Relations topologiques pour l'intégration sémantique de données et images d'observation de la Terre

Herbert Arenas, Nathalie Aussenac-Gilles, Catherine Comparot, Cassia Trojahn

IRIT, CNRS, Université de Toulouse, France
{prenom.nom}@irit.fr

RÉSUMÉ. *Les satellites d'observation de la Terre lancés récemment par l'ESA délivrent entre 8 à 10 To de données par jour, offrant de nouvelles opportunités pour la gestion de l'environnement, l'étude du climat ou de l'évolution urbaine. Les applications de ces domaines requièrent d'enrichir les méta-données d'image avec des données provenant de diverses sources (fixes ou dynamiques) pour faciliter la prise de décision sur les zones étudiées. L'intégration de données hétérogènes soulève un défi majeur. Nous présentons une approche s'appuyant sur les représentations spatio-temporelles pour enrichir des métadonnées d'images satellites avec des données ouvertes. Elle s'appuie sur un vocabulaire qui spécialise des standards (comme SOSA et GeoSPARQL) ainsi qu'un processus pour aligner et intégrer des données géospatiales hétérogènes. Le processus exploite le tuilage des images, représentant une zone fixe d'une grille associée à la surface terrestre, pour traiter les données ayant une composante spatiale fixe. Les relations temporelles, quant à elles, sont calculées à la volée à partir d'une topologie temporelle.*

ABSTRACT. *Recently launched Earth observation satellites, which deliver between 8 and 10TB of image data per day, open emerging opportunities in domains ranging from environmental monitoring to urban planning and climate studies. However, domain-oriented applications require image metadata to be enriched with data coming from various sources (either static or dynamic), in order to support decision making on the observed areas. The integration of heterogeneous data highly relying on spatio-temporal representations raises a major challenge. We present a semantic approach to support data integration thannks to spatio-temporal relations between image metadata and various open data sets. We propose a vocabulary that specializes standards (like SOSA, GeoSPARQL) as well as a process to map and integrate heterogeneous geo-spatial data sets. This process relies on image tiles, representing a fixed area of a grid associated with the Earth's surface, to handle data with a fixed spatial component. The temporal relationships are calculated on the fly based on temporal topology.*

MOTS-CLÉS : *intégration de données, vocabulaire sémantique, observations de la Terre*

KEYWORDS: *semantic vocabulary, earth observation, data integration*

1. Introduction

L'observation de la Terre offre une valeur ajoutée à un grand nombre de domaines. Récemment l'Agence Spatiale Européenne (ESA) a lancé le programme Sentinel avec deux types de satellites, Sentinel-1 and Sentinel-2, qui transmettent des images de haute qualité (entre 8 à 10 To de données quotidiennement). Ces images sont captées selon différentes technologies et libres d'accès. Cette disponibilité des données ouvre de nombreuses perspectives économiques grâce à de nouvelles applications dans des domaines aussi variés que l'agriculture, l'environnement, l'urbanisme, l'océanographie ou la climatologie. Ces applications métier ont néanmoins besoin de coupler les images avec des données sur les zones observées. Ces données sont accessibles à partir de différentes sources dans des formats hétérogènes et des temporalités différentes : elles peuvent être statiques, comme les données sur le relief ou la couverture terrestre, ou dynamiques, comme les observations météorologiques. Elles peuvent être utiles par exemple pour indiquer qu'une image contient une région touchée par un phénomène tel qu'un tremblement de terre ou une canicule, et sont alors utilisées pour décider des actions à mener dans cette zone ou conduire à des analyses à plus long terme. Plus encore, en exploitant les caractéristiques spatio-temporelles d'un phénomène (son emprise spatiale et sa date), il devient possible de savoir si une entité localisée dans l'emprise de l'image (e.g. une ville) a subi le même phénomène.

Dans ce contexte, les images étant décrites par des méta-données, une des difficultés est d'intégrer à ces méta-données, des données hétérogènes provenant de sources diverses. L'apport des technologies sémantiques pour faciliter cette tâche a été démontré dans des travaux antérieurs (Reitsma, Albrecht, 2005) (Sukhobok *et al*., 2017). En lien avec ces travaux, nous présentons une approche sémantique, basée sur un vocabulaire, pour intégrer des données en vue d'enrichir des méta-données d'images satellites avec des données provenant de sources diverses. Le vocabulaire sémantique doit être défini de manière à représenter les données et à y accéder de façon homogène. Cette approche requiert aussi des règles de transformations pour peupler le modèle avec les données de ces sources hétérogènes. Une caractéristique essentielle des observations de la Terre est qu'elles sont géo-localisées et datées. Elles peuvent donc être liées par des relations topologiques spatiales et temporelles. Le processus d'intégration des données doit gérer correctement les propriétés et relations spatiales et temporelles. Pour éviter de dupliquer des données fixes, i.e. valides pour toutes les images d'une même zone au cours du temps, il est commode d'exploiter le concept de tuile ("tile" en anglais) défini par l'ESA : la surface terrestre est associée à une grille dans laquelle une tuile représente une zone fixe de cette surface.

Nous présentons ici un cadre pour l'intégration sémantique de diverses données géographiques et des méta-données d'images satellites. Celui-ci s'appuie sur un vocabulaire que nous avons défini ; il permet d'associer les mêmes classes à ces différents types de données, et de les représenter comme des entités ayant des propriétés spatiales et temporelles. Les données proviennent d'ensembles de données géospatiales avec des formats hétérogènes (shapefile, KML, CSV, GeoJSON, TIFF). Une partie de ces données sont dites "contextuelles" et sont le résultat de mesures : nous les trai-

tons comme des données de capteurs. Ce vocabulaire spécialise ainsi des vocabulaires connus du LOD, dont SOSA [1] et GeoSPARQL (Kolas *et al*., 2013).

