l'évolutivité de l'approche et de la relation entre la résolution de l'image et la taille des unités territoriales de référence.
Les données sémantiques générées pour ces cas d'utilisation sont stockées dans une base de données à laquelle on peut accéder via une interface de recherche sémantique ou un point d'accès SPARQL [4].

4 Conclusions et travaux futurs

L'approche que nous proposons pour intégrer des données calculées à partir de rasters et d'autres données ouvertes permet de qualifier des unités territoriales sur la base de leurs caractéristiques spatiales (vecteurs) et temporelles.
Nous prévoyons d'étendre ce travail pour assurer le passage à l'échelle de l'approche, en exploitant des scénarios de big data pour la gestion des zones Natura 2000 (pertinentes pour l'étude de l'évolution de l'occupation du sol et la détection des changements dans les zones de conservation). Nous prévoyons également de traiter des données provenant de fichiers CSV, en particulier les observations météorologiques de Météo France [5]. Enfin, considérant les cubes de données comme des tableaux de données multidimensionnels fréquemment utilisés pour enregistrer des données géolocalisées, nous envisageons que le processus d'intégration gère ce type de structure, ce qui est rarement le cas dans l'état de l'art actuel.

Remerciements

Ce travail a bénéficié d'une subvention H2020 de la Communauté Européenne pour le projet CANDELA.

Références

[1] L. Ding, G. Xiao, D. Calvanese, and L. Meng. A framework uniting ontology-based geodata integration and geovisual analytics. *ISPRS International Journal of Geo-Information*, 9(9), 2020.

[2] Daniela Espinoza-Molina, Charalampos Nikolaou, Corneliu Octavian Dumitru, Konstantina Bereta, Manolis Koubarakis, Gottfried Schwarz, and Mihai Datcu. Very-High-Resolution SAR Images and Linked Open Data Analytics Based on Ontologies. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(4) :1696 – 1708, 2015.

[3] Ba-Huy Tran, Nathalie Aussenac-Gilles, Catherine Comparot, and Cassia Trojahn. Semantic integration of raster data for earth observation on territorial units. *ISPRS Int. Journal of Geo-Information*, 11(2), 2022.

[4] Jesús Villegas, Hector Sánchez Pastor, Lorena Hernanz, María Checa, and Dumitru Roman. Enabling the use of sentinel-2 and lidar data for common agriculture policy funds assignment. *International Journal of Geo-Information*, 6 :255, 08 2017.

2. http://www.candela-h2020.eu/
3. Accès à l'image docker qui encapsule ce processus :https://hub.docker.com/r/h2020candela/triplification
4. http://melodi.irit.fr/tom/
5. http://www.meteofrance.com/