Représentation sémantique de données géospaciales au service de l'analyse de changements

Jordane Dorne

THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 2 - Jean Jaurès

Présentée et soutenue par

Jordane DORNE

Le 29 octobre 2021

Représentation sémantique de données géospatiales au service de l'analyse de changements

Ecole doctorale : **EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse**

Spécialité : **Informatique et Télécommunications**

Unité de recherche :
IRIT : Institut de Recherche en Informatique de Toulouse

Thèse dirigée par
Nathalie AUSSENAC-GILLES

Jury

Mme Marlène VILLANOVA-OLIVER, Rapporteure
M. Christian SALLABERRY, Rapporteur
Mme Cassia TROJAHN DOS SANTOS, Examinatrice
Mme Catherine COMPAROT, Examinatrice
Mme Thérèse LIBOUREL, Examinatrice
M. Christophe CRUZ, Examinateur
Mme Nathalie AUSSENAC-GILLES, Directrice de thèse
M. Franck RAVAT, Président

Remerciements

Je souhaiterais remercier tout d'abord Marlène Villanova-Oliver et Christian Sallaberry pour avoir accepté d'examiner ce mémoire en tant que rapporteurs ainsi que pour leurs remarques constructives. Je remercie également Thérèse Libourel, Christophe Cruz et Franck Ravat d'avoir accepté de participer à mon jury de thèse.

Je remercie Nathalie Aussenac-Gilles, Catherine Comparot, Cassia Trojahn et Romain Hugues pour leur encadrement et leur bienveillance tout au long de cette thèse.

Je souhaite également remercier Jean-Guy Planès, Xavier Olive et Matthias Renard de m'avoir donné l'opportunité de réaliser cette thèse au sein de l'entreprise Thales Alenia Space. Je remercie également Hugues Sassier de m'avoir transmis sa passion et ses connaissances dans le domaine de la télédétection.

Enfin je remercie ma famille qui m'a toujours soutenu ainsi que Claire qui est à l'origine de nombreux succès dans ma vie.

Résumé : Grâce à des modèles d'apprentissage de plus en plus efficaces, les images satellitaires d'observation de la Terre (images EO) peuvent être plus facilement comparées afin que les changements entre des séries de deux images ou plus puissent être calculés. Les changements sont des indicateurs qui peuvent révéler des événements naturels ou des catastrophes telles que des incendies, des inondations ou des tremblements de terre ; ou l'évolution à long terme de l'occupation des sols comme la déforestation, l'urbanisation, etc. Les services évaluant les changements géographiques au fil du temps nécessitent généralement de rechercher, de découvrir et d'identifier manuellement les données pertinentes. Il s'agit d'une tâche laborieuse qui repose sur la connaissance du domaine et qui est souvent sujette à l'erreur humaine. L'identification automatique des changements grâce au traitement des images EO réduit cet effort. La plupart des premiers travaux reposaient sur des algorithmes d'apprentissage supervisés, les plus récents ne sont pas supervisés et permettent d'éviter l'annotation manuelle des échantillons . Ces détections de changements comparent des séries ou des paires d'images dans lesquelles elles identifient les changements à représenter au niveau des pixels dans des fichiers raster. Cependant, les résultats de ce processus manquent de contexte et le traitement de fichiers raster nécessite des outils spécifiques. Sans plus d'informations, un changement détecté ne peut pas aider les experts du domaine à comprendre le phénomène et prendre la décision appropriée. Les algorithmes de détection de changement sont efficaces pour détecter les changements et leur attribuer un degré d'importance mais ils n'identifient pas l'événement qui a causé ce changement. Pour ce faire, des données contextuelles sont nécessaires. Divers types de données géospatiales peuvent fournir un contexte pour identifier ces changements et mettre en évidence des phénomènes spécifiques : soit des indices calculés directement à partir d'images EO, des données liées ouvertes et des données sociales. Ce processus nécessite l'intégration de données provenant de sources diverses et de nature différente, ce qui peut être fait efficacement grâce aux technologies du Web sémantique. L'objectif de cette thèse est de proposer un processus générique afin que i) tout type de données puisse être ingéré comme connaissance contextuelle sémantique en fonction du type d'événements à identifier ; ii) il puisse prendre en entrée tout type de raster de changement quel que soit l'algorithme qui l'a généré ; iii) il puisse gérer la géométrie des événements avec une forme étant un compromis entre précision et simplicité. La tâche que nous abordons ici peut être formulée comme telle : étant donné un raster de changement, nous voulons construire un graphe de connaissances contenant diverses données contextuelles pour la description d'un événement identifiable sur cette région à la période considérée. Une difficulté que nous abordons ici est de regrouper de manière optimale les pixels pour définir les régions pertinentes, c'est-à-dire les polygones représentant les régions d'intérêt (ROI) et de trouver la division géographique précise qui guidera l'intégration des données. La notion de ROI est pratique pour être utilisée comme référence géographique pour l'intégration de données. Les contributions principales de cette thèse sont i) un algorithme pour identifier les ROI dans un raster ; ii) un

processus sémantique et un vocabulaire pour générer un graphe de connaissances à partir de différentes sources afin d'aider à qualifier les changements identifiés dans les rasters de changement ; iii) une validation de l'approche avec différents cas d'utilisation sur le suivi des changements à différentes granularités temporelles et échelles de restitution (changement du NDVI à l'aide de tuiles, changement d'occupation du sol à l'aide de polygones et suivi des incendies à partir des ROI).

Mots clés : Sémantique, Géospatial, Télédétection

Abstract : Thanks to efficient deep learning models, Earth Observation satellite images (EO images) can be more easily compared so that changes between series of two or more images can be computed. Changes are indicators that may reveal natural events or disasters such as fires, floods, or earthquakes ; or long term evolution of ground occupation like deforestation, urbanisation, etc. Services assessing geographic changes over time typically require searching, discovering, and manually identifying relevant data. This is a difficult and laborious task that relies on domain knowledge and that is often subject to human error. Automatically identifying changes thanks to EO satellite image processing reduces this effort. Whereas most early works relied on supervised ML algorithms, more recent ones are unsupervised to avoid manual tagging of examples. These automatic change detection methods compare series or pairs of images in which they identify changes to be represented at pixel level in raster files. However, results of this process lack context and reading raster files requires specific tools. Without more information, a detected change cannot help domain experts analyzing those images, understanding the phenomenon, and taking the appropriate decision. Generic change detection algorithms are efficient to detect changes, to assign them a degree of importance but they are not able to identify the nature of the event that caused the change. To do so, more contextual data is needed. Various kinds of geospatial data can provide context to detect changes and highlight specific kinds of phenomenon : either indices (e.g., vegetation, hydrography or landscape indices) directly computed from EO images ; or linked data and social data used as rich complementary sources of information. This process requires the integration of data from various sources and of different nature, which can be efficiently done thanks to semantic web technologies. The aim of this thesis is to propose a generic process so that i) given the type of events to be identified from the change impact of EO images, any kind of dataset could be selected and ingested as semantic contextual knowledge, ii) it could take as input any change raster whatever the algorithm that generated it ; iii) it could manage the geometry of events with a shape that could reach a compromise between precision (to know the event exact location) and simplicity (to make it easy to compare with the geometry of other geographic objects). The task addressed here can be formulated as such : given the results of a change detection algorithms as rasters on an area of interest, to build a knowledge graph (KG) depicting various contextual data related to each event of a specific kind identifiable on that region at the considered period. One difficulty is to optimally group pixels as for defining relevant regions, i.e., polygons representing Regions Of Interest (ROIs) and to find the accurate geographic division that will guide data integration. Similar to the notion of tiling grid defined by ESA, that divides the Earth surface into tiles representing a fixed area on this surface, the notion of ROI is convenient to be used as a geolocated reference for data integration. Unlike tiles, not all ROIs have the same size or shape. The main contributions of the thesis are i) an algorithm to identify ROIs on a raster ; ii) a semantic-driven process and a vocabulary to generate a KG

from different sources in order to help explaining the changes identified in change rasters ; iii) a validation of the approach with different use-cases on change monitoring at different temporal granularity, different scales of restitution and different contextual data (NDVI change using tiles, land cover change using polygons, and fire monitoring from ROIs).

Keywords : Semantic, Geospatial, Earth Observation

Table des matières

Chapitre 1

Introduction

1.1 Des images satellitaires à l'étude des changements

Le volume, le flux et la diversité de données géospatiales offrent des perspectives diverses pour les utilisateurs de différents secteurs d'activité tels que la gestion des ressources, l'environnement, le changement climatique ou la gestion des risques. Ces données correspondent, par exemple, aux données de stations météorologiques, aux images de vidéo surveillance, aux données de capteurs connectés ou encore aux traces collectées par Global Positioning System (GPS), téléphonie mobile, mais aussi aux données échangées sur les réseaux sociaux, qui sont de plus en plus géolocalisées. L'une des problématiques majeures liées à la grande masse de données géospatiales disponibles est d'en extraire de l'information pertinente et adaptée au besoin de chacun de ces usages qui correspondent à des attentes très différentes, pour en faire des connaissances et aider dans le processus de prise de décision.

Cette problématique s'articule à l'intersection des domaines de la géomatique et du Web sémantique. La géomatique [Langlois 2004] s'intéresse à la collecte, la gestion, l'analyse ou la représentation numérique de données géolocalisées à la surface ou autour de la Terre. Le Web sémantique [Berners-Lee *et al.* 2001], quant à lui, repose sur l'idée d'exposer les données sur le Web avec des annotations de manière à ce qu'elles puissent être mieux exploitées par les machines et par les humains pour le traitement, l'intégration et la réutilisation dans diverses applications. Cette idée repose essentiellement sur la notion d'*ontologie*, une structure de représentation des connaissances qui permet d'organiser les concepts d'un domaine d'intérêt, les relations entre ces concepts, et d'exprimer leur sémantique dans un contexte donné. Les technologies du Web sémantique peuvent améliorer l'accès, la gestion, la recherche et l'analyse de données géographiques, ce qui explique l'attention particulière qu'elles ont reçues ces dernières années [Egenhofer 2002, Reitsma & Albrecht 2005, Janowicz *et al.* 2012].

Depuis peu d'années, la disponibilité des images issues des nouveaux flux provenant des satellites placés en orbite polaire ont largement contribué à l'évolution du domaine de la géomatique, en particulier. Dans le cadre du programme européen Copernicus, qui offre des services d'information basés sur l'observation de la Terre par satellite et sur les données in situ (non spatiales), les satellites d'observation de la Terre Landsat 8 et les programmes Sentinel de l'European Space Agency (ESA) produisent en continu des images de la Terre captées selon différentes technologies et accessibles à large échelle. Ces satellites sont censés transmettre quotidiennement des images de haute qualité, soit un peu plus de 15 To de données. Ces données

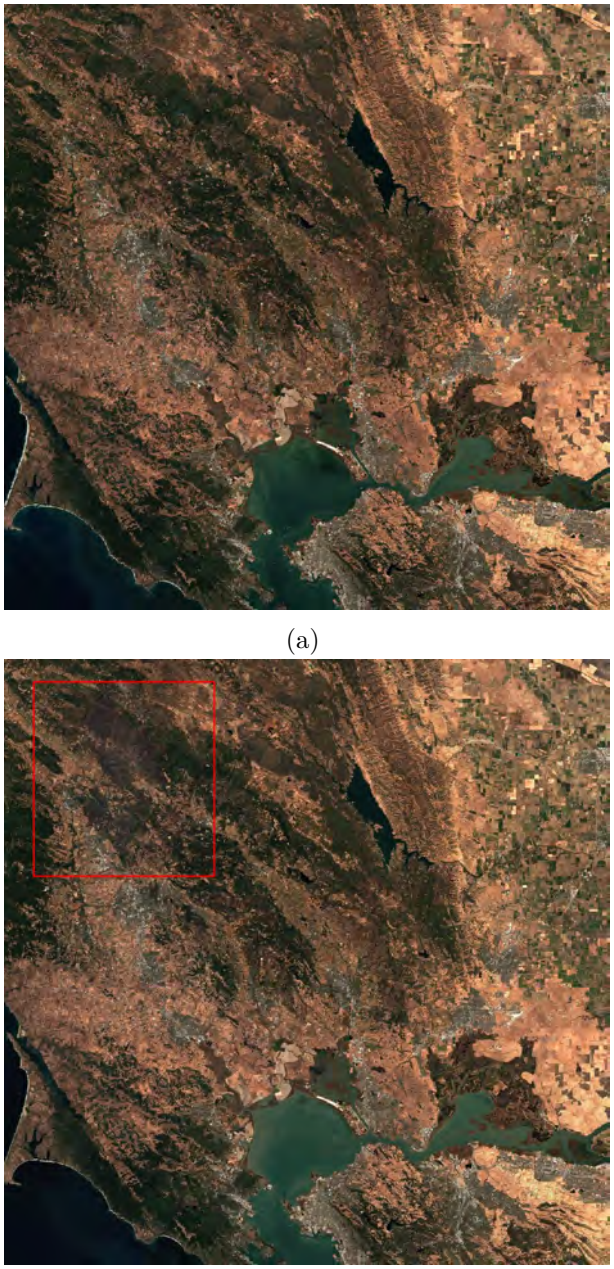
sont toutes en accès libre, levant ainsi la restriction financière liée à l'utilisation systématique des images satellitaires des anciens programmes comme SPOT.

Les applications liées à ces satellites sont nombreuses et varient de l'étude des forêts au suivi de l'agriculture en passant par la planification urbaine [Haas & Ban 2017, Behera *et al.* 2021]. Les images d'observation de la Terre ainsi produites permettent également d'étudier les catastrophes et leur impact à large échelle, qu'elles soient naturelles ou causées par l'homme. En particulier, l'étude de certains phénomènes à la surface de la Terre revient à rechercher des changements entre deux observations ou au sein d'une série d'observations d'une zone terrestre. Dans le cas d'événements ponctuels (incendie, inondations, tempête, etc.), il s'agit de les repérer en observant leur impact, ce qui revient à comparer la zone impactée avant et après une date. Dans le cas de changements sur le long terme (urbanisation, sécheresse, déforestation ...), les observations doivent être répétées et s'étaler dans le temps. Les images satellitaires peuvent apporter des solutions en offrant les observations de la Terre requises, à condition d'être captées suffisamment fréquemment et systématiquement.

Il convient de différencier deux types de changements : les changements sur la surface de la Terre et les changements dans l'imagerie [Théau 2008]. Les *changements sur la surface de la Terre* peuvent être dus à des événements variés comme des inondations ou des constructions. Ceux-ci ont des dimensions spatiales et temporelles variées. Les *changements dans l'imagerie* entre deux dates correspondent à un changement dans la radiance des images. L'hypothèse principale pour l'utilisation d'images pour la détection de changements sur la surface de la Terre est qu'un changement sur la surface de la Terre induit un changement sur les valeurs de radiance des images. Nous distinguons également la notion de *changement* de la notion d'*évolution* selon leur temporalité. Le changement est ici caractérisé par une période finie avec un début et une fin en opposition à l'évolution qui ne possède pas de limite temporelle finie.

À titre d'exemple de changement sur le court terme, considérons les Figures 1.1(a) et 1.1(b), qui représentent des images de la même zone, située au nord de la baie de San Francisco en Californie en 2019, 22 octobre et 1er novembre 2019. Cette zone a été touchée par un incendie important au cours de cette période, qui a brûlé la végétation à la surface de la Terre. Ce changement à la surface de la Terre peut à peine être détecté par l'œil humain en comparant les deux images, mais la radiance des deux images est bien différente sur la zone impactée. C'est pourquoi l'imagerie, et en particulier la détection des changements sur les images, peut être un moyen de vérifier les impacts précis des incendies après qu'ils se soient produits, ou de les suivre pendant qu'ils sont encore actifs, y compris dans des endroits difficiles d'accès.

L'étude de changements à partir d'images d'observation de la Terre est l'objet de recherches depuis des nombreuses années. Ce domaine a fait des progrès considérables grâce aux nouvelles sources de données d'images satellitaires couplées au développement d'algorithmes d'apprentissage automatique [Asokan & Anitha 2019, Radke *et al.* 2005]. Ces algorithmes sont capables de traiter différentes résolu-



(a)

(b)

FIGURE 1.1 – Images Sentinel-2 de la Californie utilisées comme données d'entrée ((a) 22/10/2019 - (b) 01/11/2019). Le rectangle rouge indique une zone touchée par l'incendie après la première image.

tions d'image [Amin *et al.* 2016, Gong *et al.* 2017] et d'appliquer différents niveaux d'optimisation [Amin *et al.* 2016, Jia *et al.* 2014]. Alors que la plupart des premiers travaux reposaient sur des algorithmes d'apprentissage supervisés, les plus récents ne sont pas supervisés pour éviter l'annotation manuelle des exemples

[de Jong & Bosman 2019]. En général, la détection de changement s'appuie sur le calcul de la différence au niveau de chaque pixel des images comparées, différence qui est estimée par une probabilité de changement plus ou moins forte. Les algorithmes produisent une carte de changement de raster en sortie, où chaque valeur de pixel (numérique) évalue le degré de changement. Ils permettent de faciliter les taches de l'opérateur qui peut identifier directement la zone de l'image impactée par un changement anormal.

Les algorithmes de détection de changements sont efficaces pour détecter les changements, pour leur attribuer un degré d'importance mais ils n'identifient pas la nature de l'événement qui a causé le changement. Sans informations supplémentaires ou contexte, il est presque impossible de connaître la nature des événements qui ont causé les changements. Dans le cas de l'exemple ci-dessus, l'algorithme de détection de changement génère un fichier raster où les zones mises en évidence sur la deuxième image sont évaluées comme fortement modifiées entre les captures de l'image 1 et de l'image 2. Mais sans aucune information supplémentaire, il est difficile de fournir une explication à ce changement. Un humain peut savoir que la raison réelle de ce changement est un incendie grâce à des informations contextuelles (contact personnel, informations, tweets ou toutes données de réseaux sociaux, cartes d'incendie ou encore des indices calculés à partir d'images). Fournir des connaissances contextuelles pour expliquer un changement détecté aide les experts du domaine à analyser ces images, à comprendre le phénomène et à prendre la décision appropriée. De retour à l'exemple, les informations pertinentes sur les changements peuvent être les unités administratives, afin que l'on sache dans quel endroit (ville et/ou département) l'incendie a eu lieu, la couverture terrestre pour connaître le type de végétation affectée ou de zone habitée, des tweets ou des informations qui pourraient confirmer que l'événement à l'origine du changement est un incendie, et la mesure de températures anormalement élevées sur la surface de la Terre relevées par des satellites météorologiques.

1.2 Problématique

La problématique majeure de cette thèse est l'association de données contextuelles à des données de changements identifiés sur des images satellitaires. Un *premier défi* est l'agrégation spatiale des données de changement, afin de repérer des zones touchées par un changement suffisamment significatif pour qu'il puisse être interprété comme un événement. Cette agrégation spatiale doit prendre en compte différents découpages territoriaux et niveaux de granularités selon le type d'évènement à traiter.

Un *deuxième défi* de cette problématique est de pouvoir qualifier chaque évènement qui a donné lieu à un changement identifié dans l'image. Cela implique que a) les sources de données contextuelles choisies soient pertinentes par rapport aux caractéristiques spatiales et temporelles des images d'observation en question ; et que b) l'intégration sémantique de données hétérogènes soit guidée par une repré-

sentation spatiale et temporelle des données de changement afin de leur associer des données contextuelles qui leur donnent du sens.

Les questions de recherches suivantes seront traitées dans cette thèse :

Question 1 Comment agréger des données calculées à l'échelle du pixel sur des images pour en faire des zones homogènes ? Comment procéder lorsqu'il s'agit de zones identifiées par l'utilisateur ? de grilles correspondant à des découpages prédéfinis comme les tuiles de l'ESA ? de zones d'intérêt trouvées dans des jeux de données telles que des villes ou des unités territoriales ? ou enfin des régions d'intérêt calculées automatiquement ?