Une seconde contribution est le processus d'intégration basé sur la topologie des entités et les principes des données liées afin de gérer les problèmes d'hétérogénéité. Pour chaque ensemble de données à intégrer, nous avons défini des patrons et fonctions de transformation. Les propriétés temporelles contribuent à l'intégration des données dynamiques. Pour traiter la composante spatiale des données statiques et dynamiques, le processus s'appuie sur le tuilage des images qui permet de réduire le volume de données à traiter. Enfin, le processus d'intégration génère plusieurs entrepôts de données et des fichiers JSON de méta-données enrichies ou de mesures qui peuvent être réutilisés à des fins diverses. Nous illustrons notre approche par un cas d'étude qui exploite des méta-données d'images Sentinel-2 fournies par le CNES, les tuilages d'images de l'ESA et des données contextuelles : des données de météorologie fournies par Météo France, la couverture végétale terrestre et les entités administratives. Grâce à la représentation sémantique de toutes ces données [2], nous avons lié les méta-données de chaque image aux données s'appliquant à la zone terrestre définie par l'emprise (ou *footprint*) de cette image à la date de sa saisie.

Le reste de l'article est organisé comme suit. La Section 2 discute des travaux liés. La Section 3 offre un aperçu de notre approche. La Section 4 présente le modèle proposé, et la Section 5 détaille les processus de sélection, d'alignement et d'intégration de données. Nous concluons et présentons des perspectives à ce travail en Section 6.

2. Publication et mise en relation de données d'observation de la Terre

Publication de données liées d'observations de la Terre. Rendre disponibles sous forme de données ouvertes des données géo-localisées et les relier à des bases de connaissances couvrant d'autres aspects du domaine facilite le développement de services ayant une grande valeur environnementale et commerciale (Smeros, Koubarakis, 2016). Dans ce but, les principes du LOD (Linked Open Data) définissent des bonnes pratiques pour exposer, partager et intégrer des données au format RDF et identifiées par des URI déréférençables sur le Web (Heath, Bizer, 2011 ; Blázquez *et al*., 2014 ; Sukhobok *et al*., 2017). Le W3C fournit d'ailleurs des recommandations pour publier des données spatiales sous forme de LOD et gérer les relations spatiales. Il propose aussi des systèmes de référence (CRS ou *Coordinate Reference Systems*) pour leur réprésentation (Tandy *et al*., 2017). Des projets européens tels que LEO et TELEIOS ont commencé à publier des données liées au sein d'*Observatoires Virtuels* promus par l'initiative internationale IVOA (International Virtual Observatory Alliance) (Koubarakis *et al*., 2012). Grâce aux nouveaux liens identifiés entre les données et aux connaissances inférées, ces observatoires virtuels fournissent des en-

1. https://www.w3.org/2015/spatial/wiki/SOSA_Ontology
2. Les données sont publiées sur http://melodi.irit.fr/sparkindata/

sembles d'informations plus riches que les images d'observation de la Terre et leurs méta-données seules (Koubarakis *et al*., 2012).

Les données géospatiales sont souvent disponibles au format "raster", un format défini initialement pour les images. Pour représenter ces données, souvent volumineuses, dans le LOD, le W3C suggère l'ontologie QB (RDF Data Cube) (Brizhinev *et al*., 2017) combinée à d'autres ontologies standards du W3C et de l'OGC dont SSN (Semantic Sensor Network) [3], OWL-Time [4], SKOS [5], PROV-O [6] et la récente extension de DataCube pour les entités spatio-temporelles, QB4ST [7]. Pour représenter sous forme sémantique les données spatio-temporelles des CRS, les entités topographiques et leurs géométries, plusieurs modèles existent. A cette fin, Atemezing a défini quatre vocabulaires qui étendent des vocabulaires existants et offrent deux avantages supplémentaires (Atemezing, 2015) : une utilisation explicite du CRS identifié par des URI pour la géométrie, et la possibilité de décrire des géométries structurées en RDF. De même, le projet GeoKnow a tiré parti des données spatiales du LOD et mis à disposition des outils pour collecter, fusionner, agréger des données spatiales ainsi qu'une architecture pour les publier, les réutiliser et les visualiser (García-Rojas *et al*., 2013).

Mise en relation de données spatio-temporelles et découverte automatique de liens. Lier des données d'observation de la Terre signifie découvrir des liens spatiaux et temporels au sein du graphe RDF obtenu après la publication des données (Blázquez *et al*., 2012). Grâce aux propriétés spatiales, les données d'observations peuvent être associées aux tuiles et ainsi aux images d'observation de la Terre. Grâce aux propriétés temporelles, les observations temporelles peuvent aussi être liées aux images. Lorsque des entités de même nature sont collectées à partir de diverses sources, un algorithme d'association d'entités peut identifier des alignements entre des entités spatiales similaires ou identiques. On peut définir ces algorithmes de manière à ne prendre en compte que les propriétés spatiales et temporelles pour produire ces alignements.

L'OGC a introduit la notion de *données géo-liées* ("geolinked data") pour faire référence aux données liées géographiquement. Dans les premiers travaux, la géométrie était stockée dans un ensemble de données géospatiales séparé, et non directement comme valeur d'attribut. Cette option est plus contraignante lorsqu'il faut comparer la géométrie de chaque entité. C'est pourquoi les entrepôts actuels mémorisent ensemble une représentation RDF de la géométrie et une représentation RDF des entités spatiales. Suivant la source, la géométrie de chaque donnée est décrite par un ou des points, lignes ou polygones. Atemezing a identifié des outils pour construire une représentation RDF de la géométrie, comme Geometry2RDF [8] ou TripleGeo [9] (Atemezing,

3. http://purl.oclc.org/NET/ssnx/ssn
4. https://www.w3.org/TR/owl-time
5. http://www.w3.org/2004/02/skos/core
6. https://www.w3.org/TR/prov-o
7. https://www.w3.org/TR/qb4st/
8. https://github.com/boricles/geometry2rdf
9. https://github.com/GeoKnow/TripleGeo

2015). Le processus défini par Vilches-Blázquez et ses collègues compare précisément les géométries de données de telle sorte que les données spatiales puissent être retrouvées et reliées à un haut niveau de granularité (Blázquez *et al*., 2014). Pour aller plus loin et retrouver précisément tout ce qui est localisé à un endroit précis, les images de satellites peuvent être classées et enrichies de données externes dans un format sémantique qui permet de raisonner sur ces données grâce à des règles de raisonnement spatial spécifiques au domaine (Alirezaie *et al*., 2017).

Pour calculer des liens entre des ressources LOD possédant des propriétés temporelles, et donc assimilables à des événements, Georgala et ses collègues utilisent les intervalles de l'algèbre des intervalles d'Allen (Georgala *et al*., 2016). Leur proposition peut s'appliquer aux données géolocalisées ayant une dimension temporelle. Leur approche, AEGLE, réduit le nombre de relations temporelles d'Allen de 13 à 8, et les implémente de façon optimale pour effectuer plus rapidement les comparaisons de propriétés temporelles nécessaires pour calculer les relations temporelles.

Une autre facette de la mise en relation des données est traitée dans l'état de l'art par l'appariement d'entités (*entity resolution*) (Shen *et al*., 2015). Il s'agit d'associer entre elles des entités équivalentes, ce qui a un enjeu dans des domaines comme les bases de données relationnelles, la recherche d'information ou encore l'annotation de textes. Plus généralement, la découverte de liens (*entity linking*) vise à trouver des liens sémantiques entre des entités issues de différentes bases de connaissances (Auer *et al*., 2011 ; Smeros, Koubarakis, 2016). Selon (Smeros, Koubarakis, 2016), les approches de l'état de l'art se concentrent sur la recherche d'équivalence entre les entités (mêmes étiquettes, mêmes noms ou mêmes types), laissant d'autres types de relations, par exemple les relations spatiales ou temporelles, inexploitées. Ces auteurs proposent donc d'utiliser les liens spatio-temporels pour calculer plus de relations. Or la représentation spatiale de la plupart des données géo-localisées est complexe, sous forme de polygone. Le calcul des relations entre polygones au sein de très grands jeux de données est particulièrement complexe et long. Une étape de pré-traitement est nécessaire pour transformer les données (issues de vocabulaires RDF, de CRS, de sérialisations, etc.) selon un modèle unique. Ensuite, une technique de *blocking* vise à réduire la complexité du calcul. Elle consiste à découper en "blocs" (rectangles incurvés) la surface terrestre, puis à évaluer les relations topologiques entre entités en se basant sur ce découpage. De même, Sherif et ses collègues (Sherif *et al*., 2017) proposent de découvrir des liens topologiques encore plus efficacement grâce à une indexation des entités à l'aide de tuiles découpant la surface terrestres en rectangles. Cette méthode accélère le calcul de relations topologiques entre deux géométries d'entités dans le plan.

3. Une approche sémantique pour l'intégration spatiale et temporelle de données d'observations de la Terre

3.1. Principes de l'approche d'intégration

Reprenant certains des principes de cet état de l'art, nous proposons une approche sémantique pour l'intégration de données d'observations de la Terre qui s'appuie sur

leurs propriétés spatiales et temporelles. Nous nous intéressons en particulier à des données géo-localisées tirées de sources ouvertes et aux méta-données d'images satellites. Comme (Atemezing, 2015), nous proposons une ontologie, à savoir un vocabulaire formel qui étend des vocabulaires standards présents sur le LOD pour mieux représenter ces données comme des entités associées à des classes et possédant des propriétés spatiales (une géométrie) et temporelles (une datation). Pour intégrer ces données, nous nous appuyons d'abord sur leur dimension spatiale et comme (Smeros, Koubarakis, 2016), nous avons recours à la notion de tuilage pour réduire les coûts de calcul des relations spatiales entre entités représentant les données et les images. Cependant, nous avons choisi de nous limiter aux relations spatiales définies par GeoSPARQL afin d'utiliser ce langage pour interroger les données. Dans un deuxième temps, l'intégration prend en compte les propriétés temporelles des données pour associer les données pertinentes par rapport à la prise de vue d'une image.

Ce travail a été réalisé dans le cadre du projet SparkinData [10] visant à construire une plate-forme cloud de données d'observations de la Terre et à valoriser les images de la collection Sentinel-2. Nous avons évalué notre approche grâce à un cas d'étude où nous exploitons les dimensions spatiales et temporelles pour lier les méta-données d'images avec les unités administratives publiées sur le LOD de l'INSEE et des données météorologiques fournies par MétéoFrance. Nous rendons accessibles ces données via un point d'accès [11] SPARQL.

3.2. Architecture

Cette approche d'intégration est mise en oeuvre au sein d'une plate-forme dont l'architecture est modulaire (Figure 1). Ses différents niveaux permettent de découpler les étapes du processus permettant de passer des données brutes aux données sémantisées. Elle est composée des modules suivants :

– **Sélection des données** : la première étape du processus d'intégration des données est l'identification et l'accès aux sources de données à collecter. Un ensemble de données est soit un fichier, soit le résultat d'une requête d'interrogation d'un entrepôt de données. Les formats traités pour le moment sont CSV, JSON, RDF, XML, GeoTIFF, et Shapefile. Les sources de données utilisées sont décrites en Section 5.1.

– **Conversion des données** : Les données des sources sélectionnées sont dans un premier temps converties dans une représentation pivot en JSON. Pour cela, nous avons soit réutilisé des scripts dédiés soit développé nos propres scripts, selon le type de données de la source considérée. Les fichiers JSON intermédiaires sont stockés dans une base de données MongoDB comme sauvegarde de sécurité. Des exemples de conversion de données sont présentés en Section 5.2.