Question 2 Comment intégrer et comparer des données géographiques et contextuelles dont les dimensions spatiales et temporelles sont exprimées à différentes échelles ?

Question 3 Comment les données contextuelles peuvent-elles aider à expliquer un changement ou une série de changements détectés grâce au traitement d'image reposant sur l'apprentissage automatique ? Quelles données sont pertinentes selon le phénomène à étudier et le découpage territorial choisi ?

FIGURE 1.2 – Objectif visé pour l'annotation sémantique de changements

La figure 1.2 montre l'objectif souhaité pour l'annotation sémantique de changements. D'une part, il s'agit de délimiter sur l'image la zone au sol qui a subi un changement. Ensuite, le résultat de l'annotation revient à associer des informations à cette zone, à savoir la nature et éventuellement le nom de l'événement qui s'est

déroulé, sa date, sa localisation ainsi que des mots-clés associés. Ce résultat fournit une représentation formalisée d'un changement détecté à partir des images satellitaires et une première qualification de ce changement. Cette représentation peut être exploitée par des traitements informatiques et par des experts en télédétection afin de les aider dans le repérage de changements et leur qualification précise. Cette étape est aujourd'hui réalisée manuellement et elle s'avère très chronophage. Or elle est très utile pour des applications de gestion des risques par exemple, et elle est indispensable pour retrouver l'image représentant l'évènement au sein d'un catalogue.

1.3 Contributions

Pour répondre aux problématiques énoncées ci-dessus, cette thèse propose un processus afin que, étant donné un type d'événements à identifier à partir de l'étude du changement entre deux images satellitaires, (i) tout type de jeu de données puisse être sélectionné et intégré en tant que connaissance contextuelle sémantique, (ii) tout type de raster de changements, quel que soit l'algorithme qui l'a généré, puisse être traité ; (iii) la géométrie des événements puisse être estimée de différentes manières, en proposant une forme "optimale" de cette géométrie correspondant à un compromis entre précision (pour connaître l'emplacement exact de l'événement) et simplicité (pour faciliter la comparaison de cette géométrie à celles d'autres objets géographiques).

Les contributions de cette thèse sont les suivantes :

Algorithme de ROI : un algorithme pour regrouper de manière optimale les pixels pour définir les régions pertinentes, c'est-à-dire les polygones représentant les régions d'intérêt ou Region Of Interest (ROI) et de trouver la division géographique précise qui guidera l'intégration des données. Cette technique permet notamment de générer uniquement de la connaissance sur une zone géographique d'intérêt.

Ontologies de changement : des ontologies modulaires pour l'association de données contextuelles de nature diverse (indices calculés à partir d'images, tels que le Normalized Difference Vegetation Index (NDVI) ou la couverture terrestre, des données ouvertes des réseaux sociaux ou encore issues d'autres capteurs) à différents découpages territoriaux (tuile, polygones représentant des unités administratives, ou zones de changement ROI).

Intégration sémantique de données : un processus d'intégration de données sémantique, guidé par les ontologies proposées, pour générer un graphe de connaissances à partir de différentes sources de données contextuelles afin d'aider à l'analyse et à la qualification des changements. Cette chaîne de traitement est composée d'un ensemble de modules logiciels ayant chacun pour fonction de générer ou enrichir un graphe de connaissances à partir de différents types de données. Cette chaîne permet d'obtenir une représentation sémantique des changements annotés par des données contextuelles. Cette

représentation pourra par la suite être exploitée dans un autre système ou bien être visualisée directement grâce à des outils spécialisés.

Validation sur différents cas d'usage : une validation de l'approche avec différents cas d'utilisation sur le suivi des changements à différentes granularités temporelles et échelles de restitution, en utilisant diverses sources de données contextuelles : changement de l'indice de couvert végétal NDVI à l'échelle de tuiles, changement d'occupation du sol sur des polygones représentant des unités administratives, et suivi des incendies à partir de zones d'intérêt (ROI).

Les travaux de cette thèse s'inscrivent dans la convention Cifre ANR n°2017/1399 entre Thales Alenia Space et le Centre National de la Recherche Scientifique (CNRS). Thales Alenia Space est concerné par la valorisation des images de la collection Sentinel 2. Une des pistes étudiées pour cette valorisation correspond à l'étude de changements. Pour cela, Thales Alenia Space a développé plusieurs algorithmes d'apprentissage automatique non supervisé qui calculent les changements entre deux images d'observation de la Terre. Les résultats obtenus sont disponibles sous forme de fichier image (format raster). Ces fichiers constituent un des points d'entrée de cette thèse, et ont servi en particulier aux différentes expérimentations validant l'approche proposée. La faisabilité et la pertinence de l'approche retenue en milieu industriel étaient des critères complémentaires qui ont orienté les solutions envisagées.

1.4 Plan du manuscrit

Cette thèse est organisée selon les chapitres suivants :

Chapitre 2 introduit les notions nécessaires pour le reste de la thèse, tirées des deux domaines auxquels nous apportons des contributions : l'observation de la Terre d'une part, le Web sémantique d'autre part.

Chapitre 3 discute les principaux travaux similaires sur la détection de changements à partir d'images satellitaires, sur la représentation sémantique du temps et de l'espace, et sur l'étude du changement à l'aide des approches du Web sémantique. Il nous permet de positionner notre travail.

Chapitre 4 détaille les propositions de la thèse, à savoir un algorithme d'agrégation de changements pour définir des régions d'intérêt, des modèles sémantiques adaptés pour leur représentation, la qualification des changements selon différentes granularités temporelles et échelles de restitution ; et enfin, le processus d'intégration sémantique de données contextuelles issues de différentes sources guidé par ces représentations.

Chapitre 5 présente les expérimentations menées et leurs résultats afin d'évaluer l'approche sur différents cas avec, pour chacun, des régions d'agrégation des changements différentes : changement du couvert végétal via l'indice NDVI à l'aide de tuiles, changement de l'occupation du sol à l'aide de polygones et suivi des incendies à partir des régions d'intérêt (ROI).

Chapitre 6 discute les limites et les perspectives des contributions de la thèse.

Chapitre 2

Des images satellitaires à l'étude des changements sur la Terre à l'aide du Web sémantique

Content

Ce chapitre vise à fournir certaines notions nécessaires pour la compréhension technique de cette thèse, tirées des deux domaines auxquels nous apportons des contributions : l'observation de la Terre d'une part, le Web sémantique d'autre part. Il se compose de deux parties : la première présente quelques notions clés de la géomatique ainsi que des références historiques sur l'évolution de ce domaine. Cette partie a pour but d'expliquer le rôle de la télédétection ainsi que les détails techniques en rapport avec l'exploitation d'images satellitaires. La deuxième partie du chapitre introduit les notions importantes liées au Web sémantique. Cette partie présente, dans un premier temps, l'évolution du Web sémantique ainsi que les principaux acteurs et, dans un second temps, les technologies, modèles et langages standardisés permettant la représentation de connaissances associées à des éléments du Web.

2.1 Observation de la Terre et géomatique

Comme nous l'avons dit en introduction, la géomatique s'intéresse à la collecte, la gestion, l'analyse ou la représentation numérique de données géolocalisées

à la surface ou autour de la Terre. Ce domaine a débouché sur un foisonnement de propositions hétérogènes. Il s'est progressivement structuré, avec l'apparition d'organismes de standardisation d'une part (partie 2.1.1), et de formats pour la représentation d'information spatialisée (partie 2.1.3.1), pour déclarer les coordonnées (partie 2.1.2) ou encore pour représenter les informations de localisation d'objets spatialisés 2.1.3.

2.1.1 Les organismes de standardisation des données géospatiales

Dans cette section seront abordés les standards existants pour la représentation des données géospatiales. Les données rasters et vectorielles représentent la quasi totalité des données exploitées dans la géomatique. La mission de l'Open Geospatial Consortium (OGC) est de définir des standards ouverts afin d'harmoniser les formats et services existants pour ces données et ainsi de faciliter leur exploitation. L'OGC a été fondée en 1994 afin de garantir l'interopérabilité dans les outils de type Système d'Information Géographique (SIG) et compte aujourd'hui plus de 500 acteurs. Ce consortium est à l'origine de nombreux standards ouverts pour décrire des objets géographiques ainsi que des services. Parmi ceux-ci, le Web Processing Service (WPS) est très largement utilisé dans les catalogues d'images satellitaires par exemple afin d'exécuter des traitements sur des serveurs distants. L'OGC possède également l'objectif de rendre les données géospatiales Findable, Accessible, Interoperable, and Reusable (FAIR)[Wilkinson & et al 2016].

En partenariat avec l'OGC, l'Open Source Geospatial Foundation (OSGeo) développe des logiciels libres pour la communauté de la géomatique. L'OSGeo [1] est une Organisation Non Gouvernementale (ONG) fondée en 2006 qui organise chaque année les conférences Free and Open Source Software for Geospatial (FOSS4G) [2]. Ces conférences regroupent chercheurs et industriels pour présenter et promouvoir le logiciel libre dans le géospatial. OSGeo est à l'origine de la librairie open source Geospatial Data Abstraction Library (GDAL) qui est la plus utilisée pour le traitement de données géospatiales.

2.1.2 Les systèmes de coordonnées

Afin de représenter un objet dans l'espace, il est nécessaire d'utiliser un système de coordonnées. Ce système référence un objet dans l'espace grâce à un couple de coordonnées. On distingue deux types de systèmes de coordonnées différents : les systèmes de coordonnées géographiques et les systèmes de coordonnées projetées.

Les systèmes de coordonnées géographiques utilisent un référentiel sphérique en trois dimensions pour définir une position sur la surface de la Terre. Ce référentiel en trois dimensions est défini grâce à une géoïde et une ellipsoïde. La géoïde est une représentation plus détaillée de la surface de la Terre par rapport à une sphère classique (voir Figure 2.1). La géoïde prend en compte la pesanteur

1. `https://www.osgeo.org`
2. `https://foss4g.org/`

pour sa modélisation. Les coordonnées sont définies par la longitude et la latitude, qui sont calculées à partir de la valeur de l'angle par rapport au centre de la Terre. Le système de coordonnées le plus répandu est le World Geodetic System 1984 (WGS84) [3] qui est notamment utilisé par le système GPS.

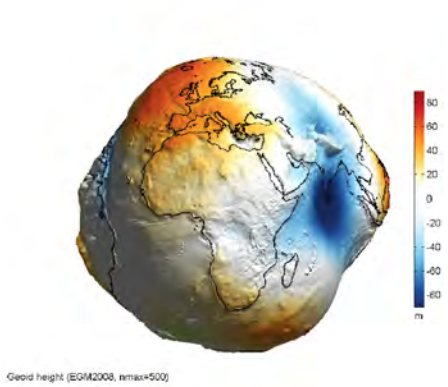


FIGURE 2.1 – Représentation de la géoïde EGM2008 par [Pavlis *et al.* 2012]

Les systèmes de coordonnées projetées se basent sur un référentiel à deux dimensions. Chaque position est définie par des coordonnées X et Y dans le référentiel en prenant compte la forme de la projection. On distingue trois types de projections : cylindrique, conique et planaire (voir Figure 2.2). Le système de coordonnées Lambert-93 [4] utilise une projection conique. Le second système standard utilisé pour les coordonnées projetées est l'Universal Transverse Mercator (UTM) qui permet de représenter la surface de la Terre par un découpage en zones. Ce système de coordonnées géographiques possède une projection cylindrique et est calculé à partir du système de coordonnées géographique WGS84.

Afin d'harmoniser ces différents types de systèmes, l'European Petroleum Survey Group (EPSG) [5] a mis au point un standard de codification pour chaque système. Le WGS84 possède le code 4326 (WGS84 = EPSG :4326) et le Lambert-93 possède le code 2154 (RGF93 = EPSG :2154). Ce système de codification est largement utilisé dans les SIG et permet de n'utiliser qu'un seul référentiel sans avoir à connaître l'ensemble des systèmes existants.

2.1.3 Représentation des entités géo-localisées

2.1.3.1 Les données raster

Une image multi-spectrale est composée de plusieurs matrices, aussi appelées rasters, qui correspondent chacune à une bande spectrale spécifique et où chaque pixel est géo-localisé. Les rasters peuvent également stocker des informations calculées depuis les images satellitaires tels que le land cover ou les indices calculés

3. `https://spatialreference.org/ref/epsg/wgs-84/`
4. `https://geodesie.ign.fr/contenu/fichiers/Lambert93_ConiquesConformes.pdf`
5. `https://epsg.org`

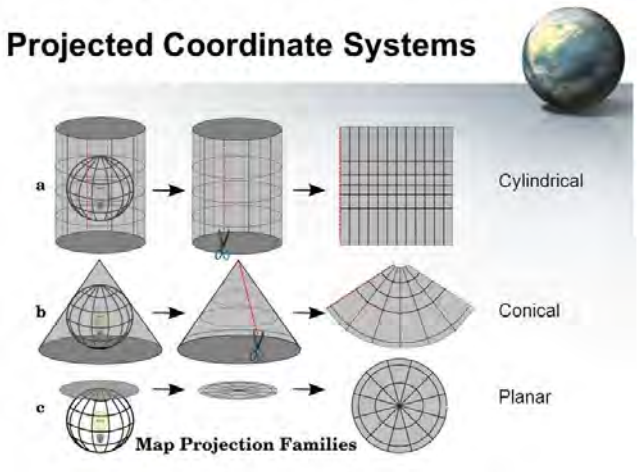


FIGURE 2.2 – Illustration des différents type de système de coordonnées projetées (docs.qgis.org)

comme le NDVI. Nous allons maintenant présenter comment il est possible d'enregistrer ces données issues de capteurs variés. Les formats disponibles pour stocker ces images matricielles sont nombreux [6].

Un des formats les plus répandus pour les images satellitaires est Geographic Tagged Image File Format (GeoTIFF) [7] développé dans les années 1990 par [Ritter & Ruth 1997]. Ce format est un standard ouvert de l'OGC qui étend le format Tagged Image File Format (TIFF) très répandu en photographie. GeoTIFF permet d'ajouter des informations géographiques telles que la projection, le système de coordonnées ou la résolution spatiale de l'image.

Un second format largement répandu est JPEG2000 [8] qui est un standard de codage et de compression d'image développé entre 1997 et 2000 par le groupe Joint Photographic Experts Group (JPEG). Celui-ci est plus performant que le standard JPEG déjà existant et permet donc d'obtenir des images de meilleurs qualité et plus légères. Comme GeoTIFF ce format permet de stocker des métadonnées géographiques directement dans le fichier. C'est le standard adopté par l'ESA pour les images du programme Sentinel.

En plus du raster qui la compose, une image satellitaire contient des métadonnées. Comme décrit précédemment, certaines métadonnées peuvent être directement enregistrées dans le fichier image. Il existe également parfois des métadonnées associées à l'image se trouvant dans un fichier distinct. Ces fichiers peuvent contenir de très nombreuses informations techniques sur les images comme la position du satellite au moment de l'acquisition, la couverture nuageuse ou le nom et la version des processus utilisés par le segment sol pour construire l'image. Le format de ces fichiers peut être Extensible Markup Language (XML), JavaScript Object Notation

6. https://svn.osgeo.org/gdal/trunk/gdal/frmts/formats_list.html
7. https://www.ogc.org/standards/geotiff
8. https://jpeg.org/jpeg2000/

(JSON), ou un simple fichier texte.

Le format et le contenu de ces fichiers sont hétérogènes ce qui pose une problématique majeure pour les organismes souhaitant centraliser ces données dans un catalogue par exemple. Dans une optique d'homogénéisation de ces données, l'entreprise Airbus DS a mis au point le format DIMAP [9] pour formaliser les métadonnées de ses satellites des gammes SPOT et Pléiades. La commission Européenne a également lancé une directive européenne appelée INSPIRE [10] visant à harmoniser l'ensemble des données géographiques publiées par les organismes publiques européens. Elle a mis en ligne un validateur INSPIRE [11] qui permet de vérifier si des données sont conformes aux directives INSPIRE. Les métadonnées des satellites du programme Sentinel respectent les normes INSPIRE.

2.1.3.2 Les données vectorielles

En plus des données images au format raster présentées précédemment, la géolocalisation d'une entité peut s'exprimer au format vectoriel. Le principe est très différent de celui des rasters : au lieu de représenter l'information spatialisée sous forme de grille, le ou les vecteurs permettent de définir l'emprise au sol d'un objet ou d'un groupe d'objets, sous forme d'une ligne (pour les routes, voies d'eau, etc.) ou d'une surface (parcelle agricole, ville ou étendue d'eau par exemple. Ces données vectorielles sont très utilisées en géomatique notamment en cartographie car elles permettent de décrire des entités géographiques et de les repérer sur une carte.

Un vecteur possède deux composantes qui sont d'une part la composante attributaire et d'autre part la composante graphique. La *composante attributaire* permet de décrire les informations propres à l'entité que l'on souhaite représenter. Ces attributs sont définis grâce à des champs textuels pour la plupart. Par exemple, si l'on souhaite représenter une route au format vectoriel, une information attributaire peut être le nom de cette route.

La *composante graphique* sert quant à elle à décrire la forme de l'entité appelée géométrie. Cette forme est composée de noeuds qui sont des points référencés dans l'espace grâce à des coordonnées (X, Y) ainsi que des relations qui lient ces noeuds entre eux. Les trois géométries les plus répandues sont : le point, la ligne et le polygone. Le point est représenté par un seul noeud, la ligne est représentée par deux noeuds, ou plus liés, entre eux mais sans former une forme géométrique fermée. Le polygone est représenté par un ensemble de noeuds reliés entre eux en formant une forme géométrique fermée. La figure 2.3 représente les géométries les plus utilisées dans les SIG.

D'autres géométries plus complexes existent telles que le Multipoint qui permet de représenter une entité composée d'un ensemble de points. Le Multipolygon permet quant à lui de représenter une entité constituée d'un ensemble de polygones disjoints comme un archipel par exemple.

9. `https://www.intelligence-airbusds.com/en/8722-the-dimap-format`
10. `https://inspire.ec.europa.eu/about-inspire/`
11. `https://inspire.ec.europa.eu/validator/about/`

Geometry primitives (2D)

Type	Examples	
Point		POINT (30 10)
LineString		LINESTRING (30 10, 10 30, 40 40)
Polygon		POLYGON ((30 10, 40 40, 20 40, 10 20, 30 10))
		POLYGON ((35 10, 45 45, 15 40, 10 20, 35 10), (20 30, 35 35, 30 20, 20 30))

FIGURE 2.3 – Les trois géométries les plus utilisées dans les SIG

L'avantage principal d'utiliser une représentation vectorielle par rapport au format raster est la taille du fichier. En effet, une représentation vectorielle étant composée d'un ensemble de points, celle-ci ne contiendra qu'une liste de coordonnées. Au contraire, le fichier raster permet de connaître les coordonnées de tous points de la zone couverte par l'image, au sein d'une représentation compacte mais volumineuse.

2.1.3.3 Les formats des données vectorielles dans les SIG

Il existe de nombreux formats de fichiers pour enregistrer ce type de données. Les formats les plus répandus sont le Shapefile, le Well-Known Text (WKT), le Geography Markup Language (GML) et le Geographic JSON (GeoJSON).

Le format **Shapefile** est un format ouvert de données vectorielles développé par l'entreprise Américaine ESRI [12]. Ce format est très utilisé dans un grand nombre de logiciels libres comme QGis [13] ou MapServer [14]. Les fichiers au format Shapefile (.shp) sont accompagnés d'au moins deux autres fichiers, DBF et SHX : le SHP contient les différentes géométries des objets à représenter, le fichier DBF contient l'ensemble des attributs comme le nom de l'entité ou son identifiant et le fichier SHX contient l'index de la géométrie.

Le **format WKT** [15] est un format texte permettant de décrire des objets par leur géométrie au format vectoriel. Il s'agit d'un standard de l'OGC. Ce format est très répandu dans les bases de données spatiales comme PostGIS [16] ou Oracle Spatial.

Le **format GML** est un dérivé de XML. C'est un standard développé par l'OGC

12. https://www.esri.com
13. https://www.qgis.org
14. https://mapserver.org/
15. https://docs.opengeospatial.org/is/12-063r5/12-063r5.html
16. https://postgis.net/

qui a l'avantage d'être lisible de la même manière qu'un fichier XML. Il permet de décrire des géométries complexes ainsi que des topologies au format vectoriel.

Enfin le **format GeoJSON** n'est pas un standard de l'OGC. Il a été développé par l'Internet Engineering Task Force (IETF) qui tente de le spécifier avec notamment la rédaction d'une Request for comments (RFC) [Butler *et al.* 2016]. Ce format est le plus récent et a été élaboré par la communauté du Web mais il est aujourd'hui très répandu en géomatique notamment dans les SIG.

2.1.4 Production d'images d'observation de la Terre par satellite

2.1.4.1 Satellites d'observation de la Terre

Aujourd'hui, selon l'Union of Concerned Scientists (UCS) [17], il existe plus de 3370 satellites en orbite autour de la Terre. Une partie de ces satellites peuvent être visualisés en temps réel via des interfaces Web comme Satmap [18]. Ces satellites de différentes nationalités ont des missions variées et des fins commerciales, gouvernementales ou militaires. Au 1er avril 2019, 38% des satellites en orbite concernaient des missions d'observation de la Terre et d'étude du climat. 37% des satellites concernaient la télécommunication et 7% pour la navigation avec les systèmes GPS. Le reste des satellites ont des objectifs divers tels que la recherche spatiale ou l'Internet of Things (IoT).

Ces satellites possèdent des caractéristiques différentes en fonction de leurs missions. En effet, un satellite peut être en orbite géostationnaire à 36 000 km d'altitude ou en orbite basse jusqu'à 2000 km d'altitude. Les satellites géostationnaires permettent d'observer une zone fixe de la Terre comme les satellites du programme Meteosat [19] qui envoient des images hémisphériques qui couvrent l'Europe, l'Afrique et une partie de Amérique du Sud. La zone couverte par une image satellitaire est appelée **fauchée**. Les satellites défilants sont en mouvement et prennent des images de la zone qu'ils sont en train d'observer au moment de leur passage. Étant en orbite basse, ils possèdent une fauchée plus restreinte que les satellites géostationnaires. Une autre notion importante est la **résolution temporelle** d'un satellite. Cette résolution temporelle correspond au temps que met le satellite pour effectuer un cycle orbital complet. Cela permet de savoir au bout de combien de temps un satellite peut reprendre une image de la même zone. Cette période peut varier de un à quelques jours. Pour les satellites géostationnaires, la résolution temporelle correspond au temps minimum qu'il peut y avoir entre deux prises d'images.

Une autre notion importante dans l'imagerie spatiale est la **résolution spatiale** d'une image. Cette résolution correspond à la taille du plus petit élément qu'il est possible de distinguer sur l'image. Elle dépend du capteur utilisé et plus la résolution spatiale augmente plus la superficie totale couverte par l'image diminue. Une image du satellite Sentinel-2 possède une résolution spatiale de 10 m et peut couvrir une

17. `https://www.ucsusa.org/`
18. `https://satmap.space/`
19. `https://www.eumetsat.int/meteosat-second-generation`

largeur au sol de 290 km. Au contraire, une image du satellite Pléiade [20] possède une résolution spatiale de 0.7 m et peut couvrir une surface au sol de 60 km de large.

L'évolution des différentes technologies de capteurs permet d'élargir le champs des missions possibles à l'aide des images satellitaires. En effet, les premiers satellites d'observation du programme Landsat [21] 1 à 3 de la National Aeronautics and Space Administration (NASA) lancés entre 1972 et 1978 avaient pour mission la surveillance des terres et océans. Ces satellites embarquaient deux capteurs et avaient une résolution spatiale de 38m. D'autres capteurs développés dans les 1980 permettent d'effectuer de l'imagerie radar. Ces satellites avaient des missions principalement de recherche et les images n'étaient pas accessibles au public.

Aujourd'hui, des capteurs beaucoup plus puissants sont disponibles. Les satellites d'imagerie optique peuvent avoir une résolution inférieure à 1m comme le satellite WorldView-3 [22] possédant une résolution spatiale de 0.3m. Les satellites à très haute résolution ont des buts commerciaux où les images sont commandées comme des produits à l'entreprise chargée de les fournir.

Il existe d'autres missions comme le programme Sentinel de l'ESA, présenté précédemment, ou les missions Landsat 7/8 qui fournissent aujourd'hui des images gratuites en libre accès. Les images des satellites Landsat sont accessibles via le portail de l'United States Geological Survey (USGS) [23] et les images Sentinel via le portail de l'ESA [24]. Ces satellites possèdent des résolutions moins élevées comprises entre 10 m et 50 m mais embarquent une large gamme de capteurs. Leurs images sont principalement utilisées à des fins scientifiques ou dans le milieu agricole.

2.1.4.2 La télédétection

Pour nos travaux, nous nous intéressons aux satellites d'observations de la Terre car ceux-ci permettent la télédétection.

> « La télédétection est la technique qui, par l'acquisition d'images, permet d'obtenir de l'information sur la surface de la Terre sans contact direct avec celle-ci. La télédétection englobe tout le processus qui consiste à capter et à enregistrer l'énergie d'un rayonnement électromagnétique émis ou réfléchi, à traiter et à analyser l'information, pour ensuite mettre en application cette information. » - Centre Canadien de Télédétection (CCT) [25]

Le processus de télédétection se compose des sept étapes suivantes :

— **A : Illumination ou source d'énergie** - La source d'énergie permet d'illuminer l'objet ou la surface à étudier.

20. https://pleiades.cnes.fr
21. https://landsat.gsfc.nasa.gov/
22. http://worldview3.digitalglobe.com/
23. https://earthexplorer.usgs.gov/
24. https://scihub.copernicus.eu/dhus/
25. http://www.rncan.gc.ca

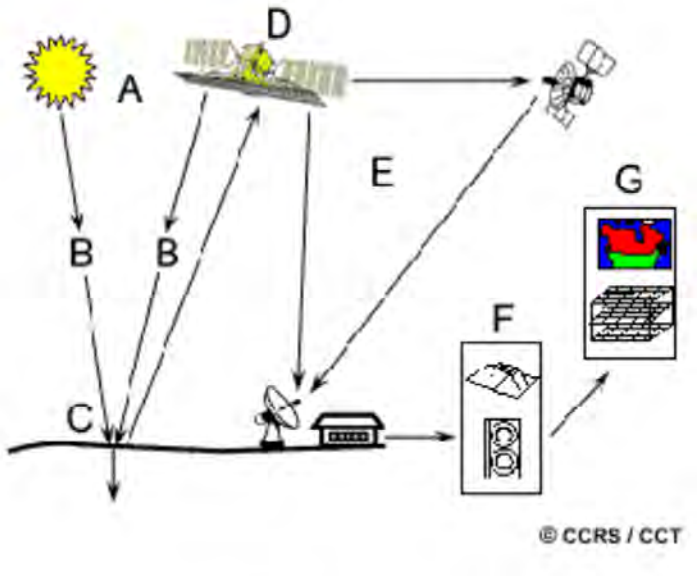


FIGURE 2.4 – Les sept étapes du processus de télédétection. source : CCT

— **B : Rayonnement et atmosphère** - Le rayonnement interagit avec l'atmosphère, une première fois entre la source et la cible et une seconde fois entre la cible et le satellite.

— **C : Interaction avec la cible** - La nature de l'interaction avec la cible à étudier dépend de la caractéristique du rayonnement ainsi que les propriétés de l'objet à étudier.

— **D : Enregistrement via le capteur** - Le capteur intercepte l'énergie diffusée ou émise par la cible et l'enregistre.

— **E : Transmission, réception et traitement** - L'information enregistrée par le capteur est transmise au segment sol qui transforme les données en images raster.

— **F : Interprétation et analyse** - Une interprétation numérique et/ou visuelle permet d'obtenir des informations sur la cible.

— **G : Application** - L'interprétation est exploitée afin de résoudre une problématique ou découvrir de nouvelles informations sur la cible.

L'ensemble de nos travaux portent sur les étapes F et G du processus de télédétection. Ces étapes sont réalisables dès lors que le segment sol fournit l'image comme produit fini. Il appartient à l'entreprise qui gère le segment sol de fournir les images gratuitement ou à des fins commerciales.

En télédétection, il peut y avoir deux sources d'onde :

— La première source est le soleil qui émet des rayons qui seront réfléchis par la surface de la Terre puis retournés au capteur. Il s'agit d'une *télédétection passive* et le résultat obtenu est une *image optique*.

— La seconde source d'émission de rayons peut être un émetteur placé sur le satellite qui envoie des ondes qui seront, elles aussi, réfléchies par la cible et retournées au capteur. On appelle cela une *télédétection active* et le résultat est une *image radar*.

La télédétection repose sur deux composantes du rayonnement électromagnétique qui sont la **longueur d'onde** et la **fréquence**. L'oeil humain est capable de voir la lumière se trouvant dans le spectre du visible. Les longueurs d'ondes qui composent ce spectre sont comprises entre 0,4 μm et 0,7 μm comme montré dans la Figure 2.5.

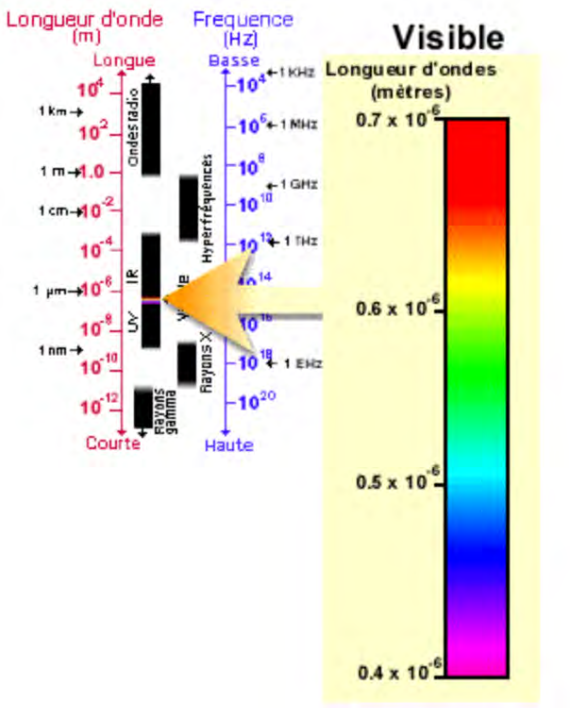


FIGURE 2.5 – Illustration des longueurs d'ondes et fréquence (CCT)

En télédétection radar, les domaines spectraux des micro-ondes vont être exploités alors que la télédétection optique va exploiter les longueurs d'ondes du visible ainsi que de l'infrarouge.

Le principe du rayonnement est que le capteur intercepte la lumière qui a été réfléchie par la cible. En télédétection, l'atmosphère joue un rôle important dans cette étape car les gaz qui la composent dévient les rayons. Ce sont les molécules de gaz comme l'azote ou l'oxygène qui perturbent ce rayonnement mais aussi les particules de poussières. Il faut donc les prendre en compte lors de l'interprétation.

Lorsque les rayons atteignent la cible, celle-ci peut absorber, transmettre ou réfléchir ces rayons. On appelle la capacité d'un objet à réfléchir les rayons la **réflectance**. Cette réflectance est le rapport entre le flux lumineux reçu par le capteur et le flux lumineux émis en pourcentage.

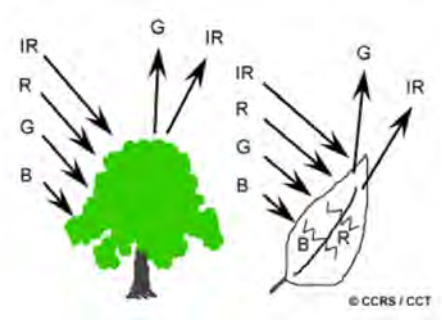


FIGURE 2.6 – Illustration de la réflectance des feuilles (CCT)

La Figure 2.6 montre un exemple de flux lumineux arrivant sur une feuille d'arbre. La chlorophylle présente dans la feuille va fortement absorber les longueurs d'ondes rouge et bleu et va réfléchir une partie des longueurs d'ondes verte et infrarouge.

Dans une image satellitaire, les informations sur les couleurs sont décomposées en bandes spectrales. Chaque bande est enregistrée dans un fichier raster en niveau de gris qui correspond à une couleur. Une bande couvre une partie du spectre lumineux que l'on peut qualifier comme "bande du bleu" pour le spectre du bleu par exemple. Ce raster est composé de pixels qui auront une valeur élevée pour les surfaces ayant une réflectance importante pour cette partie du spectre. Pour qu'une image puisse être interprétée en "vraies couleurs", il faut donc qu'elle soit composée des bandes du rouge, du vert et du bleu. En opposition il est possible de représenter une image en "fausses couleurs" comme dans la figure 2.7. Ce principe permet de discriminer visuellement de la végétation par exemple. Pour représenter une image en fausse couleur, il faut associer la bande infrarouge à la place de la bande rouge, la bande rouge à la place de la bande verte et la bande du vert à la place du bleu.

FIGURE 2.7 – Exemple d'une image satellitaire de la ville de Copenhague en fausses couleurs

2.1.4.3 Principaux traitements dans les images satellitaires

Le terme "analyse d'images" renvoie à des réalités et des processus très différents suivant la nature des images et les domaines d'application, d'une grande di-

versité. L'analyse peut concerner des images photographiques (par exemple, issues de réseaux sociaux, de systèmes de surveillance ou de vues aériennes), des images radiologiques (en santé, en histoire de l'art ...). Nous nous intéressons ici seulement aux travaux relatifs aux images d'observation de la Terre à partir de satellites, et donc au traitement d'images en télédétection. En télédétection, il existe plusieurs processus associés au traitement d'images pour en extraire des informations.

La classification : Un traitement très connu en télédétection est la classification des pixels au sein des images. Lorsque ces algorithmes sont spécialisés pour l'étude de la couverture des sols, ils sont appelés *land cover classification*. Ce processus consiste à classer numériquement chaque pixel dans l'image pour l'associer à une classe possédant des propriétés particulières puis à qualifier ces classes. Le but est ici d'obtenir une classe ou un thème à partir de la valeur de l'intensité des pixels. Aujourd'hui il existe deux types de classification : la classification supervisée et la classification non-supervisée. La figure 2.8 montre une classification de l'occupation des sols en France pour l'année 2014 réalisée par le laboratoire Centre d'Etudes Spatiales de la Biosphère (CESBIO).

Le principe de la *classification supervisée* est de définir les classes ou thèmes (comme par exemple : forêt, eau ou champ) et d'associer à chaque pixel de l'image une de ces classes. Le principe de la supervision est de construire un grand jeu d'exemples déjà classés qui vont servir à entraîner l'algorithme. L'idée est d'exploiter un ensemble d'images contenant des types de surfaces variées pour identifier un maximum de classes. Les pixels de ces images sont classés manuellement par un expert qui indique, pour chaque pixel, la classe à lui associer. Une fois ces exemples construits, un algorithme de classification supervisée, comme la classification naïve bayésienne ou Support Vector Machines (SVM), est entraîné à partir de ces exemples. On peut ensuite l'appliquer à l'image souhaitée pour déterminer la classe la plus proche à laquelle pourrait appartenir chaque pixel. Les travaux de [F.Y *et al.* 2017] font état des algorithmes supervisés les plus répandus et comparent leurs performances.

Concernant la *classification non-supervisée*, le principe est légèrement différent. Les classes ne sont plus déterminées par un utilisateur mais calculées par l'algorithme lui-même. Cet algorithme va construire des classes en regroupant les valeurs similaires ou proches de chaque pixel. L'utilisateur doit ensuite associer un label aux classes ainsi identifiées. Il peut également spécifier en paramètre de ces algorithmes le nombre de classes à identifier dans l'image pour en améliorer les performances. Ces algorithmes fonctionnent par itération et peuvent parfois générer des classes que l'utilisateur final devra séparer ou fusionner. L'avantage est ici que l'utilisateur n'a aucun exemple à fournir, mais en contre partie, il doit être capable d'identifier le nom des classes générées et d'indiquer si celles-ci sont correctes. Les travaux de [Abbas *et al.* 2016] montrent qu'il est possible d'appliquer une classification non-supervisée à des images satellitaires afin de déterminer les différentes classes de couverture des sols.

Il existe également des organismes qui calculent ces classifications de la surface de la Terre à une date donnée, et les mettent à disposition au format raster. Le

service Corine Land Cover [26] du projet Copernicus fournit la valeur du couvert des sols sur toute l'Europe pour les années 1990, 2000, 2006, 2012 et 2018 avec une résolution spatiale de 100 m. Ces classifications ont été établies à partir de différentes images provenant des satellites comme Landsat-7/8 ou Sentinel 2 pour chaque année. L'organisation Food and Agriculture Organization (FAO) des Nations-Unies met à disposition un classement du couvert des sols sur l'ensemble de la Terre appelé Global Land Cover-SHARE (GLC-SHARE). Ce land cover est disponible pour l'année 2014 et possède une résolution spatiale de 1000 m. Il a été calculé à partir d'images issues de satellites comme MERIS ou MODIS. Le CESBIO a calculé un classement du couvert des sols sur la France à partir d'images Sentinel-2 pour les années 2016 et 2018 avec une résolution spatiale de 10 m. La méthodologie et les algorithmes utilisés sont détaillés dans [Stoian *et al.* 2019].

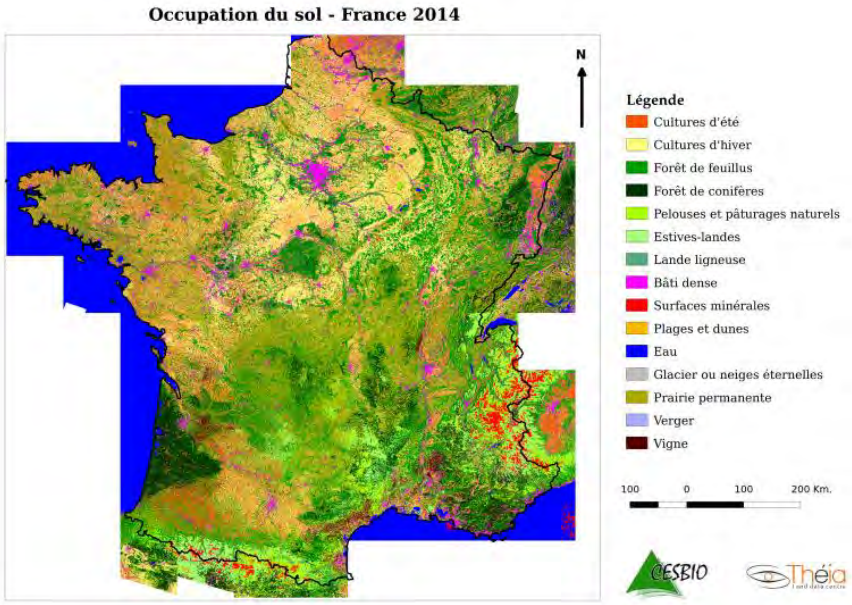


FIGURE 2.8 – Classification de l'occupation du sol en France par le CESBIO pour l'année 2014.