10. SparkInData fait partie du programme français d'Investissement d'Avenir (FUI).
11. http://melodi.irit.fr/sparkindata/

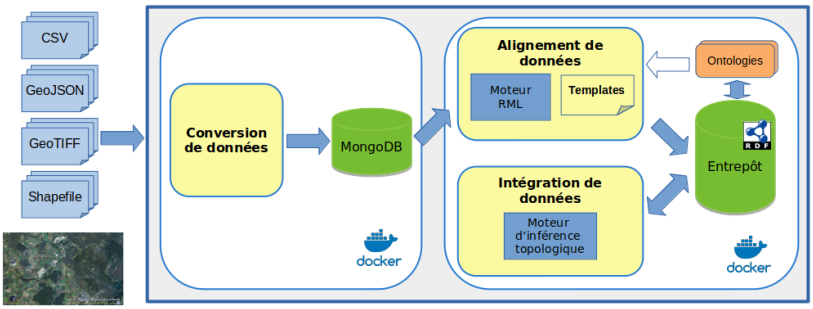

FIGURE *1. Architecture d'intégration des données issues de sources hétérogènes.*

– **Alignement des données** : Les données des fichiers JSON sont transformées en instances de classes de l'ontologie présentée en Section 4. Nous avons défini pour cela un template d'alignement (i.e. un modèle RDF de triplets à produire) et implémenté un mécanisme de traitement au sein d'un module Python. Nous fournissons dans la Section 5.2, des exemples de templates. A partir des valeurs présentes dans les documents JSON, les fonction du module Python réalisent des opérations sophistiquées qui ne sont pas possibles dans les approches alternatives telles que RML.

– **Intégration des données** : Le processus d'intégration s'appuie sur les relations topologiques, soit spatiales, soit temporelles, entre les instances des classes du modèle. A ce stade, il est possible de calculer les relations topologiques entre toutes les instances ayant une représentation spatiale, et de les stocker comme des assertions dans le triplestore. Une alternative est d'évaluer les relations topologiques à la volée, notamment les relations temporelles. Dans ce cas, pour réduire le temps et le coût de calcul, on peut opérer une sélection des relations à considérer ou des instances à lier.

4. Un modèle pour l'intégration de données d'observations de la Terre

4.1. Vocabulaires réutilisés

Notre modèle ontologique pour l'intégration des données s'appuie sur deux vocabulaires existants, l'ontologie noyau SOSA et l'ontologie GeoSPARQL. Dans un de nos précédents travaux (Arenas *et al.*, 2016), nous avions utilisé respectivement DCAT et SSN pour représenter les enregistrements de méta-données et les données météorologiques. Désormais, nous adoptons SOSA comme ontologie noyau pour ces deux types de données.

SOSA est une ontologie légère, indépendante, représentant les classes et propriétés élémentaires de SSN (Semantic Sensor Network). SOSA décrit des capteurs et leurs observations, les procédures mises en oeuvre, les éléments d'intérêt étudiés, les échantillons utilisés, et les propriétés mesurées. SOSA est pertinent pour une vaste gamme d'applications, dont l'imagerie satellite. Nous l'avons donc adoptée pour décrire les

méta-données d'image comme des *observations de la Terre* (instances de *EarthObservation*) et les observations météo comme des *observations météo* (instances de *MeteoObservation*) (Figure 2). Nous avons néanmoins spécialisé SOSA pour mieux typer les instances de ces concepts, même si la tendance dans les domaines où SOSA a été adopté, comme l'IoT, est d'éviter ce type construction et d'utiliser directement SOSA comme vocabulaire principal (Pomp *et al*., 2017).

GeoSPARQL, un standard de l'OGC, définit une petite ontologie pour représenter des caractéristiques, des relations et des fonctions spatiales (Kolas *et al*., 2013) (Battle, Kolas, 2012). Il existe des alternatives à GeoSPARQL comme GeoRDF qui permet de représenter des données simples telles que la latitude, la longitude, l'altitude, comme des propriétés de points (en utilisant WGS84 comme référentiel), ou encore GeoOWL qui permet d'exprimer des objets plus complexes (lignes, rectangles, polygones). Nous avons retenu GeoSPARQL qui permet de raisonner sur des géométries, et de proposer ainsi des relations (inclusion, recouvrement, etc.) entre des entités sur la base des relations entre leurs géométries.

4.2. Un modèle étendu pour l'intégration de données aux méta-données d'images

Le modèle ontologique que nous proposons est détaillé sur la Figure 2 et est organisé en modules. Il intègre certaines classes et propriétés (en particulier les propriétés temporelles) de SOSA (module `sosa` qui réutilise la classe `time:TemporalEntity` du vocabulaire OWL Time), ainsi que des classes et propriétés de GeoSPARQL (module `geo`) comme `SpatialObject`, `Feature` et `Geometry`. Ce modèle comporte aussi des classes et propriétés spécifiques à notre modèle. Deux modules sont dédiés à la représentation des images d'observation de la Terre : `eom` pour les méta-données d'images (qui spécialise `sosa` et `geo`) et `grid` pour représenter les tuiles (qui spécialise `geo`). Ensuite, il convient de définir des classes pour décrire chaque jeu de données que l'on souhaite intégrer, et de les relier aux classes de `geo` et si besoin de `sosa`. Dans notre cas d'étude, les classes de `mfo` (qui spécialise `sosa`) servent à représenter les données météorologiques et les stations météo, alors que `admin` permet de représenter des unités administratives.

Toute instance de la classe `sosa:Observation` possède une dimension temporelle. Nous définissons donc un enregistrement de méta-données d'image par la classe `eom:EarthObservation` qui spécialise `sosa:Observation`. Sa dimension temporelle identifie le moment où l'image a été prise. De même, les stations météo enregistrent périodiquement des mesures. Nous définissons donc la classe `mfo:MeteoObservation` comme une sous-classe de `sosa:Observation` pour représenter les données mesurées par une station météo. Nous pouvons alors lier par une relation temporelle (before, after) un enregistrement de méta-données d'image et des mesures météo ou mémoriser des périodes d'intérêt (e.g., une semaine après la prise de l'image).