Le calcul d'indices : Un autre traitement très répandu en télédétection est le calcul d'indices comme le NDVI, le Normalized Difference Snow index (NDSI), le Normalized Difference Water index (NDWI) ou encore le Normalized Difference Cloud Index (NDCI). Ces indices normalisés sont très utiles lorsque l'on souhaite étudier un type de sol connu. Le NDVI permet par exemple d'étudier la couverture de végétation présente sur le sol dans la zone captée par une image, le NDSI permet d'étudier la couverture neigeuse et le NDWI permet d'étudier les surfaces composées d'eau. Enfin, le NDCI permet de localiser les nuages qui masquent la surface terrestre et peuvent perturber certaines études ou d'autres calculs. Chacun de ces

26. `https://land.copernicus.eu/pan-european/corine-land-cover`

indices repose sur un calcul réalisé à partir de plusieurs bandes d'une image multispectrale. Les bandes concernées et le calcul à effectuer sont propres au phénomène étudié. Par exemple, le NDVI est calculé grâce à une opération arithmétique entre la bande du proche infrarouge (PIR) qui est un indicateur de la haute réflectivité des matières végétales et la bande du rouge (R) qui est révélatrice de l'absorption du pigment chlorophyllien d'une image. L'opération est détaillée dans le calcul 2.1.

$$NDVI = \frac{(PIR - R)}{(PIR + R)} \tag{2.1}$$

Le résultat obtenu est une matrice de la même dimension que les bandes utilisées pour cette opération. Les valeurs de cette matrice sont comprises entre -1 et 1. Les valeurs négatives comprises entre -1 et 0 ont une forte probabilité de représenter des surfaces composées d'eau comme les lacs ou la neige. Les valeurs positives comprises entre 0 et 1 représentent les surfaces végétales comme les champs ou les forêts. Plus la valeur est élevée plus la végétation est dense. Ces indices ne sont pas assez fiables pour déterminer une classification précise des images mais ils permettent d'évaluer la quantité de végétation, de neige ou surface d'eau présentes dans les images sans avoir à les visualiser. Dans les travaux de [Gandhi *et al.* 2015], il est montré comment étudier les évolutions des types de végétation en calculant le NDVI sur plusieurs années.

La détection d'objets : Un des traitements les plus répandus est la reconnaissance d'objets à partir de l'analyse du contenu de ces images. Ces objets peuvent être des bateaux, des bâtiments ou encore des champs. Beaucoup de travaux comme [Pritt & Chern 2020, Pathak *et al.* 2018] reposent sur un apprentissage automatique à partir des images, et utilisent par exemple certains des algorithmes de classification présentés précédemment, des algorithmes à base de réseaux de neurones Convolutional Neuron Network (CNN) et de modèles entraînés. Même si ces algorithmes obtiennent de bons résultats, il faut entraîner au préalable les modèles sur lesquels ils s'appuient en utilisant des vignettes d'images contenant les objets que l'on souhaite pouvoir détecter. Cette étape est très contraignante car les traitements requièrent une banque d'images conséquente. De plus, pour certains types d'apprentissage, comme l'apprentissage supervisé, il faut étiqueter les images pour indiquer les objets reconnus et leur localisation dans l'image. Plus la banque d'images est importante, plus la détection d'objet sera fiable, mais plus les ressources de calcul Graphics Processing Unit (GPU) nécessaires seront importantes.

La détection de changements : La détection de changements consiste à comparer deux ou plusieurs images prises sur la même localisation et à des dates différentes. Il existe aujourd'hui plusieurs méthodes afin de détecter un changement et comme pour la détection d'objets, les derniers travaux dans le domaine utilisent aussi les réseaux de neurones [Amin *et al.* 2016, Jia *et al.* 2014, Gong *et al.* 2017]. L'avantage de la détection de changement par rapport à la détection d'objet est que celle-ci ne requiert pas d'entraînement à l'aide de vignettes. Ces algorithmes utilisent des modèles entraînés sur des images entières et ne demandent aucun découpage en vignettes. Le résultat obtenu par ces algorithmes prend la forme d'un raster

de changement, à chaque pixel étant associé un indice de changement indiquant la probabilité qu'un changement ait eu lieu entre les images comparées. L'algorithme développé dans [Aubrun *et al.* 2020] explique comment obtenir ce type de raster sans supervision.

2.2 Le Web sémantique

Dans cette partie, nous rappelons les objectifs du Web sémantique, et les définitions des principaux concepts relatifs à ce domaine. Nous présentons également les standards définis par le World Wide Web Consortium (W3C) pour mettre en oeuvre pratiquement ces concepts. Ils nous permettrons ensuite (chapitre 3) de préciser comment ces technologies facilitent la réponse aux nouveaux défis de partage de données d'observation de la Terre et de leur croisement avec d'autres données pour différents types d'applications, en particulier l'étude des changements calculés à partir de ces images.

2.2.1 Le projet du Web sémantique

A l'inverse du Web, le Web sémantique n'a pas pour but d'être lu, mais d'être exploité par des applications et des services. L'objectif du Web sémantique est de transformer la grande quantité de données non structurées ou mal documentées disponibles sur le Web de telle sorte qu'elle soit facilement traitée par des services et des programmes. Pour cela, ce projet vise l'extraction de connaissances à partir de textes ou de données, l'utilisation de formats standards pour représenter ce qui est extrait, ou pour annoter ces données, et le choix de vocabulaires partagés, formels et communs pour produire une représentation sémantique qui facilite les échanges et l'interopérabilité [Berners-Lee 1998].

L'expression Web sémantique, fait d'abord référence à la vision du Web de demain comme un vaste espace d'échange de ressources entre êtres humains et machines permettant une exploitation, qualitativement supérieure, de grands volumes d'informations et de services variés. Les utilisateurs se trouvent déchargés d'une bonne partie de leurs tâches de recherche, de construction et de combinaison des résultats, grâce aux capacités accrues des machines à accéder aux contenus des ressources et à effectuer des raisonnements sur ceux-ci [Laublet *et al.* 2004].

Le résultat attendu est de disposer de représentations des contenus en ligne permettant leur exploitation de manière plus précise et plus intelligente par des solutions informatiques. Cela est possible grâce à la construction et le partage de données structurées, disponibles en ligne dans des portails ou des référentiels.

2.2.1.1 Ontologies et Web sémantique

La structuration de ces données passe par la définition d'ontologies. L'équipe de [Studer *et al.* 1998] a proposé une définition du terme ontologie qui allie la définition

de [Gruber 1993] et celle de [Borst 1997] : "An ontology is a formal, explicit specification of a shared conceptualization". L'équipe de [Guarino *et al.* 2009] a choisi d'orienter leur définition de l'ontologie sur les notions de *conceptualization* et *explicit specification*.

Idéalement, les ontologies définissent les concepts ou classes d'entités d'un domaine et leurs propriétés essentielles, qui font qu'une entité sera rattachée à cette classe si on reconnaît qu'elle possède ces propriétés. L'ensemble des concepts et de leurs propriétés forme un vocabulaire formel qui doit être partagé par une communauté d'usage [Noy & Mcguinness 2001]. La sémantique d'une ontologie restitue la compréhension partagée de ces concepts et propriétés. Elle se traduit par un ensemble d'axiomes qui explicitent comment déduire de nouvelles connaissances à partir de celles formalisées dans l'ontologie. Dans la pratique, les ontologies du Web sémantique sont des modèles au sens où elles simplifient le réel selon un point de vue, une finalité, adaptée à des applications informatiques et leurs usagers. Elles représentent les concepts de ce domaine et leurs relations de spécialisation, ainsi que leurs propriétés, en fonction de la manière dont on va les utiliser, ainsi que des axiomes explicitant la sémantique des relations (hiérarchiques ou non) entre concepts ou la sémantique des contraintes sur les propriétés qui les relient (cardinalité, types d'entités reliées les propriétés, métapropriétés des propriétés, etc.)

Le mot *ontologie* est utilisé pour désigner une gamme de modèles de connaissances plus ou moins précisément formalisées. Lorsque le besoin de représentation des données ne nécessite pas des définitions poussées pour produire des raisonnements, il est commun d'établir un *Vocabulaire* (au sens du Web sémantique). Un vocabulaire ou *vocabulaire contrôlé* est composé d'une hiérarchie de concepts, à savoir de classes organisées selon la relation d'inclusion entre elles et reliées par des propriétés étiquetées [Arano 2005]. La définition des concepts est donnée en langage naturel et non sous forme de conditions nécessaires et suffisantes sur la présence de ces propriétés. Dans la plupart des cas, les vocabulaires contrôlés sont mis à disposition en ligne afin d'être réutilisés sans être reliés à d'autres données sémantiques.

Pour une représentation plus complète des concepts, il est nécessaire d'utiliser une ontologie. Celle-ci permet des représentations plus détaillées des concepts avec notamment les notions de domaine, co-domaine et cardinalité associés aux relations. La définition d'une ontologie est un procédé complexe et il est nécessaire d'avoir connaissance de l'ensemble des concepts pour pouvoir les lier ensemble via des propriétés. Une fois définie, une ontologie peut être publiée en ligne afin qu'elle soit réutilisée par la communauté. Les ressources décrites en utilisant ces ontologies sont représentées sous forme de graphes de connaissances formés d'instances des concepts de l'ontologie, de relations entre ces instances et éventuellement de règles ou d'axiomes utilisant ces relations. Ces graphes de connaissances peuvent également être publiés sur le Web ou dans le Linked Open Data (LOD).

Afin d'harmoniser les ontologies disponibles en ligne et les ontologies créées pour un besoin métier par exemple, il est possible de les lier en spécifiant l'équivalence d'un concept d'une ontologie à un autre concept dans une seconde ontologie. Ce pro-

cédé s'appelle l'alignement d'ontologie et il existe aujourd'hui de nombreux travaux portant sur cette problématique [Thiéblin 2019]. Il existe des ontologies de plusieurs niveaux pour permettre de lier des concepts plus ou moins précisément. Les ontologies dites "haut-niveaux" comme DOLCE détaillée dans [Borgo & Masolo 2010], ont un niveau d'abstraction plus élevé et permettent de décrire plus de concepts. Les ontologies de domaine sont quant à elles plus précises sur la description des concepts et les propriétés et sont orientées sur un domaine spécifique.

Le site du W3C [27] rappelle à quel point, dans la mise en oeuvre du Web sémantique, les mots *vocabulaire contrôlé* et *ontologie* sont utilisés pour désigner des modèles de même type et l'un pour l'autre .

2.2.1.2 Les Linked Open Data, données liées du Web

Le LOD (présenté dans l'introduction) est une évolution du Web qui a été définie pour mettre en avant les données disponibles sur le Web, en les rendre plus facilement accessibles, compréhensibles et réutilisables. Dans la pratique, l'idée est de représenter les données au sein de graphes de données liées, en utilisant des vocabulaires contrôlés pour les décrire et indiquer leur origine, ainsi que des formats standards pour faciliter leur réutilisation et leur mise en relation.

Tout comme le Web qui peut comporter plusieurs fois le même jeu de données, le LOD peut contenir des redondances, la même entité ou la même classe pouvant être définie dans plusieurs graphes. Dans le LOD, l'accent est mis sur la mise en relations, les correspondances entre entités de jeux de données différents, pour rendre explicite qu'il s'agit bien de mêmes entités. Le fait de définir de simples vocabulaires sans axiomatique ni relation à des ontologie de plus haut niveau, suffit à donner du sens. La modélisation par ontologies est moins utilisée car plus complexe à réaliser. Pour faciliter réutilisabilité et interopérabilité, le LOD repose sur des normes, standards et recommandations dont une grande majorité provient du W3C : Resource Description Framework (RDF) et RDF Schema (RDFS) pour la représentation des graphes de connaissances, SPARQL Protocol And RDF Query Language (SPARQL) pour leur interrogation, le protocole HyperText Transfer Protocol (HTTP) pour l'accès et les Uniform Resource Identifier (URI) pour leur identification. La figure 2.9 est une représentation visuelle des jeux de données disponibles dans le LOD en 2021. On constate une forte densité de liens entre les jeux de données, le grand nombre de jeux de données en sciences de la vie et de données linguistiques, et le rôle clé des publications pour établir des liens entre thématiques. Une comparaison avec les visualisations des années précédentes permettrait de confirmer la croissance rapide du LOD depuis 10 ans.

2.2.1.3 Les structures de connaissances du Web sémantique

La structure de représentation de base du Web sémantique est le *triplet* représenté sous la forme <s, p , o> pour <sujet, prédicat, objet> dont le sujet est une

27. `https://www.w3.org/standards/semanticweb/ontology`

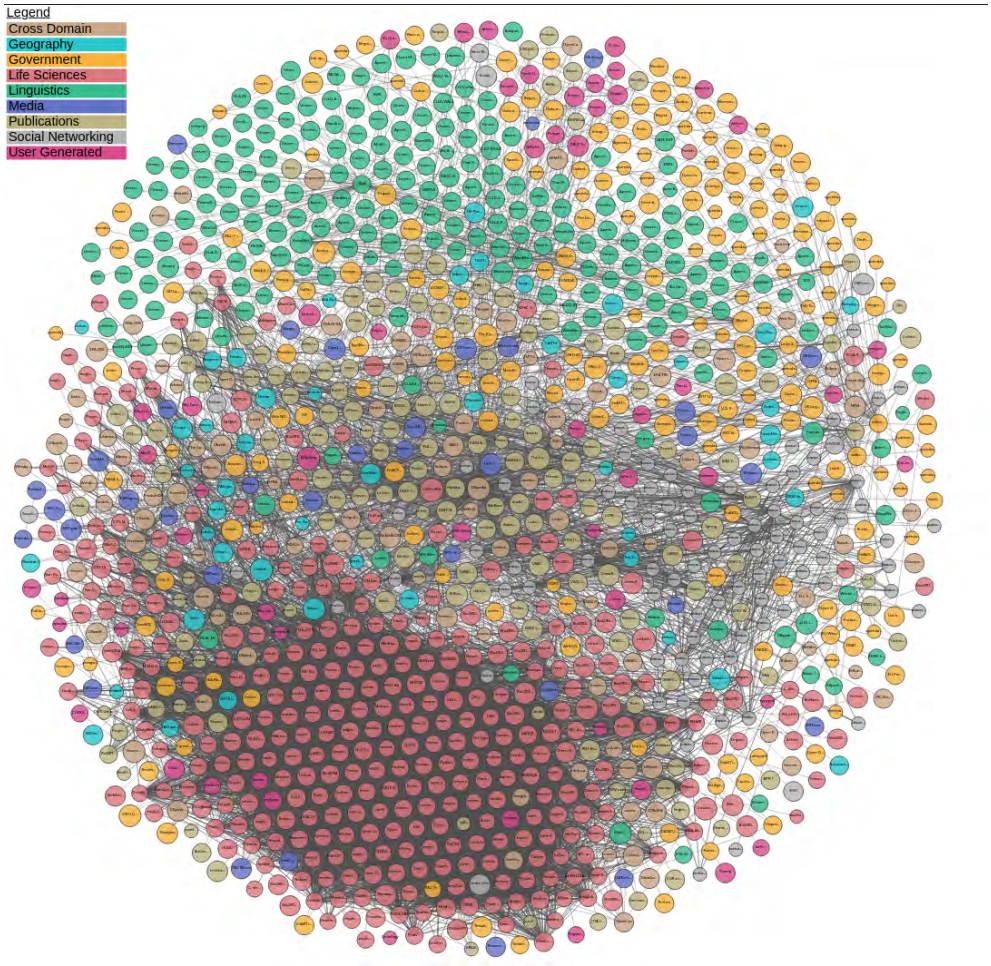


FIGURE 2.9 – Représentation visuelle du LOD en 2021 (source : lod-cloud.net)

ressource du Web désignée par un URI, le prédicat une propriété définie dans un vocabulaire contrôlé, et l'objet est soit une ressource, soit une valeur (numérique, chaîne de caractères ..). Plusieurs triplets liés entre eux forment un ou plusieurs *graphes de connaissance*. Une ressource peut faire référence à une entité précise (par exemple Paris) ou à une classe d'entités (par exemple une ville, une capitale) définie dans un vocabulaire contrôlé ou une ontologie. L'ontologie définit les classes nécessaires pour décrire un domaine, et leurs propriétés, jouant le rôle d'un schéma de structuration des données du graphe. Ces représentations de modèles sous formes de concepts ou classes qui pourront ensuite être instanciées par des entités ou individus, pour former une base de connaissances. Les classes sont organisées en hiérarchie selon des relations d'inclusion d'une sous-classe dans une classe. Cette hiérarchisation permet de raisonner sur l'appartenance d'une entité à d'autres classes que celle à laquelle elle est directement rattachée. La sémantique de la relation d'inclusion entre classes (subClass-of) est que Si une classe B est définie comme sous-classe de A alors toute entité de la classe B appartient aussi à la classe A. Les ontolo-

gies permettent aussi d'exprimer des méta-propriétés sur les relations, comme la transitivité ou la symétrie.

On distingue deux types d'ontologies selon la richesse de la formalisation des concepts : les *ontologies lourdes* composées de classes et propriétés mais aussi de règles et axiomes qui viennent ajouter de la complexité dans la description ; les *ontologies légères* qui se concentrent sur l'organisation hiérarchiques des concepts formant un vocabulaire des données et qui fournissent une description plus simple des données. Une des particularités majeures de ces modèles de connaissances par rapport aux bases de données relationnelles est que le schéma qui organise les connaissances est explicitement déclaré dans le fichier des données et accessible sur le Web.

Le Web sémantique repose sur l'hypothèse du **monde ouvert** pour la représentation des connaissances, contrairement aux bases de données relationnelles qui supposent que la base fonctionne dans un **monde fermé**. Le paradigme du monde ouvert signifie que les données qui ne sont pas présentes dans le système ne sont pas connues mais peuvent exister ailleurs [Drummond & Shearer 2006]. Donc par défaut, tout fait, même non explicité, et potentiellement déductible, est supposé vrai, ce qui peut conduire à des incohérences.

L'approche du monde fermé, utilisée dans les bases de données relationnelles, signifie qu'une donnée qui n'est pas présente dans le système n'existe pas dans celui-ci ni dans un autre. Le monde ouvert est utile lorsque l'on souhaite modéliser un grand nombre de données de manière incrémentale, en prenant en compte de nouvelles sources, et sans connaître a priori leur nombre exact ni leur emplacement de stockage. L'inconvénient de cette approche est qu'il n'est pas possible de référencer l'ensemble des informations pour vérifier si celles-ci sont toutes présentes. Dans le cas où l'ensemble des données doit être vérifié, l'approche du monde fermé est plus adaptée car chaque information est connue et localisée dans le système.

2.2.2 Les technologies du Web sémantique

Pour représenter ces modèles/structures de connaissances, l'organisme de standardisation du Web, le W3C, a proposé plusieurs standards d'expressivité et de complexité croissante représentés sous forme de couches présentes dans la Figure 2.10.Le W3C a fait le choix de définir ces standards à partir des standards existants d'identification sur le web (les URI) et du langage standard de balisage de documents XML.

Pour écrire des ontologies, on utilise les langages Web Ontology Language (OWL) ou RDFS tous deux basés sur RDF. RDFS convient pour définir des ontologies légères avec des relations simples, alors que OWL permet de décrire des ontologies lourdes avec des relations plus complexes.

Dans la suite, nous présentons 4 des standards du W3C : RDF, RDFS, OWL et ses différentes déclinaisons, et SPARQL. Chacun de ces langages a été défini à un niveau logique par des recommandations du W3C qui ont été implémentées dans

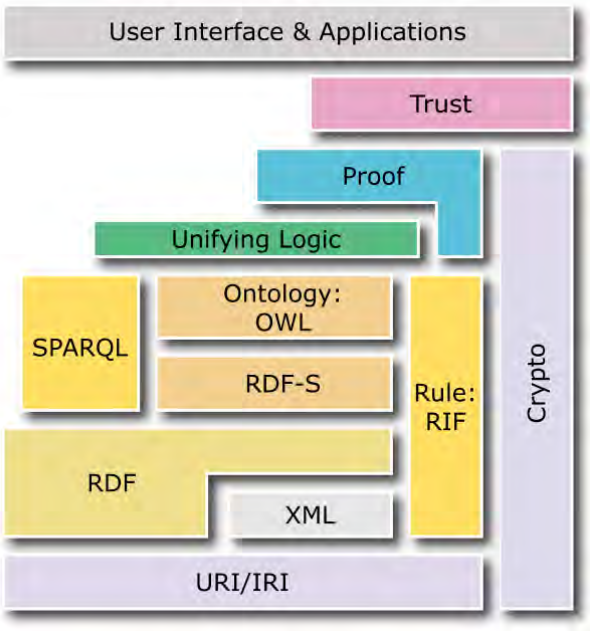


FIGURE 2.10 – Pile du web sémantique selon le W3C

plusieurs langages de sérialisaton : XML, la notation simplifiée Notation3 (N3) [28] et sa simplification Turtle [29]. La syntaxe recommandée par le W3C est basée sur la syntaxe XML (RDF/XML, OWL/XML).

2.2.2.1 RDF

RDF est un langage créé par le W3C dans le but de décrire des métadonnées à l'aide de triplets qui relie deux ressources du web à l'aide d'une propriété étiquetée. Ces triplets sont en relation et forment un graphe. C'est ensuite devenu un format général pour la description conceptuelle ou la modélisation de données. Il est aujourd'hui le standard pour la description de données et connaissances du Web sémantique.

Chaque ressource représentée en RDF est associée à un URI unique permettant son identification. Chaque triplet est composé des éléments suivants :

— **Sujet** : La ressource que l'on souhaite décrire

— **Prédicat** : Le type de propriété applicable au sujet

— **Objet** : Une donnée ou une autre ressource qui est la valeur de la propriété.

Lorsque le sujet ou l'objet sont des ressources, ils peuvent être décrits par un URI ou un noeud anonyme qui ne possède pas d'URI. Le prédicat doit, quant à lui, toujours être identifié par un URI. L'URI est un identifiant unique qui peut être de la forme d'un Uniform Resource Locator (URL) pour être disponible en ligne

28. https://www.w3.org/DesignIssues/Notation3
29. https://www.w3.org/TR/turtle/

comme c'est le cas dans le LOD. Un URI qui permet de localiser une ressource sur le Web est dite *déréférençable*.

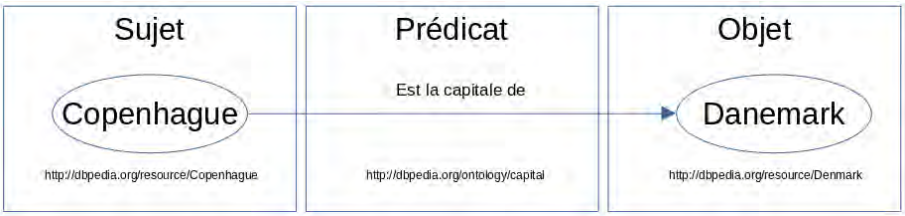


FIGURE 2.11 – Exemple de triplet RDF

L'exemple de la Figure 2.11 représente la relation : "Copenhague est la capitale du Danemark" en utilisant la syntaxe RDF. Le sujet est la ville de Copenhague qui est représentée par l'URI `http://dbpedia.org/resource/Copenhague`. Ce sujet possède la propriété "Est la capitale de" identifiée par l'URI `http://dbpedia.org/ontology/capital`. L'objet de la propriété est le pays Danemark qui est identifié par l'URI `http://dbpedia.org/resource/Danemark`

2.2.2.2 RDFS

La modélisation des connaissances en Web sémantique se fait via des ontologies, qui nécessitent de manipuler explicitement la notion de classe, vue comme un ensemble d'entités regroupées selon des propriétés communes qu'on leur assigne, et les relations d'inclusion entre classes. La représentation formelle d'ontologie requiert en effet de distinguer les ressources jouant le rôle de classes de celles qui renvoient à des éléments de ces classes ou à des entités spécifiques du monde réel. Elle suppose aussi de définir la sémantique de ces nouveaux éléments (ce que va impliquer le fait de définir une ressource comme une classe ou non par exemple). Les langages RDFS et OWL ont été définis dans cet objectif.

La syntaxe RDFS [Brickley & Guha 2014] est un standard du W3C qui est une extension de RDF : c'est langage de définition de schémas de données RDF. La version initiale a été publiée en 1998 et une version finale en 2014. Le vocabulaire RDFS permet d'organiser les données RDF en identifiant explicitement des types de ressources définis a priori, comme "Ressource" (`rdfs:Ressource`), "Class" (`rdfs:Class`) et "Property" (`rdfs:Property`), ainsi que les relations qui existent entres elles grâce à des propriétés comme `rdfs:subClassOf` pour relier des classes, ou `rdfs:domain` et `rdfs:range` pour préciser la sémantique des relations en indiquant le type des entités qu'elles relient. La notion d'instance peut alors être définie grâce à `rdf:type` ; i `rdf:type` C où C `rdf:type rdfs:Class`. Une sémantique est associée à ces ressources, qui permet à un moteur d'inférence de produire de nouveaux triplets à partir d'une base initiale : par exemple la sémantique de `rdfs:subClassOf` est que si i `rdf:type` C et C `rdfs:subClassOf` D alors i `rdf:type` D. Ces éléments ont été jugés suffisants par le W3C pour décrire des ontologies légères. Les limites du RDFS apparaissent pour la description d'ontolo-

gies complexes qui nécessiteraient l'ajout de contraintes sur les relations entre les entités, ou la définition d'axiomes plus complexes.

2.2.2.3 OWL

Le langage OWL a été créé dans le but d'améliorer RDFS. Une première version a été publiée en 2004 et une seconde version en 2009 OWL-2 [30]. OWL permet de représenter des ontologies complexes grâce à un vocabulaire plus varié et grâce aux notions de contraintes et raisonnement. Ce langage permet également d'ajouter des notions de cardinalité, équivalence et contraire entre les entités. dans sa version initiale, OWL se déclinait en trois sous-langages qui sont : OWL-Lite, OWL-DL et OWL-Full.

— **OWL-Lite** : Il s'agit de la version la plus simple d'OWL. Comme RDFS OWL-Lite permet d'établir des relations hiérarchiques entre les entités via les notions de classes et superclasses. Il permet également l'ajout de cardinalités aux propriétés entre les classes mais limitées aux valeurs 0 ou 1.

— **OWL-DL** (Descriptive Logics) : Ce langage est basé sur la logique de description et possède un niveau de d'expressivité plus riche mais aussi une complexité supérieure à OWL-Lite avec un nombre de classes plus important. OWL-DL permet de garder la complétude des calculs, ce qui signifie que toutes les inférences sont garanties calculables.

— **OWL-Full** : Il s'agit de la version la plus développée qui permet une plus grande complexité pour la description des entités. L'un des avantages majeur de cette version est qu'elle est compatible avec RDFS. Ce lien peut s'effectuer via la classe "Thing" de OWL-Full qui est l'équivalent de "Resource" en RDFS.

Dans sa version OWL2, seules deux sémantiques ont subsisté :

— **OWL2-DL** (Description Logics) : On parle dans ce cas de *Sémantique Directe* de OWL2, qui correspond à une interprétation des classes, propriétés et axiomes en logique de description.

— **OWL2-Full** : Cette sémantique s'appuie sur RDF, étend RDFS et considère une ontologie OWL-2 comme un graphe RDF.

A côté de cette dichotomie sémantique, le W3C a prévu plusieurs profils, des implémentations de OWL2 qui assurent chacune un compromis différent entre expressivité et calculabilité, à savoir une bonne capacité de raisonnement : **OWL2-EL**, qui est proche de OWL2-DL, est adapté à la représentation de grandes ontologies en nombre de classes mais avec une axiomatique élémentaire ; **OWL2-RL** est beaucoup plus riche et convient pour des ontologies nécessitant l'expression de règles et d'axiomes ; **OWL2-QL** a été conçu pour faciliter l'interrogation des bases de connaissances en proposant des capacités d'expression proches de celle de RDFS ou encore analogue à une base de données relationnelle.

30. `https://www.w3.org/TR/owl-overview/`

2.2.2.4 Bases de connaissances et graphes de connaissances

Une *base de connaissance* regroupe un ensemble des connaissances d'un domaine sous forme exploitable par des machines. Chaque base comporte une formalisation des concepts et propriétés de ce domaine, formant une ontologie, mais aussi l'ensemble des instances des classes de l'ontologie, aussi appelés individus, et des relations entre ces individus.

Un *triplestore* est un entrepôt de données formalisée, généralement au format RDF. La fonction principale d'un triplestore est similaire à celle d'une base de données relationnelle qui est de stocker un ensemble de données structurées. La différence avec les bases de données relationnelles réside dans le format des données stockées et dans la logique de ce stockage. Un triplestore stocke sous forme de graphes de connaissances formés de triplets RDF qui représentent des vocabulaires, ontologies ou des bases de connaissances. Un entrepôt comporte également contenir des fonctionnalités permettant de raisonner sur les données. Le raisonnement infère de nouvelles connaissances comme de nouvelles relations de subsomption entre des classes, mais aussi elle indique éventuellement la présence d'incohérences au sein de cette base.

2.2.2.5 SPARQL et la notion de *endpoint*

Tout comme Structured Query Language (SQL) dans base de données relationnelles, les triplestores possèdent des possibilités d'interrogation des données qu'ils stockent. SPARQL est le langage de requête standard défini par le W3C pour interroger les bases de connaissances représentées sous forme de graphes RDF. Les entrepôts peuvent être interrogés directement en ligne via un point d'accès (*endpoint*) qui est un service Web permettant de saisir des requêtes SPARQL ou via un programme grâce à l'API de l'entrepôt. On peut ainsi récupérer des données du LOD en vue de leur mise en relation.

La première version de SPARQL, SPARQL 1.0 ne permettait que d'interroger les graphes RDF avec les commandes *SELECT*, *CONSTRUCT*, *ASK* et *DESCRIBE*. Une requête est composée de triplets et retourne un extrait d'un graphe de connaissances de l'entrepôt qui correspond aux critères définis par ces triplets. Comme SQL, SPARQL permet d'organiser le résultat obtenu par la requête via les opérateurs comme *DISTINCT* ou *ORDER BY*.

Depuis la version SPARQL 1.1, il est possible de modifier ou supprimer des données du graphe de connaissances ou bien d'en ajouter de nouvelles. Les modifications du contenu des graphes se font via des commandes comme *INSERT DATA*, *DELETE DATA* ou *CLEAR* et les modifications sur le graphe entier se font via des opérations comme *CREATE*, *DROP* ou *MOVE*.

Dans l'exemple de la Figure 2.12, on interroge un graphe utilisant l'ontologie Friend Of A Friend (FOAF) définie pour décrire une personne physique et ses liens avec d'autres personnes. La requête récupère l'ensemble des noms (désignés par la variable `?name`),issus des triplets de la forme `?person foaf:name ?name`, i.e. liant

```
PREFIX foaf:  <http://xmlns.com/foaf/0.1/>
  SELECT ?name
  WHERE {
      ?person foaf:name ?name .
  }
```

FIGURE 2.12 – Exemple d'une requête SPARQL

une ressource (désignée par `?person`) à un nom via la propriété de l'ontologie FOAF appelée `foaf:name`.

2.2.2.6 Inférence et SWRL

Dans le Web sémantique, l'inférence est la capacité de générer de la connaissance à partir des données existantes. L'inférence se base sur un ensemble de règles qui sont exécutées selon des algorithmes implémentés au sein de moteurs d'inférence. Cette méthode permet de découvrir de nouvelles relations entre les entités ou bien découvrir des incohérences entre les entités qui auraient été enregistrées par les utilisateurs dans les graphes. Les règles d'inférence génériques sont celles liées à la sémantique du langage utilisé. Par exemple, à la propriété `rdfs:subClassOf` est associée une règle qui permet d'inférer que toute entité appartenant à une classe appartient aussi à toutes les classes "mères" de cette classe. Des règles spécifiques peuvent être définies au sein d'une ontologie, et sont alors écrites grâce à des langages de règles comme Semantic Web Rule Language (SWRL). SWRL est un langage de règles combinant les primitives de OWL-DL et un autre langage de règles, Rule Markup Language (RuleML). Le but de ce langage est d'étendre OWL-DL avec les clauses de Haurn réduites aux prédicats unaires et binaires. Une règle SWRL se compose d'un ou plusieurs antécédents et d'une ou plusieurs conséquences. Le moteur de raisonnement considère que si les antécédents d'une règle sont vrais alors les conséquences sont vraies également.

$$hasParent(?x1, ?x2) \wedge hasBrother(?x2, ?x3) \implies hasUncle(?x1, ?x3) \qquad (2.2)$$

L'exemple 2.2 est un exemple de règle sous forme d'équation logique. Si l'antécédent "hasParent" entre x1 et x2 est vrai et l'antécédent "hasBrother" entre x2 et x3 est aussi vrai alors la conséquence "hasUncle" entre x1 et x3 est donc vraie aussi. La syntaxe SWRL pour cet exemple est la suivante :

```
 Implies(Antecedent(hasParent(I-variable(x1) I-variable(x2))
hasBrother(I-variable(x2) I-variable(x3)))
Consequent(hasUncle(I-variable(x1) I-variable(x3))))
```

Certaines de ces règles peuvent également être définies comme des requêtes SPARQL qui seront exécutées automatiquement lors de l'ajout de nouvelles connais-

sances via un raisonneur. Cette action est possible grâce à la commande `INSERT` de SPARQL.

2.3 Conclusion

Ce chapitre a permis de présenter les domaines de la géomatique et du Web sémantique. Tout deux possèdent des normes et standards définis par des organismes internationaux. Ces standards évoluent encore aujourd'hui ainsi que les technologies permettant de les exploiter. La télédétection joue un rôle majeur dans l'étude de la Terre et l'évolution des capteurs va permettre de nouvelles applications dans les années à venir. La mise à disposition des images satellitaires comme données ouvertes a permis d'élargir le champs des exploitants qui n'est plus composé uniquement de chercheurs et industriels. Le rôle du Web sémantique est d'organiser et gérer d'importantes quantités de connaissance. L'enjeu de cette thèse est de définir comment modéliser et exploiter des données géospatiales liées à des des changements calculés sur des images grâce aux technologies du Web sémantique.

Chapitre 3

Etat de l'art

Content

La problématique de la thèse est de montrer l'apport des technologies sémantiques à l'identification, la qualification puis l'utilisation de changements géolocalisés à la surface de la Terre. Ces changements concernent des entités datées et géolocalisées, dont on compare des caractéristiques, afin de détecter si celles-ci ont changé. Le processus de *détection de changement* définit la manière de sélectionner les caractéristiques et de les comparer. La *représentation des changements* s'intéresse à la manière de les stocker en lien avec les images dont ils sont extraits, et en leur associant éventuellement aussi d'autres informations localisées sur les mêmes zones et aux mêmes dates. Cette représentation est dite *sémantique* lorsque les changements sont décrits à l'aide d'une représentation formelle des entités ou des notions qui les caractérisent ainsi que de leurs propriétés. Pour parvenir à une représentation *sémantique* des changements, les technologies et langages du Web sémantique peuvent intervenir à différents niveaux.

Dans ce chapitre sont présentés, dans un premier temps (partie 3.1), des travaux menés dans le cadre du Web sémantique pour proposer des représentations des entités géospatiales et datées. Les composantes spatiales et temporelles sont abordées à travers plusieurs modèles et ontologies que nous présentons dans les parties 3.1.1 et 3.1.2, ainsi que des ressources du Web sémantique disponibles en ligne fournissant des données géospatiales au format RDF (partie 3.1.3). La seconde partie du chapitre (partie 3.2) détaille des travaux traitant le sujet de la thèse, la représentation sémantique de changements géolocalisés et identifiés à partir de données géospatiales. Chacune de ces approches se distingue par la nature

des changements étudiés, la manière dont ils ont été identifiés, le format sous lequel ils sont accessibles, la façon dont est produite leur représentation sémantique et son utilisation. La dernière partie (3.3) présente le positionnement scientifique adopté dans cette thèse en regard des différentes approches présentées ainsi que les choix technologiques associés.

3.1 Géomatique et Web sémantique

Dans cette partie, nous abordons les travaux portant sur la représentation de données géospatiales, datées et localisées à la surface de la Terre grâce aux standards du Web sémantique. Ces éléments peuvent être géolocalisés sous différentes formes : des données associées à des grilles numériques géolocalisées (format raster) ou des données auxquelles on associe des points ou des surfaces géolocalisées (format vecteur). Le temps peut être une information propre à chaque donnée (date de mesure ou de saisie) ou commune à un jeu de données. Afin de représenter des données comportant des propriétés temporelles et spatiales, la solution la plus répandue aujourd'hui est l'utilisation de base de données relationnelles de type PostGIS. Ces bases de données reposent sur la logique du monde fermé et ne possèdent pas une grande flexibilité sur la modélisation contrairement aux technologies du Web sémantique. De plus, chaque base de données est organisée selon son propre schéma, qui n'est pas toujours explicité, ses primitives et son type de coordonnées. Or pour accéder aux données, il faut connaître ce schéma. Enfin, si on interroge plusieurs sources avec des schémas différents, et des systèmes de géolocalisation différents, les différents schémas sont indispensables afin de mettre en correspondance ces données. Les avantages attendus des principes du Web sémantique ont conduit à développer des vocabulaires et des ontologies pour représenter de manière unifiée les informations spatiales et les informations temporelles, à définir des opérations de calcul spatial ou temporel à intégrer dans les raisonnements des raisonneurs logiques, ou encore à produire des données déjà géolocalisées et datées au format RDF, en respectant des standards. Ces vocabulaires peuvent être utilisés pour traduire le contenu de bases de données au format RDF avant de les exploiter, par exemple en vue d'utiliser des sources hétérogènes, ou bien comme couche ontologique unique afin d'unifier l'accès à différentes bases de données dont on ne duplique pas le contenu (on parle d'Ontology-based Data Management) [Ding *et al.* 2020].

Nous avons organisé cette partie en 3 sous-parties : modèles pour représenter la dimension spatiale (partie 3.1.1), modèles pour représenter la dimension temporelle (partie 3.1.2) puis nous exposons des ressources de données géospatiales disponibles sur le LOD (partie 3.1.3).

3.1.1 Représentation sémantique de l'espace

Dans le cadre du web sémantique, la modélisation de données spatiales repose sur des ontologies qui permettent de rendre explicites les primitives utilisées pour représenter les coordonnées ou la dimension spatiales d'entités et les relations entre

ces primitives pour traduire les relations spatiales entre entités. Dans la suite, nous avons choisi de présenter des travaux faisant référence, dans un ordre historique, qui correspond à la prise en compte d'informations plus riches : Region Connection Calculus 8 (RCC8) fait appel à la notion de *région* ; deux recommandations du W3C, Basic Geo et GeoRSS, utilisent des *polygones* ; GeoSPARQL, qui répond aux exigences de l'OGC et GeoSPARQL+.