Pour représenter la géo-localisation des images, le modèle s'appuie sur leur emprise (ou footprint), qui est un polygone fermé (une géométrie) correspondant à la

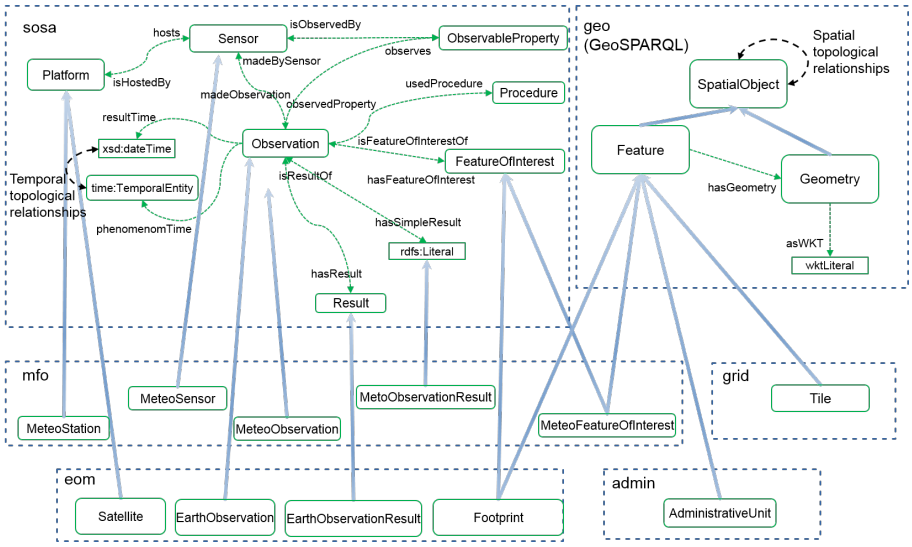

FIGURE 2. *Le modèle d'intégration. SOSA et GeoSPARQL sont spécialisés dans 4 modules dédiés à chaque source de connaissances et au tuilage des images.*

zone géographique couverte par l'image. Elle est représentée à l'aide des classes `eom:Footprint` et des tuiles, représentées comme des `grid:Tile`. Ces deux classes sont définies comme des spécialisations de `geo:Feature`. `eom:Footprint` spécialise aussi `sosa:FeatureOfInterest`.

De même, les données à intégrer sont géo-localisées, et donc définies comme des sous-classes de `geo:Feature`. C'est le cas des données météo via la classe `mfo:MeteoFeatureOfInterest` et des données sur les unités administratives `admin:AdministrativeUnit`. Images et données météo peuvent aussi être liées par des relations spatiales calculées à partir des coordonnées `geo:wktLiteral` de leur `geo:Geometry`. Pour cela, nous utilisons les relations topologiques (contient, recouvre, etc.) proposées par GeoSPARQL. Ces relations associent deux ressources (deux `geo:Geometry` ou deux `geo:Feature`) grâce aux propriétés topologiques (propriétés directes) ou à des fonctions topologiques (propriétés calculées).

Au sein du modèle `mfo`, une station météo est représentée comme une instance de la classe `mfo:MeteoStation` alors que la position géographique des stations est une propriété `hasPosition` (non mentionnée sur la figure) ayant respectivement pour domaine et co-domaine les classes `mfo:MeteoStation` et `geo:Feature`. Les capteurs fonctionnant sur une station météo sont représentés comme des instances de `mfo:MeteoSensor`, sous-classe de `sosa:Sensor`. La position géographique d'une station est représentée comme instance de la classe `mfo:MeteoFeatureOfInterest`, une sous-classe de `sosa:FeatureOfInterest`. `mfo:MeteoFeatureOfInterest` est aussi une sous-classe de `geo:Feature`. Ainsi, connaissant

la position d'un `mfo:MeteoFeatureOfInterest`, il est facile d'identifier les caractéristiques d'un autre type qui recouvrent les observations météo.

Afin de lier les observation de la Terre à des unités administratives françaises (régions, départements et villes) à partir de leur position géographique (point ou polygone), nous avons enrichi le modèle avec la classe `admin:AdministrativeUnit`, une sous-classe de *geo:Feature*. Enfin, pour les images Sentinel 2 Single Tile (S2ST), les tuiles correspondent aux *features of interest* des images. Un S2ST correspond à un fragment de l'image originale d'une taille approximative de 100 x 100 km. L'intérêt par rapport à une image S2 normale est que l'utilisateur peut sélectionner la surface qui l'intéresse et ne télécharger que l'information souhaitée [12]. Nous avons également enrichi le modèle avec les classes principales de *Global Land Cover* : `Artificial Surfaces`, `Cropland`, `Tree Covered Areas`, etc.

5. Sélection, conversion et alignement de données

5.1. Sélection des données

La finalité du processus étant l'intégration de données via des relations spatiales ou temporelles, les propriétés requises pour calculer ces relations (localisation, datation) doivent être accessibles. Selon le cas, ces données sont fournies avec la source à intégrer ou bien mémorisées à part. Dans ce dernier cas, il est nécessaire de recourir à des sources de données complémentaires. Nous distinguons les sources de données dynamiques, pour lesquelles la dimension temporelle est importante (comme c'est le cas des données de capteurs), des sources de données statiques, pour lesquelles seule la dimension spatiale est requise dans notre processus.

Les sources des données dynamiques. Dans le projet SparkInData nous utilisons des enregistrements de méta-données d'images Sentinel [13]. La périodicité de Sentinel-1 est de douze jours, tandis que celle de Sentinel-2 est de cinq jours. Les enregistrements de méta-données sont obtenus au format GeoJSON (format JSON pour encoder des données géospatiales) à partir de l'API RESTO, un service de données géré par le CNES (Gasperi, 2014). L'URL suivante par exemple retourne tous les enregistrements de méta-données de la collection Sentinel-2 Single Tile pour la France, réalisés entre le 19/09/2017 23:00 et le 25/09/2017 00:00 :

https://peps.cnes.fr/resto/api/collections/S2ST/search.json?q=France\&startDate=17-09-19T23:00:00\&completionDate=2017-09-25T00:00:00.

Les requêtes faites avec cette API peuvent spécifier les paramètres à retrouver, i.e. des métadonnées spécifiques comme la couverture nuageuse, l'intervalle de temps, la zone géographique d'intérêt, etc. Nous collectons ces informations toutes les nuits.

12. Un fichier S2ST est moins volumineux : il peut faire environ 500Mo alors que celui d'une image Sentinel-2 avant tuilage peut faire plus de 3Go.

13. https://sentinel.esa.int/web/sentinel/missions/ (07/2016)

Les données contextuelles que nous utilisons sont les informations météo fournies par *SYNOP Meteo France* [14] sous forme de fichiers CSV zippés. Les observations sont prises toutes les trois heures dans chacune des 62 stations françaises. Un fichier contenant la liste des stations avec leur position respective, i.e. un point fixe repéré par ses coordonnées géographiques, est fourni séparément.