3.1.1.1 RCC8

Pour pouvoir représenter les différents aspects de l'espace comme la topologie, la forme, la distance via des ontologies, il est nécessaire d'intégrer certaines théories mathématiques de l'espace comme l'algèbre RCC8 développé par [Randell *et al.* 1992]. Ces travaux permettent d'établir des relations entre différentes entités spatiales. Dans l'algèbre RCC8, une connexion entre deux régions a et b est notée C(a,b). Il existe 8 relations topologiques de base :

— Régions déconnectées : $DC(a, b)$

— Régions connectées : $EC(a, b)$

— Régions partiellement superposées : $PO(a, b)$

— La région a est partie tangentielle propre de b : $TPP(a, b)$

— La région a est partie tangentielle non propre de la région b : $NTPP(a, b)$

— La région a est égale à la région b : $EQ(a, b)$

— La région b est partie tangentielle propre de a : $TPP^{-1}(a, b)$

— La région b est partie tangentielle non propre de la région a : $NTPP^{-1}(a, b)$

Ces relations et leurs transitions continues sont illustrées dans la Figure 3.1

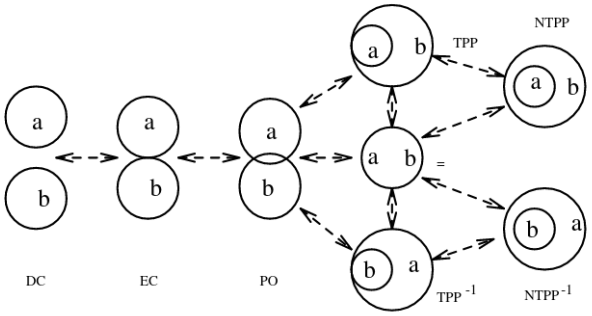


FIGURE 3.1 – Représentation des relations topologiques de l'algèbre RCC8 et leurs transitions selon [Randell *et al.* 1992]

Dans la suite, nous présenterons des ontologies qui reposent sur le modèle RCC8 mais qui peuvent, pour certaines, posséder une convention de nommage différente pour qualifier ces relations topologiques.

3.1.1.2 Basic Geo

Basic Geo est un vocabulaire du W3C créé par [Brickley 2003] qui permet de représenter des coordonnées en utilisant le référentiel WGS84. Ce vocabulaire est élémentaire et ne permet que la représentation de points grâce à leur latitude, longitude et altitude. En revanche, il ne permet pas la représentation de polygones ou de lignes et n'implémente pas l'algèbre RCC8. Le but principal de ce vocabulaire est de pouvoir positionner un objet sur une carte sans fournir plus d'information.

3.1.1.3 GeoRSS

Le modèle GeoRSS [1] est une recommandation du W3C initialement développée pour représenter des objets géographiques dans un flux Really Simple Syndication (RSS) et aussi de géolocaliser ce dernier [Lieberman *et al.* 2007]. Aujourd'hui, GeoRSS se décline en deux modèles : GeoRSS-Simple et GeoRSS-GML et dispose de plusieurs encodages. Parmi ces encodages, on note RDF et OWL ce qui permet une intégration complète de ce modèle dans le Web Sémantique.

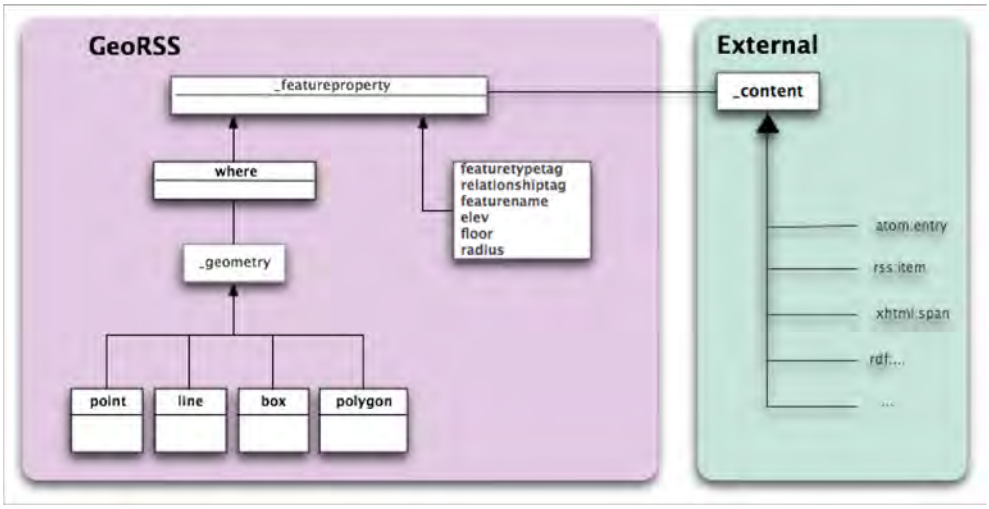


FIGURE 3.2 – Vue graphique du modèle GeoRSS-Simple [Lieberman *et al.* 2007]

GeoRSS-Simple : Il s'agit d'un modèle très léger pour la description de données géo-référencées. Ce modèle est illustré dans la Figure 3.2. La relation générique *_featureproperty* peut être utilisée pour définir l'emprise spatiale d'une ressource externe appelé *_content* grâce à la spécification d'une géométrie. La relation *where* est une spécialisation de *_featureproperty*. Elle est utilisée pour associer l'entité spatiale *_content* à un type de géométrie *_geometry* pouvant être un des types spatiaux de base : *point*, *ligne*, *rectangle*, *polygone*. Les coordonnées sont exprimées par la latitude/longitude conformément à WGS84.

GeoRSS-GML : Ce modèle se base sur GML présenté dans la partie 2.1.3.3 dont il permet de représenter un ensemble complet des géométries avec différents

1. `https://www.w3.org/2005/Incubator/geo/XGR-geo/#model`

systèmes de coordonnées. Les travaux de [Miron 2009] ont débouché sur une représentation au format OWL d'une extension de ce modèle, débouchant sur l'ontologie ONTOAST présentée sur la Figure 3.3.

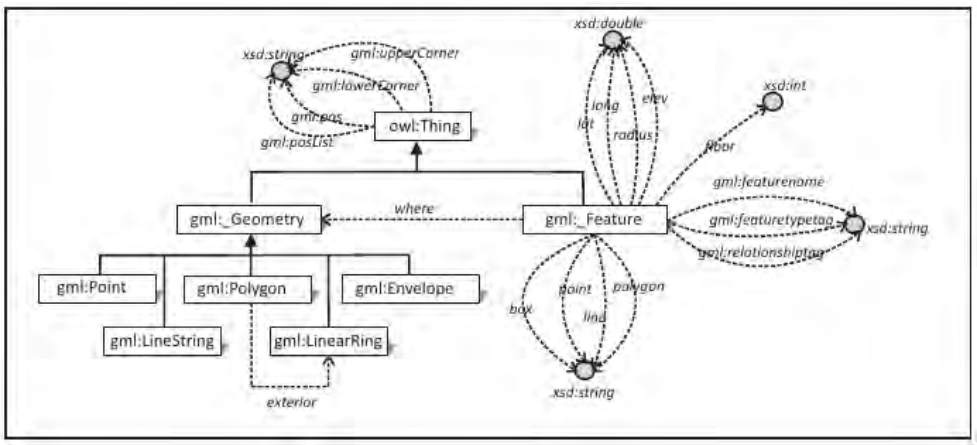


FIGURE 3.3 – Modèle GeoRSS-GML au format OWL selon [Miron 2009]

Dans cette représentation, les géométries sont attachées aux objets appelés `gml:_Feature` via des attributs `box`, `line` et `polygon` par la classe abstraite `gml:_Geometry`. Ces polygones sont représentés par des chaînes de caractères contenant les coordonnées des points arrêtes du polygone (latitude et longitude séparées par un espace). L'attribut `elev` permet de représenter l'altitude d'un point en mètres et l'attribut `radius` permet, au format WGS84, de représenter en mètres un rayon autour de la géométrie d'un objet.

3.1.1.4 GeoSPARQL

Pour représenter sémantiquement des données ayant une composante spatiale, ou données géolocalisées, le vocabulaire GeoSPARQL [2] est le standard actuellement le plus utilisé. En effet, GeoSPARQL est à la fois un vocabulaire pour représenter des données géolocalisées dans le Web sémantique, et une extension du langage de requête SPARQL pour interroger ces données et produire des raisonnements utilisant la dimension spatiale. Développé par l'OGC en 2011 [Perry *et al.* 2012], il est régulièrement maintenu et implémenté dans plusieurs triple-stores. Ce standard permet de représenter et d'interroger une grande variété de données géospatiales comme des polygones, des points ou des lignes en RDF. Il gère également une grande quantité de systèmes de coordonnées comme le WGS84. Cette ontologie a aussi pour but de lier entre elles des données géospatiales grâce à des relations spatiales. En devenant un standard, elle facilite l'interopérabilité entre des jeux de données hétérogènes [Battle & Kolas 2012, Kolas *et al.* 2013]. La Figure 3.4 représente les classes les plus importantes du modèle GeoSPARQL.

2. `https://www.ogc.org/standards/geosparql`

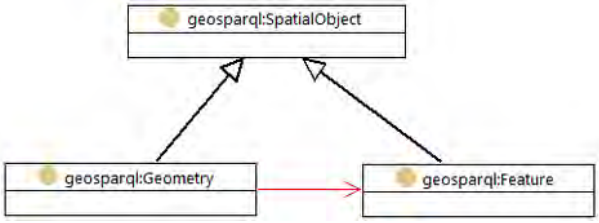


FIGURE 3.4 – Modèle GeoSPARQL selon le W3C

La classe `geo:Feature` peut être instanciée avec n'importe quelle entité et permet ainsi de la géo-référencer si ses informations sont connues. Cette classe est associée à la classe `geo:Geometry` via la propriété `geo:hasGeometry`. Les instances de la classe `geo:Geometry` contiennent les représentations textuelles des géométries au format WKT. Ce modèle permet également de définir des relations topologiques entre les entités comme `geo:sf-contains`, `geo:sf-intersects` ou `geo:sf-overlaps`. Grâce à ces propriétés, le moteur de raisonnement associé à GeoSPARQL peut inférer de nouvelles connaissances. Ce processus se base sur les relations connues pour déduire de nouvelles relations automatiquement. L'équation 3.1 montre un axiome réalisé par [Alirezaie *et al.* 2017] dans le vocabulaire `ontocity` qu'il a défini. Cet axiome crée un segment grâce à la propriété `intersects` qu'il a défini comme héritant de la propriété `geo:sf-intersects` de GeoSPARQL. Cet axiome définit un segment est défini lorsqu'un élément rectangulaire est en intersection avec une région. L'espace de nom `geos` fait ici référence à l'ontologie GeoSPARQL.

$$\begin{aligned} \texttt{ontocity:Segment} \sqsubseteq (&texttttgeos : Feature \sqcap \\ &\exists\ \texttt{geos:hasGeometry.geos:Rectangle} \sqcap \\ &\exists\ \texttt{ontocity:intersects.ontocity:Region}\ ) \end{aligned} \tag{3.1}$$

Aujourd'hui une grande majorité des triplestores comme Virtuoso, Stardog ou Fuseki intègrent le standard GeoSPARQL comme le montre l'étude faite par [Jovanovik *et al.* 2021]. Ces triplestores n'implémentent pas tous les mêmes fonctionnalités, en particulier pour raisonner avant d'exécuter une requête SPARQL et ne réagissent pas de la même manière face à un volume important de données géolocalisées.

3.1.1.5 stRDF et stSPARQL

Dans [Koubarakis & Kyzirakos 2010], il est présenté une alternative à GeoSPARQL pour la représentation de données géospatiales avec le modèle stRDF (pour Space and Time RDF) et d'interrogation avec stSPARQL (pour Space and Time SPARQL). Ces deux modèles ont été mis au point dans le cadre d'un projet européen à la même période où SPARQL a été défini. Ce modèle n'est pas standardisé mais il permet, tout comme GeoSPARQL, de représenter les relations topologiques issues du modèle RCC8 entre différentes entités géoréférencées. L'implémentation de

ces extensions sont disponibles dans le triple-store Strabon [Kyzirakos *et al.* 2012] développé par la même équipe.

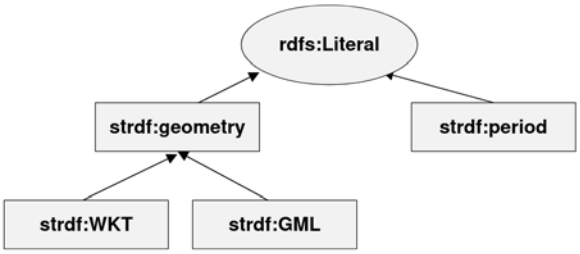


FIGURE 3.5 – Extrait du modèle stRDF selon [Koubarakis & Kyzirakos 2010]

Comme montré dans la Figure 3.5, stRDF permet de représenter des géométries en WKT et GML via les types `strdf:WKT` et `strdf:GML` sous forme de chaînes de caractères (`rdfs:Literal`). stSPARQL est une extension de SPARQL permettant d'interroger des graphes spatio-temporels. En plus d'intégrer la composante spatiale, stRDF intègre également la composante temporelle avec la classe `strdf:Period`.

D'après [Perry *et al.* 2012], stRDF/stSPARQL et GeoSPARQL ont en commun plusieurs fonctionnalités : dans les deux cas, les géometries sont représentées sous forme de `rdfs:Literals` à l'aide de types de données spatiaux, qui font appel aux géométries WKT et GML ; dans les deux cas, les mêmes familles de fonctions sont disponibles pour faire des requêtes sur les géométries. GeoSPARQL offre certaines fonctionnalités non disponibles en stSPARQL, comme le fait de s'inspirer de la terminologie GIS ou le fait de pouvoir expliciter l'axiomatique des relations topologiques et ainsi de pouvoir raisonner sur ces relations. De plus, GeoSPARQL dispose d'un mécanisme de réécriture de requête. A l'inverse, stSPARQL propose des fonctions non fournies par GeoSPARQL comme l'agrégation géospatiale et une meilleure gestion de la dimension temporelle.

3.1.1.6 GeoSPARQL+

Plus récemment, [Homburg *et al.* 2020] a proposé une extension de GeoSPARQL appelée GeoSPARQL+ afin de représenter un fichier raster au format RDF. Le but de cette extension est de considérer un fichier raster comme un nouveau type de donnée géospatiale dans le Web sémantique. L'ontologie GeoSPARQL+ spécialise l'ontologie GeoSPARQL présentée dans la section 3.1.1.4. La classe principale de ce vocabulaire est la classe `geo2:Raster` qui est elle-même une spécialisation de la classe `geo2:Coverage`. La classe `geo2:Coverage` est une sous-classe de la classe GeoSPARQL `geo:SpatialObject`. En plus de pouvoir définir des couvertures de zones géographiques sous forme de polygones, il est possible d'associer une couverture sous forme d'un ou de plusieurs rasters via la propriété `geo2:hasCoverage`. Cette couverture a une représentation au format CoverageJSON avec la propriété `geo2:asCoverageJSON`. Le format CoverageJSON est un standard du W3C et de l'OGC peu utilisé aujourd'hui [Blower *et al.* 2017]. GeoSPARQL+ permet aussi

d'effectuer certaines opérations sur les rasters ainsi que de définir des filtres pour l'interrogation.

3.1.2 Représentation sémantique du temps

La plupart des logiques temporelles et des travaux formalisant des propriétés temporelles reprennent l'algèbre des intervalles de Allen [Allen 1983], dont nous présentons les grandes lignes dans la partie suivante. L'ontologie OWL-Time, devenue un standard du W3C pour la représentation du temps dans des graphes de connaissances, en reprend les relations. En revanche, l'ontologie Semantic Web for Earth and Environmental Terminology (SWEET) propose des relations temporelles plus élémentaires adaptées de l'algèbre RCC8. Nous présentons ici ces deux ontologies.

3.1.2.1 Algèbre des intervalles temporels d'Allen

Afin de pouvoir raisonner sur le temps, la théorie la plus répandue est celle présentée par [Allen 1983]. Tout comme le modèle RCC8 présenté dans la section 3.1.1.1 sur la dimension spatiale, Allen propose d'établir des relations entre différents intervalles temporels. Cet algèbre ne prend pas en compte la modélisation du concept d'instant mais uniquement des intervalles. L'ensemble de ces relations est utilisé dans plusieurs domaines et logiques de l'intelligence artificielle pour produire les raisonnements temporels grâce à l'union ou l'intersection entre intervalles.

X Y	X before Y, Y is after X
X Y	X meets Y, Y is met by X
X Y	X overlaps with Y, Y is overlapped by X
X Y	X starts Y, Y is started by X
X Y	X during Y, Y contains X
X Y	X finishes Y, Y is finished by X
X Y	X is equal to Y

FIGURE 3.6 – Représentation des 13 relations entre intervalles temporels selon [Allen 1983]

Grâce à la Figure 3.6, on peut voir que deux intervalles satisfaisant une relation ne peuvent en satisfaire une seconde.

3.1.2.2 OWL-Time

Pour représenter une entité en précisant sa dimension temporelle, le standard le plus répandu est l'ontologie OWL-Time [3] créée en 2006 et présentée pour la première fois dans [Hobbs & Pan 2004]. Cette ontologie est un standard reconnu par l'OGC et le W3C. Elle est régulièrement mise à jour et sa dernière version date de mars 2020. Il est possible de représenter des relations de l'algèbre des intervalles d'Allen présentées dans la section 3.1.2.1 grâce à des propriétés comme `time:intervalContains`, `time:intervalOverlaps` et `time:before`.

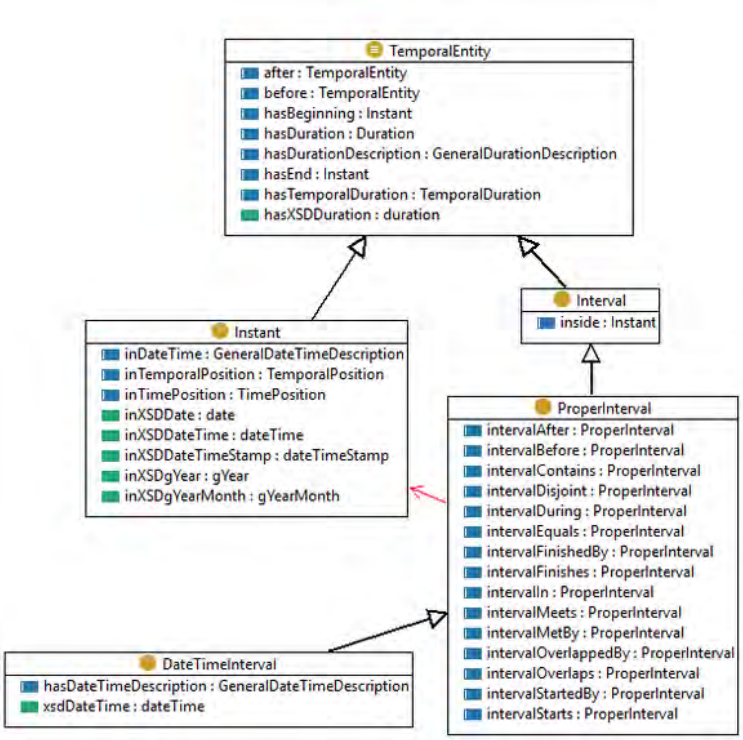


FIGURE 3.7 – Extrait de l'ontologie OWL-Time selon le W3C

Sur la Figure 3.7, on remarque que la classe principale est la classe abstraite `time:TemporalEntity` pour représenter des instants et des intervalles. Les instants sont représentés via la propriété `time:inXSDDateTimeStamp` au format `xsd:dateTimeStamp` [4]. Les intervalles sont définis par deux instants via les propriétés `time:hasBeginning` et `time:hasEnd`.

3.1.2.3 SWEET : Semantic Web for Earth and Environmental Terminology

En marge de l'ontologie standard OWL-Time, il existe d'autres ontologies permettant la description de données temporelles. Nous avons vu le modèle stRDF

3. `https://www.w3.org/TR/owl-time/`
4. `https://www.w3.org/TR/xmlschema11-2/#dateTimeStamp`

dans la section 3.1.1.5 qui possède une composante temporelle. De même, l'ontologie SWEET, développée par la NASA, permet de décrire des concepts sur une dimension spatio-temporelle. Cette ontologie a été créée initialement dans le but de décrire des processus scientifiques ainsi que leurs données associées [Raskin 2005]. La partie temporelle de ce modèle permet de représenter des dates et des durées comme une saison, une année ou un instant. Certaines relations temporelles issues de l'algèbre RCC8 comme `before` ou `after` sont disponibles dans cette ontologie.