Les sources des données statiques. Pour les images S2ST, des informations sur la couverture spatiale de l'image sont obtenues à partir des méta-données sous deux formes : 1) le *footprint* de l'image, 2) l'identifiant de la tuile qui lui correspond. Le fichier grid KML qui indique l'emprise de chaque tuile ainsi que son nom est fourni par l'ESA [15]. Nous avons donc traité ce fichier.

Une autre source de données que nous avons exploitée est le GLC-SHARE (Global Land Cover SHARE) produit par le FAO, qui donne des informations sur la couverture terrestre. Elle s'appuie sur une nomenclature qui classe les zones en fonction du type d'occupation des sols ou du type de surface ; 11 classes sont définies telles que surface artificielle (01), terre cultivée (02), zone forestière (03), etc. Les données du GLC-SHARE sont fournies sous forme d'image au format *GeoTIFF* (format TIFF incluant des informations de géo-référencement) dont chaque pixel correspond à une surface d'environ 1 km^2. La valeur d'un pixel est un entier indiquant la classe la plus fréquente pour la zone couverte par le pixel. Nous avons exploité cette source pour associer des données aux tuiles des images S2ST. Nous avons ainsi calculé la composition de la couverture terrestre de chacune de ces tuiles : sous forme d'un pourcentage des différentes classes GLC-SHARE sur la surface couverte par la tuile.

Finalement, nous collectons des données RDF sur les unités administratives françaises à partir de la base de connaissances de l'INSEE [16]. Ces données n'étant pas géo-localisées, il n'était pas possible de les intégrer aux méta-données d'images. Nous avons donc utilisé la plate-forme française des données publiques "data.gouv.fr" [17] pour obtenir ces informations, accessibles au format shapefile.

5.2. Transformation et alignement des données

Nous venons de présenter les données que nous exploitons, leurs diverses sources et la variété de leurs formats originaux. Afin de standardiser les traitements, nous représentons toutes ces données au format JSON puis nous les convertissons en RDF à l'aide de mécanismes d'alignement. Toutefois, certaines données, par exemple la couverture terrestre ou les données météo, sont aussi exploitées en amont du processus de transformation RDF pour produire d'autres données (moyennes, etc.) adaptées aux besoins. Nous décrivons à présent chacun de ces processus.

---

14. https://donneespubliques.meteofrance.fr/ (07/2016)
15. https://sentinel.esa.int/web/sentinel/missions/sentinel-2/news/-/article/sentinel-2-tiling-grid-updated
16. http://rdf.insee.fr/def/index.html
17. https://www.data.gouv.fr

Les unités administratives. Un langage classique pour convertir des données en RDF est RML (*RDF Mapping Language*), un langage de règles de transformation d'un format donné (e.g. JSON) en une représentation RDF. Une des limites de RML est qu'il ne permet pas de faire appel facilement à des fonctions spécialisées pour traiter des informations spécifiques. Le document JSON suivant illustre ce défaut :

```
{"wkt": "MULTIPOLYGON(((
-1.0988062299633785 45.64032288975508, ...
-1.0988062299633785 45.64032288975508))...)",
"name": "Poitou-Charentes", "geomType": 5,
"inseeInfo": {"adminType": "region", "insee": "54"}}
```

Ce code décrit une unité administrative française comme un ensemble d'attributs et de valeurs. La valeur de l'attribut `wkt` est la géométrie codée en WKT (Well known text), et la clé `name` est une chaîne de caractères donnant le nom de cette unité ("Poitou-Charentes"). La clé `inseeInfo` contient des informations en référence à l'identification INSEE de cette unité. A partir de la valeur de `inseeInfo`, on peut récupérer l'URI de l'unité administrative dans la base de connaissances de l'INSEE. Ceci nécessite tout de même de créer une requête SPARQL pour interroger la base de données de l'INSEE, ce qui ne peut pas être réalisé avec une règle RML.

Tout en conservant une approche similaire à RML, nous avons développé une solution alternative et mieux adaptée pour transformer le JSON en RDF. Cette solution comprend un template de triplets et un processeur codé en Python. Le code suivant est un exemple de template qui transforme l'extrait de document JSON montré plus haut en RDF à l'aide du vocabulaire `admin` de l'ontologie.

```
@prefix geo: <http://www.opengis.net/ont/geosparql#> .
@prefix admin: <http://melodi.irit.fr/ontologies/administrativeUnits.owl#> .
# ce template definit la structure d'une unite administrative
<dummy> a getUrlAdministrativeUnitType(\$.inseeInfo.adminType) .
<dummy> admin:hasInseeCode stringToLiteral(\$.inseeInfo.insee) .
<dummy> admin:hasName stringToLiteral(\$.name) .
# representation spatiale de l'unite administrative
<dummy> a geo:Feature .
<dummy> geo:hasGeometry <dummy_geo> .
<dummy_geo> a geo:Geometry .
<dummy_geo> geo:asWKT valueToWktLiteral(\$.wkt) .
# l'instance est liee a l'unite administrative de l'INSEE
<dummy> owl:sameAs getInseeUrl(\$.inseeInfo) .
```

Le template est constitué de triplets dont les variables sont remplacées par les valeurs lues dans le document JSON. Pour les valeurs contenues qui nécessitent des traitements supplémentaires, nous avons développé des fonctions auxquelles nous passons en paramètres le chemin JSON vers les informations à extraire du fichier. En ce qui concerne *getInseeUrl($.inseeInfo)*, la fonction crée une requête SPARQL à partir des valeurs en paramètre, l'envoie au endpoint SPARQL, traite le résultat et retourne l'URI de l'unité administrative qui correspond à ces valeurs. Voici la requête SPARQL générée par la fonction *getInseeUrl()* pour cet exemple :

```
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
```

```
PREFIX igeo:<http://rdf.insee.fr/def/geo#>
SELECT ?adminUnit  WHERE {
?adminUnit rdf:type igeo:Region .
?adminUnit igeo:codeINSEE "54"^^<http://www.w3.org/2001/XMLSchema#token> .}
```

Le graphe RDF résultant est fourni dans cet extrait :

```
@prefix geo: <http://www.opengis.net/ont/geosparql#> .
@prefix admin: <http://melodi.irit.fr/ontologies/administrativeUnits.owl#> .
@prefix l_admin: <http://melodi.irit.fr/lod/administrativeUnit/> .
l_admin:region_54 a admin:Region .
l_admin:region_54 admin:hasInseeCode "54"^^xsd:String .
l_admin:region_54 admin:hasName "Poitou-Charentes"^^xsd:String .
l_admin:region_54 owl:sameAs <http://id.insee.fr/geo/region/54> .
l_admin:region_54 a geo:Feature .
l_admin:region_54 geo:hasGeometry l_admin:region_54_geo .
l_admin:region_54_geo geo:asWKT "MULTIPOLYGON(((
-1.0988062299633785 45.64032288975508, ...
-1.0988062299633785 45.64032288975508))...)"^^wkt:Literal .
```

Les observations météorologiques. Le caractère temporel des données météorologiques a une importance particulière. Les observations contenues dans l'entrepôt SYNOP ont plusieurs temporalités. Les observations codées `tminsol` représentent la plus petite température relevée durant les 12 dernières heures, alors que celles codées `t` correspondent à la température relevée au moment de la mesure. Selon notre approche il suffit d'implémenter des fonctions pour traiter la diversité temporelle de ces données. L'extrait de code JSON suivant représente une observation de type `tminsol` relevée par la station météo `07747` le `03/06/2017 à 3h`.

```
{    "temporalInfo" :
    {   "timeStamp" : 1512529200,
        "month" : "12",
        "day" : "06",
        "hour" : "03",
        "year" : "2017" },
    "tminsol" : 271.45,
    "numer_sta" : "07747" }
```

Pour traiter des observations SYNOP de Météo France nous avons écrit le template suivant qui appelle la fonction *getMFO_PhenomenonTime(doc)* :

```
<dummy> sosa:phenomenonTime getMFO_PhenomenonTime(doc) .
```

La fonction scanne le document JSON, trouve le type de l'observation à partir de sa clé (`tminsol`), et crée une instance de la classe `time:Interval` (spécialisation de `time:TemporalEntity`). Elle examine ensuite l'élément `temporalInfo` et calcule le début de l'intervalle de temps (la fin étant fournie par la clé `temporalInfo` elle-même). Début et fin de l'intervalle sont représentés comme des instances de `time:Instant`. Pour l'exemple précédent, le résultat est le suivant :

```
gmfo:Obs_07747_20171206030000_tminsol sosa:phenomenonTime
```

```
                        gmfo:TimeInterval_1512486000_1512529200 .
gmfo:TimeInterval_1512486000_1512529200 a time:TemporalEntity .
gmfo:TimeInterval_1512486000_1512529200 time:hasBeginning
                        gmfo:TimeInterval_1512486000_1512529200_beginning .
gmfo:TimeInterval_1512486000_1512529200_beginning time:inXSDDateTime
                        "2017-12-05T15:00:00+0100"^^xsd:dateTime .
gmfo:TimeInterval_1512486000_1512529200 time:hasEnd
                        gmfo:TimeInterval_1512486000_1512529200_end .
gmfo:TimeInterval_1512486000_1512529200_end time:inXSDDateTime
                        "2017-12-06T03:00:00+0100"^^xsd:dateTime .
```

La couverture terrestre. Pour intégrer cette source de données, nous avons dont l'API REST prend en entrée un polygone WKT en SRS EPSG:4326. A partir de ce polygone, il récupère les données originales dans un fichier temporaire, puis crée une table des fréquences. La réponse du serveur est un document JSON contenant le pourcentage de chaque classe de la couverture terrestre pour la zone délimitée par le polygone. Ce document JSON est ensuite transformé en RDF.

Les méta-données d'images satellites. Ces métadonnées sont obtenues sous forme de fichiers GeoJSON et transformée en RDF à l'aide d'un template particulier, selon le même principe en utilisant le vocabulaire `eom` (Section 4).

Les tuiles d'images. Les tuiles sont fournies dans le fichier `grid` de l'ESA que nous transformons en JSON, puis en RDF en utilisant le vocabulaire `grid` décrit dans la Section 4. Il est possible de calculer les relations topologiques entre les éléments spatiaux et les tuiles, puis d'extrapoler ces informations pour les images. Par exemple, sachant que les images $[img1, img2, img3]$ partagent la tuile $tile_1$, et que cette tuile recouvre $adminUnit_i$, il est possible d'inférer que $[img1, img2, img3]$ recouvrent aussi $adminUnit_i$.

5.3. Intégration des données

Intégration des données ayant une composante spatiale fixe. Les relations spatiales sont relativement stables dans le temps. Ainsi, les relations topologiques entre les grilles (SS2) et les unités administratives (l'image Y recouvre la région R), ou les informations sur la couverture terrestre associée à chaque cellule de la grille, sont calculées une fois pour toutes et mémorisées dans l'entrepôt RDF. Nous avons développé un script Python pour calculer les relations topologiques entre les instances des classes qui utilise la librairie *shapely* pour comparer les surfaces. Nous enregistrons ensuite les relations dans l'entrepôt sous forme de triplets dans lesquels le prédicat est une propriété topologique de GeoSPARQL.

Intégration des données ayant une composante temporelle. La mise en relation temporelle d'un enregistrement de méta-données d'image et de données ayant une propriété temporelle (date ou intervalle) prend en compte la période de temps qui intéresse l'utilisateur. Par exemple, il peut vouloir associer à une image à des informations météorologiques relevées une semaine après la prise de l'image. L'intervalle de temps défini par l'utilisateur joue le rôle de buffer temporel fournissant un contexte aux enregistrements de méta-données.