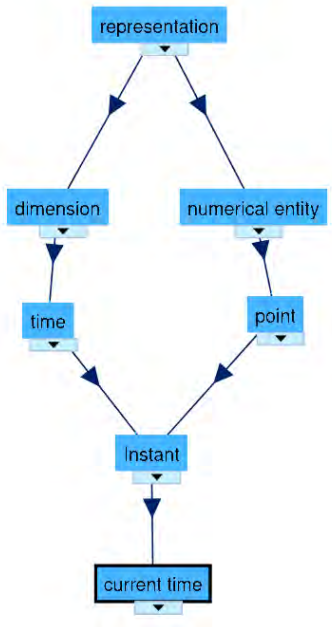


FIGURE 3.8 – Représentation d'une partie des concepts de l'ontologie SWEET (source : bioportal.bioontology.org)

La Figure 3.8 montre une partie des classes disponibles dans l'ontologie SWEET pour représenter une entité temporelle. Ici, la classe `current time` est représentée via une classe de plus haut niveau appelée `instant`. Le concept de plus haut niveau dans cet extrait est le concept `representation` qui peut être spécialisé en `dimension` ou `numerical entity`. La hiérarchisation des classes temporelles permet une description plus ou moins précise de l'information temporelle associée à des données. Cette ontologie, qui n'est pas reconnue comme standard par le W3C, est aujourd'hui toujours en développement actif [5].

3.1.3 Présentation de ressources géolocalisées du LOD

Plusieurs travaux, aux objectifs et motivations variés, visent à contribuer au LOD en fournissant aux utilisateurs un ensemble de données géolocalisées compatibles avec les technologies du Web sémantique. Pour cela, ils réutilisent les ontolo-

5. `https://github.com/ESIPFed/sweet`

gies et modèles présentés précédemment. Leur but est de publier leurs résultats en ligne sous forme d'entrepôts de données sémantiques ou dans le LOD.

3.1.3.1 GeoNames

GeoNames [6] est une base de connaissances géographiques disponible dans le LOD. Cette base fournit diverses informations géographiques et le nom dans plusieurs langues de lieux sur la Terre, et permet de retrouver ces informations à partir de leur nom. GeoNames utilise une localisation de chaque lieu par un point dont les coordonnées correspondent à la latitude et à la longitude conformément au standard WGS84. Les données sont disponibles au format RDF et accessibles via leur site Web ou des Application Programming Interface (API). L'ontologie décrivant les données est proposée en ligne [7]. Cette ontologie réutilise les concepts des ontologies Basic Geo et GeoSPARQL pour la localisation des entités. GeoNames se base sur plus de 100 sources comme la Société Nationale des Chemins de Fer français (SNCF) ou la National Oceanic and Atmospheric Administration (NOAA) aux États-Unis d'Amérique afin d'obtenir les noms et la localisation de lieux dans chaque pays, mais aussi, lorsque cela est disponible et/ou pertinent, des informations géographiques complémentaires comme l'élévation, la population, etc. Les utilisateurs peuvent les corriger ou les compléter de manière coopérative via une interface wiki très intuitive.

3.1.3.2 LinkedGeoData

Le projet LinkedGeoData [8] a pour but d'ajouter de la connaissance avec une dimension spatiale aux données déjà disponibles dans le LOD [Stadler *et al.* 2012]. Pour ce faire, les auteurs ont créé une base de connaissances à partir des données disponibles dans OpenStreetMap (OSM). Ces données sont disponibles au format RDF et peuvent être directement téléchargées depuis le site Web ou interrogées via un endpoint SPARQL. La base de connaissances complète contient environ 20 milliards de triplets RDF. Le choix a été fait de stocker l'ensemble des données OSM dans une base de données relationnelle et d'exécuter ensuite un script appelé Sparqlify [9] pour la génération de triplets RDF. Comme pour GeoNames, les données sont localisées par un point dont les coordonnées (latitude, longitude) sont représentées en WGS84 via l'ontologie Basic Geo.

3.1.3.3 Yago2Geo

Yago2Geo [10] est un autre entrepôt de données géospatiales accessible en ligne. Il s'agit d'une extension du projet Yago2 [11] développé par [Hoffart *et al.* 2013] qui a

6. `https://www.geonames.org/`
7. `https://www.geonames.org/ontology/ontology_v3.2.rdf`
8. `http://linkedgeodata.org`
9. `https://github.com/SmartDataAnalytics/Sparqlify`
10. `http://yago2geo.di.uoa.gr/`
11. `https://yago-knowledge.org/`

permis de créer tous les outils pour construire puis mettre à jour une base de connaissances à partir de Wikipedia. Yago2Geo est une base de connaissances qui reprend les principes de Yago2 mais ne contient que des entités géolocalisées, en particulier les unités administratives d'une majorité de pays du monde [Karalis *et al.* 2019]. Les données de cet entrepôt proviennent de diverses sources gouvernementales ainsi qu'OSM et Global Administrative Areas (GADM) [12], une base de données qui a pour objectif de répertorier et localiser toutes les entités administratives sur Terre, à tous les niveaux de détail. La géolocalisation est représentée grâce à des polygones correspondant à l'emprise au sol de chaque entité. Les données sont accessibles sous forme de triplets RDF via un endpoint SPARQL [13] ou à télécharger depuis leur site Web. L'ontologie définie pour la description des concepts réutilise l'ontologie GeoSPARQL pour la représentation des polygones. La génération de triplets RDF se fait à l'aide d'une application Java [14]. L'avantage majeur de cette ressource est qu'elle offre une localisation précise des entités géographiques grâce à des polygones pour représenter leurs frontières (physiques ou administratives) alors que beaucoup d'autres graphes de connaissances (comme DBPedia, GeoNames, etc.) utilisent des points pour localiser les entités.

3.2 Approches sémantiques pour l'étude de changements à la surface de la Terre

Cette section présente les travaux qui étudient l'identification automatique de changements puis leur représentation avec les langages et vocabulaires du Web sémantique, et cela en distinguant deux types de données géolocalisées : les données géospatiales liées à des changements d'une part, et les images satellitaires d'autre part. Cette notion de changement peut se représenter par les concepts d'évènements, d'évolution ou de tendances. Dans ces travaux, les données géospatiales ainsi que les images satellitaires sont une partie des connaissances qui permettent à des utilisateurs ou des experts la prise de décisions. L'approche sémantique permet d'obtenir une représentation homogène de l'ensemble des connaissances disponibles pour chaque application. Pour chaque approche présentée, nous indiquons la nature des changements étudiés, les types de données et les méthodes exploités pour identifier les changements ; nous identifions aussi les vocabulaires utilisés pour décrire les données géospatiales (données faisant l'objet de changements et/ou les changements eux-mêmes), l'approche de sémantisation retenue (c'est-à-dire les algorithmes utilisés pour produire les représentations sémantiques à partir des données sources) et comment les changements sont exploités. Nous indiquons aussi leur domaine d'application.

12. `https://gadm.org`
13. `http://test.strabon.di.uoa.gr/yago2geo`
14. `https://github.com/nkaralis/Yago_Extension`

3.2.1 Représentation de changements sur des données géospatiales grâce au Web sémantique

Cette section présente différents travaux sur la représentation sémantique de données géospatiales pour lesquelles on s'intéresse aux changements, et sur la représentation de ces changements. Ces données peuvent être variées comme les observations faites par une station météorologique dans [Baucic & Medak 2014] ou bien des données issues du trafic routier [Ding *et al.* 2020]. L'objectif est de modéliser ces données en réutilisant des standards du Web sémantique et de générer une représentation au format RDF pour pouvoir les exploiter. La notion de changement peut prendre la forme d'un évènement ponctuel dans le temps comme une inondation [Wang *et al.* 2018] ou bien la forme d'une évolution territoriale historique à travers de multiples données comme dans les travaux de [Kauppinen *et al.* 2008]. D'autres travaux comme [Blázquez *et al.* 2012] et [Smeros & Koubarakis 2016] cherchent à découvrir de nouvelles connaissances grâce aux données déjà connues via le processus d'annotation sémantique. Les travaux présentés par la suite montrent différentes approches relatives à la modélisation de données géospatiales et sont proches de notre thématique de recherche.

3.2.1.1 Des données liées sur l'évolution de la forêt amazonienne [Kauppinen *et al.* 2013]

Les auteurs de [Kauppinen *et al.* 2013] cherchent à améliorer le temps de collecte et de recherche d'informations sur la forêt amazonienne localisée au Brésil. Leur second objectif est de lier les différentes sources de données avec des formats hétérogènes pour les mettre à disposition de la communauté scientifique dans un format unique (RDF). Le phénomène étudié via les changements est la déforestation. Une partie des données utilisées provient du ministère de l'environnement brésilien. Une autre partie provient de l'institut géographique et de statistique du Brésil ainsi que d'autres organismes gouvernementaux. Les auteurs utilisent également un raster de déforestation généré à partir des images satellitaires provenant du programme Landsat Thematic Mapper [15] entre 1997 et 2007.

L'application proposée permet de télécharger les données statistiques depuis les sites gouvernementaux brésiliens et de les associer aux données Landsat Thematic Mapper. Ces données images possèdent une résolution de 30 m mais les auteurs ont choisi de découper leur zone d'étude en une grille composée de 8580 cellules de 25 km x 25 km pour couvrir une plus grande partie de la forêt amazonienne. Chaque donnée statistique est associée à une cellule de la grille.

Afin de formaliser l'ensemble de ces données au format du Web Sémantique, les auteurs ont choisi de développer leur propre ontologie appelée Open Linked Amazon (OLA) Vocabulary. Cette ontologie repose sur une seconde ontologie, Open Time and Space Core Vocabulary (TISC) [16], que cette équipe a développée auparavant

15. `https://landsat.gsfc.nasa.gov/landsat-4-5/tm`
16. `http://observedchange.com/tisc/ns/`

pour les dimensions spatiales et temporelles. Ce vocabulaire contient des propriétés similaires aux ontologies standards présentées précédemment comme la propriété "touches" mais n'est pas autant utilisée que celles-ci par la communauté. Ce vocabulaire a été établi en 2011 et a été mis à jour pour la dernière fois en 2013. Le résultat est un jeu de données unique au format RDF appelé Linked Brazilian Amazon Rainforest Data (LBARD).

3.2.1.2 Prévention d'incendie à l'aide d'images satellitaires et de données liées [Kyzirakos *et al.* 2014a]

Dans [Kyzirakos *et al.* 2014a], les auteurs montrent qu'il est possible d'associer de la connaissance aux images satellitaires. La problématique est ici de pouvoir valider, en utilisant des données contextuelles, la détection de points chauds sur la surface de la Terre provenant de l'analyse des images en vue de prévenir des risques d'incendies. Les auteurs présentent un service développé dans le cadre du projet européen TELEIOS [17] qui s'est déroulé de 2010 à 2013. TELEIOS avait pour but d'extraire de la connaissance à partir des données d'observation de la Terre par satellite, entre autres les images issues du satellite MSG/SEVIRI [18] qui possède une résolution spatiale de 3 km. L'avantage majeur de ce satellite est qu'il possède une résolution temporelle de 5 à 15 min ce qui permet d'avoir presque en temps réel des vues sur l'Europe. Les images sélectionnées sont stockées dans une base MonetDB où chaque pixel est une entrée de la base. La seconde source de données utilisée provient du LOD avec l'entrepôt linkedGeoData [19] qui possède les données shapefile OSM au format RDF. Enfin, l'index géographique GeoNames [20] est utilisé pour obtenir des informations sur un lieu donné. La dernière source de données, fournie par le gouvernement grec, contient les unités administratives grecques au format shapefile.

Afin d'homogénéiser l'ensemble de ces données, les auteurs ont fait le choix de développer leur propre ontologie appelée National Observatory of Athens (NOA) Ontology. Ce vocabulaire repose sur le modèle stRDF présenté précédemment ainsi que sur l'ontologie SWEET présentée dans la section 3.1.2.3. Pour valider leur approche, ils ont choisi un cas d'étude sur la prévention des feux de forêts en Grèce.

Le principe est dans un premier temps de détecter l'ensemble des points chauds obtenus par les pixels de l'image et sémantisés via une requête SPARQL. Ces points sont ensuite filtrés selon les géométries des unités administratives. Une seconde requête va éliminer les points chauds qui se trouvent dans des zones définies comme 'exclues' comme l'océan ou les côtes.

Le résultat final est un ensemble de points chauds possédant chacun un indice de confiance qui a été calculé en fonction de la couverture du terrain ainsi que la direction du vent et la géomorphologie du terrain. Ces points chauds sont représentés

17. `https://cordis.europa.eu/project/id/257662`
18. `https://www.esa.int/esapub/bulletin/bullet111/chapter4_bul111.pdf`
19. `http://linkedgeodata.org`
20. `http://www.geonames.org/`

au format RDF et sont ensuite intégrés dans un système en ligne permettant de les visualiser sur une carte. Cette interface permet également aux utilisateurs finaux d'effectuer des requêtes. Ce service a été utilisé pendant la saison des feux de forêts en Grèce entre 2012 et 2013 afin de définir des stratégies de lutte contre ces incendies par l'agence de protection civile, les pompiers et l'armée grecque.

3.2.1.3 Un modèle spatial pour découvrir la connaissance dans les jeux de données géospatiaux [Harbelot *et al.* 2015]

Les travaux présentés dans [Harbelot *et al.* 2015] ont pour but d'introduire un modèle sémantique permettant de découvrir de la connaissance à partir de données parcellaires. La problématique est d'identifier des motifs et extraire de la connaissance à partir d'une grande quantité de données. Ce modèle permet notamment d'analyser des phénomènes dynamiques à l'aide de données spatiales, temporelles et thématiques.

Le modèle présenté, appelé Land Cover Change Continuum (LC3), permet de représenter des entités dynamiques qui évoluent dans le temps. Ces entités sont nommées `Timeslices` (tranches temporelles) et sont définies selon quatre composantes : l'identité, l'espace, le temps et la sémantique de l'entité. Le concept de `Timeslice` est un concept de haut niveau qui peut être spécialisé. Chaque `Timeslice` est datée par un instant représenté par la classe `TemporalPoint` ou un intervalle via la propriété `hasTime`, et possède une géométrie grâce à la propriété `hasGeometry`.

Les transitions entre différentes entités sont modélisées à l'aide de la propriété `hasFiliation`. Cette propriété générale peut être spécialisée en deux propriétés : `hasContinuation` ou `hasDerivation`. Ces deux propriétés sont à leur tour spécialisées pour obtenir un niveau de précision supplémentaire pour décrire la transition. Les relations les plus spécialisées sont les suivantes : `expansion, contraction, division, séparation, fusion, annexion, stabilité, conversion`.

Grâce aux motifs propres à chaque propriété, les auteurs sont parvenus à établir des règles SPARQL 1.1 permettant l'inférence d'un phénomène. Ils ont choisi de tester leur théorie en appliquant les motifs sur un jeu de données situé en Gironde dans le sud de la France. Les données proviennent des rasters de classification Corine Land Cover pour les années 1990, 2000 et 2006. Chaque classe de couverture du sol est représentée par une `TimeSlice` (ts).

La Figure 3.9 montre les différents évènements qui ont pu être inférés grâce à leurs règles. `t1` représente l'année 1990, `t2` l'année 2000 et `t3` l'année 2006. Trois types de phénomènes ont été définis auparavant par des règles afin d'être détectés dans le jeu de données : intensification urbaine, inondation et déforestation. L'intensification urbaine est représentée par le passage de `ts3` à `ts5`. La déforestation est identifiée par la disparition de la forêt entre `ts1` et `ts3`.

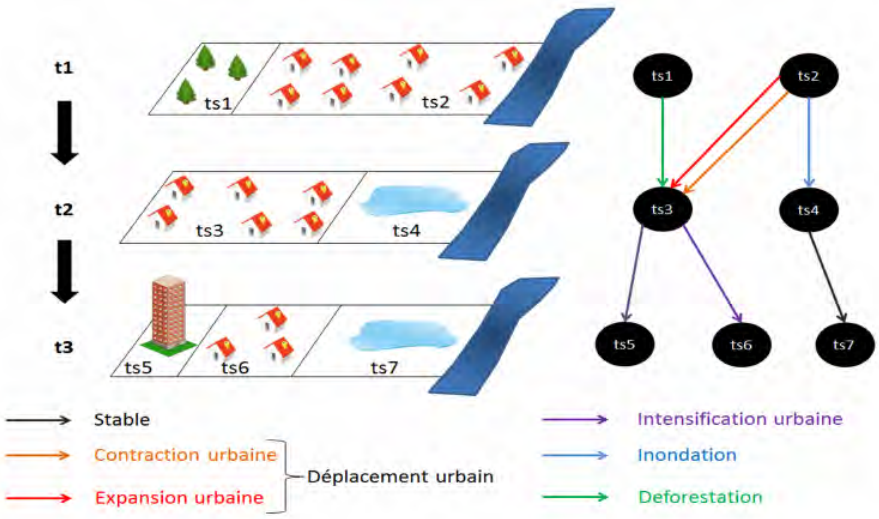


FIGURE 3.9 – Liste des phénomènes inférés dans le modèle LC3 par [Harbelot *et al.* 2015]

3.2.1.4 Etude sémantique de l'évolution d'unité territoriales [Bernard *et al.* 2018]

Les travaux de [Bernard *et al.* 2018] présentent deux ontologies pour la représentation des territoires administratifs au cours du temps. La première ontologie appelée Territorial Statistical Nomenclature (TSN) permet de représenter tout type d'unité territoriale et la seconde appelé TSN-Change permet de représenter les différents changements survenus entre ces territoires. Ces changements sont répertoriés grâce à la notion de version disponible dans l'ontologie proposée. Chaque version d'un territoire appartient à une nomenclature et possède une date.

Ces ontologies n'ont pas pour but de représenter des concepts géométriques comme l'ontologie GeoSPARQL mais des concepts géographiques comme des villes, des régions ou des départements. Ce modèle repose sur plusieurs vocabulaires standards comme Dublin Core Metadata Initiative (DCMI) [21] et Provenance, Authoring and Versioning (PAV) [22] pour la gestion des versions.

La Figure 3.10 montre les principaux concepts proposés pour décrire une unité territoriale avec l'ontologie TSN. On remarque que le concept de `Nomenclature` est directement rattaché à une version via la propriété `hasVersion`. Les concepts `TerritoryVersion` et `UnitVersion` sont des sous-classes de la classe `geo:Feature` de l'ontologie GeoSPARQL ce qui permet que ces concepts soient georéférencés via une géométrie.

L'ontologie TSN-Change, pour la description des changements, comporte un concept de haut-niveau `Change` spécialisé en concepts plus spécifiques pour distin-

21. `https://dublincore.org/specifications/dublin-core/dcmi-terms/`
22. `http://pav-ontology.github.io/pav/pav.rdf`

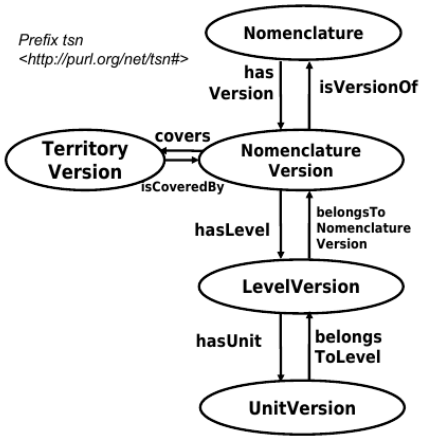


FIGURE 3.10 – Principaux concepts de l'ontologie TSN par [Bernard *et al.* 2018]

guer la représentation des changements concernant la structure (comme une fusion ou une séparation d'unités territoriales) de celle des changements concernant les caractéristiques (comme un changement de nom d'unité territoriale par exemple). Chaque changement est décrit par cinq propriétés représentées dans un tuple comme suit : `<t, INPUT, OUTPUT, causes, isCausedBy>`. `t` représente l'instant ou la période du changement, `INPUT` est l'entité qui a subi le changement, `OUTPUT` celle qui résulte de ce changement. `causes` est un ensemble d'entités de type `Change` qui ont causé le changement. `isCausedBy` représente l'entité `Change` de plus haut niveau qui a causé le changement. La qualification des changements comme `Scission` ou `Disappearance` est définie par des règles du modèle.