6. Conclusion

L'intégration de données d'observations de la Terre provenant de sources hétérogènes avec des métadonnées d'images satellites peut tirer profit des technologies du web sémantique. En publiant des ensembles de données et des méta-données d'images sous forme de LOD, l'accès aux observations de la Terre liées se trouve facilité, ce qui offre de nouvelles possibilités pour utiliser les images satellites dans une plus grande variété d'applications. De plus, pour les ensembles de données volumineux et dynamiques, en utilisant des requêtes SPARQL pour interroger conjointement des bases de données d'observations et des données liées, il est possible de créer des triplets RDF à la volée et ainsi éviter de convertir d'énormes ensembles de données en triplets RDF. Dans cet article, nous avons proposé un cadre pour intégrer des données spatiales. Nous avons conçu un vocabulaire pour représenter les données d'observations de la Terre et les méta-données d'images ; nous avons élaboré un processus de conversion RDF qui utilise des templates adaptés aux ressources et une librairie Python pour dépasser certaines limites de RML ; nous avons aussi proposé un processus d'intégration qui exploite la géométrie des données et GeoSPARQL pour lier les données géospatiales, et au final des requêtes SPARQL pour lier dynamiquement les données aux images à partir de leurs caractéristiques spatiales et temporelles. Dans la continuité de ces travaux, nous envisageons de considérer des sources propres à un domaine métier pour traiter un cas d'usage particulier (l'agriculture et des rapports bulletins agricoles) et fournir des règles et des fonctionnalités de raisonnement pour faciliter les analyses.

Bibliographie

Alirezaie M., Kiselev A., Längkvist M., Klügl F., Loutfi A. (2017). An ontology-based reasoning framework for querying satellite images for disaster monitoring. *Sensors*, vol. 17, n° 11.

Arenas H., Aussenac-Gilles N., Comparot C., Trojahn C. (2016). Semantic integration of geospatial data from earth observations. In *Knowledge engineering and knowledge management - EKAW 2016 satellite events*, p. 97–100. Bologna (It), Springer.

Atemezing G. A. (2015). *Publishing and consuming geo-spatial and government data on the semantic web*. Thèse de doctorat non publiée, Thesis. Consulté sur http://www.eurecom.fr/publication/4545

Auer S., Lehmann J., Ngonga Ngomo A.-C. (2011). Introduction to linked data and its lifecycle on the web. In *Reasoning Web. Semantic Technologies for the Web of Data: 7th International Summer School 2011, Ireland, August 23-27, 2011, Tutorial Lectures*, p. 1–75.

Battle R., Kolas D. (2012). Enabling the Geospatial Semantic Web with Parliament and GeoSPARQL. *Semantic Web*, vol. 3, n° October 2012, p. 355–370.

Blázquez L. M. V., Saquicela V., Corcho Ó. (2012). Interlinking geospatial information in the web of data. In *Bridging the geographic information sciences - international agile'2012 conference*, p. 119–139. Avignon, France.

Blázquez L. M. V., Villazón-Terrazas B., Corcho Ó., Gómez-Pérez A. (2014). Integrating geographical information in the linked digital earth. *International Journal of Digital Earth*, vol. 7, n° 7, p. 554–575.

Brizhinev D., Toyer S., Taylor K., Zhang Z. (2017). *Publishing and using earth observation data with the rdf data cube and the discrete global grid system*. Rapport technique. W3C and OGC.

García-Rojas A., Athanasiou S., Lehmann J., Hladky D. (2013). Geoknow: Leveraging geospatial data in the web of data. In *Open data on the web.* Campus London, Shoreditch.

Gasperi J. (2014). Semantic Search Within Earth Observation Products Database Based on Automatic Tagging of Image Content. In *Proc. of the Conf. on Big Data from Space*, p. 4–6. ESA/ESRIN, Frascati, Italy., EU Publications.

Georgala K., Sherif M. A., Ngomo A. N. (2016). An efficient approach for the generation of allen relations. In *Proceedings of the 22nd European Conference on Artificial Intelligence (ECAI), Including Prestigious, Applications of Artificial Intelligence (PAIS)*, p. 948–956.

Heath T., Bizer C. (2011). *Linked data: Evolving the web into a global data space ; lectures on the semantic web: Theory and technology*. Morgan & Claypool.

Kolas D., Perry M., Herring J. (2013). *Getting started with GeoSPARQL*. Rapport technique. OGC. Consulté sur http://www.ssec.wisc.edu/meetings/geosp_sem/presentations/GeoSPARQL_Getting_Started-KolasWorkshopVersion.pdf

Koubarakis M., Karpathiotakis M., Kyzirakos K., Nikolaou C., Vassos S., Garbis G. *et al.* (2012). Building virtual earth observatories using ontologies and linked geospatial data. In *Web Reasoning and Rule Systems: 6th Int. Conf. RR, Vienna, Austria*, p. 229–233.

Pomp A., Paulus A., Jeschke S., Meisen T. (2017). ESKAPE: Platform for Enabling Semantics in the Continuously Evolving Internet of Things. In *2017 IEEE 11th International Conference on Semantic Computing*, p. 262-263.

Reitsma F., Albrecht J. (2005). Modeling with the semantic web in the geosciences. *IEEE Intelligent Systems*, vol. 20, n° 2, p. 86-88.

Shen W., Wang J., Han J. (2015). Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, n° 2, p. 443-460.

Sherif M. A., Dreßler K., Smeros P., Ngomo A. N. (2017). Radon - rapid discovery of topological relations. In *Proceedings of the thirty-first AAAI conference on artificial intelligence, feb. 4-9, 2017, san francisco, cal., USA.*, p. 175–181.

Smeros P., Koubarakis M. (2016). Discovering spatial and temporal links among RDF data. In *Proceedings of the workshop on linked data on the web, LDOW 2016, co-located with 25th international world wide web conference (WWW 2016).*

Sukhobok D., Sánchez H., Estrada J., Roman D. (2017). Linked data for common agriculture policy: Enabling semantic querying over sentinel-2 and lidar data. In *Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks.*

Tandy J., Brink L. van den, Barnaghi P. (2017). *Spatial data on the web best practices, w3c working group note*. Rapport technique. W3C and OGC. Consulté sur https://www.w3.org/TR/sdw-bp/