L'expérimentation de ce modèle utilise le jeu de données ouvert fourni par la commission européenne : Nomenclature des Unités Territoriales Statistiques (NUTS) [23]. Les données exploitées sont celles disponibles pour les années 1999, 2003, 2006 et 2010 au format vecteur. La technologie RDB to RDF Mapping Language (R2RML) permet d'effectuer l'alignement entre ces données et le modèle TSN-Change grâce à un fichier de correspondances. La génération de graphes RDF est effectuée avec l'application GeoTriples [24] développée par [Kyzirakos *et al.* 2014b].

Après avoir généré l'ensemble des graphes sur les données, les changements sont extraits grâce à des requêtes SPARQL utilisant l'opérateur `DESCRIBE`. Ces requêtes retournent les valeurs des propriétés relatives à chaque changement. Par exemple, une des requêtes présentées dans leurs travaux identifie un changement de type `StructureChange` entre 1999 et 2003 pour l'unité territoriale `ES63`.

23. `https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/nuts`
24. `http://linkedeodata.github.io/GeoTriples/`

3.2.2 Exploitation sémantique d'images satellitaires

Cette section a pour but de présenter des travaux alliant la télédétection avec les technologies du Web sémantique. L'enjeu est de montrer en quoi ces deux domaines peuvent être complémentaires. Le Web sémantique peut intervenir pour plusieurs applications comme la classification des images [Gu *et al.* 2017] ou l'étude des sols [Wang *et al.* 2015] ou encore pour construire un catalogue sémantique pour les images satellitaires [Lin *et al.* 2016], [Augustin *et al.* 2018] ou [Espinoza-Molina *et al.* 2015]. Dans ce cas, les données sémantiques servent à indexer et à rechercher les images via des requêtes en vue de les exploiter. Une autre approche comme [Bouyerbou *et al.* 2019] et [Neptune 2020] consiste à extraire de l'information depuis les images grâce à des algorithmes de détection de changement ou de classification et représenter le résultat sémantiquement. Ces travaux partagent des objectifs similaires au sujet d'étude porté par la thèse puisqu'ils proposent une exploitation sémantique des images satellitaires au service de la détection de changements. Nous en présentons plusieurs dans cette partie.

3.2.2.1 Gestion de catastrophe naturelle grâce à la géolocalisation sémantique à partir d'images satellitaires [Alirezaie *et al.* 2017]

Le framework présenté par [Alirezaie *et al.* 2017], SemCityMap, a pour but d'ajouter à des images satellitaires de la connaissance géolocalisée, représentée à l'aide d'ontologies, puis d'exploiter des requêtes SPARQL et le raisonnement sur les connaissances afin de calculer des trajets entre les lieux identifiés sur ces images. L'article décrit le principe du framework SemCityMap, et en illustre l'utilisation avec une simulation d'inondation dans la ville de Stockholm.

Afin d'organiser et définir les différents concepts nécessaires pour représenter des lieux et des catastrophes pouvant survenir au sein d'une ville, une nouvelle ontologie, `OntoCity`, a été élaborée à partir de différents standards existants. Cette ontologie est une extension de GeoSPARQL et réutilise DOLCE+DnS Ultra Lite (DUL) [25] pour représenter le concept d'évènement avec la classe `DUL:Event`.

La classe `geo:Feature` de GeoSPARQL a été spécialisée par trois sous-classes : `Area`, `Region` et `Segment`, comme montré dans la Figure 3.11. Les relations spatiales sont représentées grâce à la propriété `ontocity:hasSpatialRelation` entre deux entités de type `geo:Feature`.

Ce framework a été mis en oeuvre sur un cas d'étude. Les données d'entrée sont des images satellitaires à très haute résolution spatiale de 0.5 m par pixel. Ces images sont, dans un premier temps, traitées par un algorithme de classification à base de réseaux de neurones, CNN, qui identifie la nature des éléments au sol. Le résultat de cet algorithme est un raster dont les pixels sont étiquetés par un ensemble de labels comme : routes, végétation ou bâtiments. Ces labels sont ensuite utilisés pour définir des instances géolocalisées des classes de l'ontologie comme `ontocity:Region`. D'autres entités localisées dans la zone couverte par l'image

25. `http://ontologydesignpatterns.org/wiki/Ontology:DOLCE%2BDnS_Ultralite`

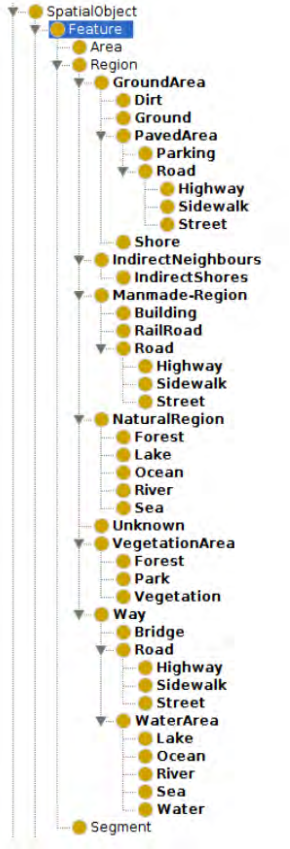


FIGURE 3.11 – Hiérarchie des classes de l'ontologie Ontocity [Alirezaie *et al.* 2017]

sont extraites de OSM et viennent enrichir la base de connaissances en spécifiant le type de bâtiment par exemple.

Enfin, des lieux particuliers situés dans cette zone et sur lesquels sont survenus des événements à surveiller sont rajoutés à la base de connaissances en utilisant une des classes de l'ontologie. Dans le cas d'étude choisi, les évènements étudiés sont des zones inondées. Un simulateur d'inondation est utilisé pour générer une carte d'inondation à partir d'un modèle 3D de la ville de Stockholm. Les zones inondées sont ensuite représentées dans la base de connaissances à l'aide de la classe `ontocity:WaterArea`, sous-classe de `ontocity:Area`.

Le processus de raisonnement est alors lancé afin de déterminer les meilleurs chemins possibles pour l'intervention des secours dans une zone inondée. Étant donné que chaque entité est géo-référencée par sa géométrie, il est possible d'évaluer les distances entre chacune des entités et d'estimer si celles-ci sont voisines. Pour le calcul d'un chemin sans obstacle à partir d'un point donné jusqu'à une destination, le choix a été fait d'utiliser l'algorithme Rapidly-exploring Random Tree (RRT) développé par [LaValle 1998]. Cet algorithme retourne l'ensemble des noeuds praticables entre le point de départ et le point d'arrivée.

3.2.2.2 Savia : ontologies et images satellitaires pour évaluer le fonctionnement de zones Natura 2000 [Pérez Luque *et al.* 2015]

Savia est un système qui facilite la surveillance des zones naturelles (Natura 2000) grâce aux images satellitaires et aux technologies du Web sémantique [Pérez Luque *et al.* 2015]. Ce système utilise des images produites par le satellite optique MODIS, des ontologies standards pour la représentation des données ainsi qu'un entrepôt de triplets RDF pour leur interrogation. Savia a été validé à l'aide d'un cas d'utilisation portant sur les forêts de la Sierra Nevada en Espagne sur une surface de plus de 2000 km^2. Ce site fait partie des territoires protégés par le programme Européen Natura 2000 [26].

Les images utilisées par Savia ont été produites par le satellite MODIS de l'année 2000 à 2012. Deux types d'images MODIS sont exploitées : les produits MOD13Q1 et MOD10A2. Les images MOD13Q1 possèdent une résolution spatiale de 250 m et une résolution temporelle de 16 jours. Les images MOD10A2 possèdent une résolution spatiales de 500 m et une résolution temporelle de 8 jours. Afin d'homogénéiser les deux résolutions spatiales, les deux grilles de pixels sont croisées afin d'assigner les identifiants de la grille la moins précise (celle du produit MOD10A2) aux données du produit le plus précis (MOD13Q1). Pour homogénéiser les résolutions temporelles, les deux types données sont agrégés selon l'échelle de temps la moins précise (au moins 16 jours).

Les images MOD13Q1 fournissent des informations sur la végétation grâce au calcul de l'indice NDVI pour chaque pixel. Les images MOD10A2 permettent d'étudier la couverture neigeuse grâce au calcul de l'indice NDSI pixel par pixel. Le calcul de cet indice est similaire à celui du NDVI mais en utilisant d'autres bandes spectrales, celles du vert et l'infrarouge à ondes courtes.

Une fois ces indices calculés et associés à la résolution spatiale et temporelle commune, des tendances sont établies, qui seront elles aussi stockées dans une base de données relationnelle. A partir de la collection d'images, plusieurs indicateurs peuvent être calculés en utilisant le NDSI ou le NDVI tels que : pour le NDSI, le nombre de jours couverts de neige par année, le premier et le dernier jour de neige dans l'année et le nombre de cycle de fonte de neige par année ; pour le NDVI, le couvert végétal aux différentes saisons, le couvert végétal minimal ou maximal etc. (cf Figure 3.12).

L'ontologie développée afin de représenter sémantiquement ces données comporte donc trois concepts de haut niveau : le concept `Pixel` pour représenter les `pixels` des images satellitaires MODIS, le concept `IndicatorValue` pour les `indices` calculés et le concept `IndicatorTrend` pour les `tendances` dérivées des indices. Cette ontologie repose sur le standard OWL-Time présenté en section 3.1.2.2 pour la dimension temporelle et sur le standard Basic Geo présenté en section 3.1.1.2 pour la dimension spatiale. On remarque sur la Figure 3.12 que le concept `IndicatorValue` est spécialisé en 2 sous-classes : `Snow` et `Vegetation`. Le concept `IndicatorTrend` est lui aussi spécialisé en deux sous-classes : `SnowTrend`

26. `https://www.ecologie.gouv.fr/reseau-europeen-natura-2000-1`

et `VegetationTrend`. Enfin, le concept `Pixel` est rattaché à un `Patch` qui lui-même est rattaché à un `Group` pour faciliter l'exploitation.

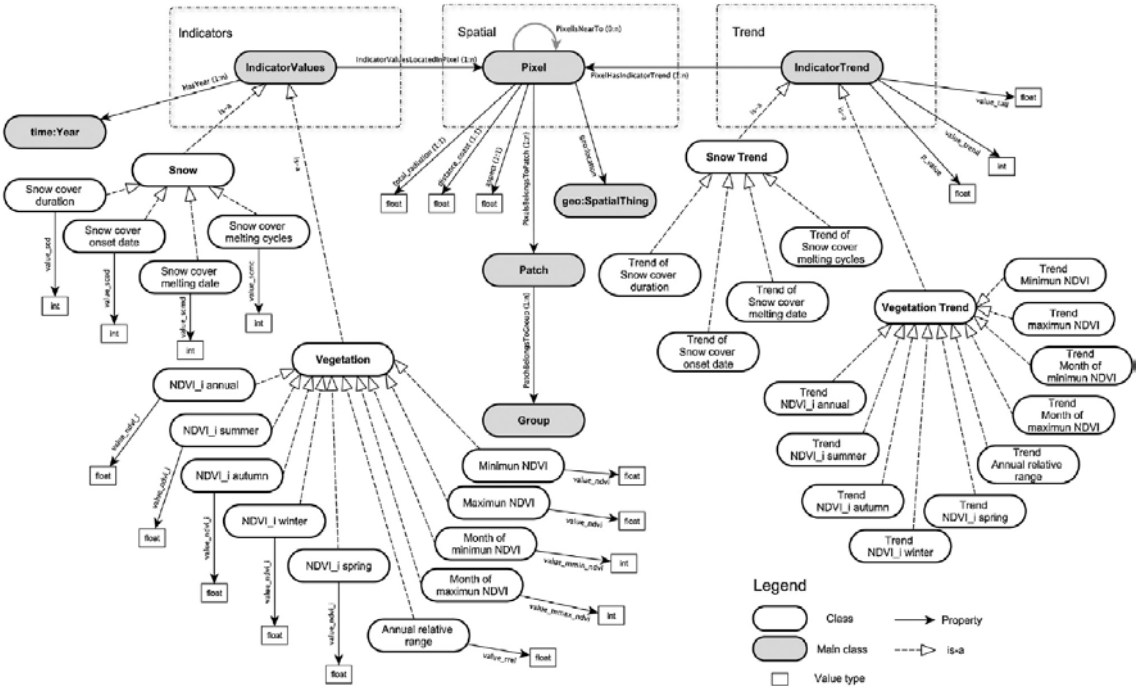


FIGURE 3.12 – Représentation de l'ontologie de Savia [Pérez Luque *et al.* 2015]

L'approche choisie pour peupler cette ontologie consiste à utiliser D2RQ [27] qui est un outil permettant de générer des triplets RDF à partir d'une base de données relationnelle en se basant sur des modèles. Ces triplets sont ensuite stockés dans le triplestore Allegrograph [28] car la solution Apache Jena [29] n'était pas assez robuste pour le nombre de triplets à stocker. Le moteur d'inférence contenu dans Allegrograph permet de trouver des relations entre les différents type d'indicateurs et des propriétés implicites comme `PixelsNearTo`.

Pour la partie exploitation, une interface Web a été développée afin de construire des requêtes SPARQL. Des exemples de requêtes correspondent à des questions comme *Which pixels show a trend towards higher productivity in summer ?* ou *Which pixels show a trend towards an earlier snow melt ?*. Les pixels et leurs valeurs retournés par la requête sont ensuite affichés sur une carte dans l'interface Web.

3.2.2.3 GeoSensor : sémantiser la détection de changements et d'événements au sein de grands jeux de données [Pittaras *et al.* 2019]

Les travaux réalisés par [Pittaras *et al.* 2019] ont pour but d'enrichir la détection de changement obtenue par images satellitaires avec des données ouvertes. Pour

27. `http://d2rq.org/`
28. `https://allegrograph.com/`
29. `https://jena.apache.org/`

ce faire, cette équipe a mis au point une chaîne de traitement des données, appelée GeoSensor, qui se décompose en trois couches : une couche *détection de changements*, une couche *détection d'évènements* et une couche *sémantique*. La Figure 3.13 représente l'ensemble des composants de cette chaîne de traitement.

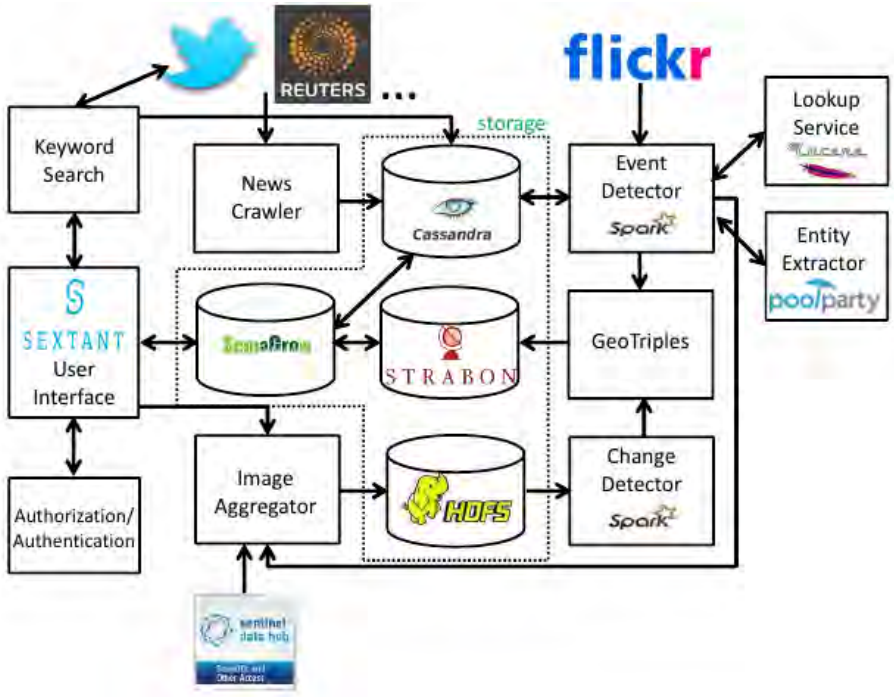


FIGURE 3.13 – Architecture du système GeoSensor [Pittaras *et al.* 2019]

La partie *détection de changements* est chargée de récupérer et comparer deux images afin d'obtenir une localisation des changements dans l'occupation et l'utilisation des sols. Un module télécharge les images directement depuis le portail de l'ESA [30] en fonction de la zone et de la période choisies par l'utilisateur. Les images sont ensuite stockées dans un entrepôt Hadoop Distributed File System (HDFS). La détection de changements entre les images est ensuite réalisée par un module de la suite logicielle SNAP Toolbox [31]. Cette suite développée par l'ESA comprend un ensemble d'algorithmes dédiés aux traitements des images issues des satellites du programme Sentinel. Le résultat obtenu par cet algorithme est un raster contenant le taux de changement de l'occupation et de l'utilisation des sols pour chaque pixel du raster. En dernière étape, l'algorithme Density-Based Spatial Clustering of Applications with Noise (DBSCAN) est utilisé pour organiser les pixels en clusters, en regroupant les pixels dont le taux de changement est similaire.

Pour la partie *détection d'évènements*, un News Crawler a été développé. Celui-ci scrute, toutes les 30 minutes, différentes sources de données textuelles comme Twitter ou un flux RSS fourni par l'agence de presse Reuters. Les titres les plus

30. `https://scihub.copernicus.eu`
31. `http://step.esa.int/main/download/snap-download/`

importants sont extraits. De plus, les noms des unités administratives sont identifiés dans les articles grâce à Apache openNLP [32]. Une fois ces unités identifiées, leur géolocalisation sous forme de polygone est récupérée à partir du jeu de données GADM. Les homonymes sont gérés par un classement réalisé selon le taux de similarité entre chaînes de caractères (l'unité localiseée et celle mentionnée dansvle texte) puis compariason de la surface de l'entité. D'autres données contextuelles sont ajoutées comme des images disponibles sur le site Flickr [33].

La partie *sémantique* est assurée par un ensemble de composants utilisant les technologies du Web sémantique. Le premier composant est GeoTriples [34] développé par la même équipe. Ce composant permet de convertir un grand nombre de données géospatiales au format RDF [Kyzirakos *et al.* 2018]. Le second composant utilisé est le triplestore Strabon capable de stocker une grande quantité de triplets RDF. Strabon intègre GeoSPARQL ainsi que stRDF présenté dans la section 3.1.1.5. Le dernier composant de cette couche sémantique est l'outil SemaGrow développé par [Charalambidis *et al.* 2015]. Ce composant permet d'optimiser les requêtes SPARQL et d'interroger plusieurs triplestores en ligne en même temps.

Une expérimentation a été conduite afin d'évaluer la performance de l'ensemble de la chaîne de traitement. Les images satellitaires utilisées proviennent du satellite Sentinel-1A. Deux paires d'images ont été utilisées : une paire concerne Los Angeles et l'autre paire concerne l'Arabie Saoudite. Les tests ont été effectués pour mesurer le temps de calcul d'une détection de changements avec l'outil SNAP avec multithreading 8 coeurs et avec 2 ou 4 machines virtuelles avec Apache Spark [35]. Les résultats obtenus montrent que la parallélisation en machines virtuelles avec Spark est 3 fois plus performante que le multithreading. Pour la partie détection d'évènements grâce aux données contextuelles, une évaluation des performances a également été effectuée. Les tests ont porté sur deux corpus d'articles, le premier contenant 4000 articles et le second 8000 articles. Ici aussi les performances ont été évaluées entre une machine en multithreading 8 coeurs, 2 machines virtuelles Spark et 4 machines virtuelles Spark. Le résultat est similaire, les machines virtuelles Spark sont plus performantes que le multithreading en temps de calcul. La conclusion de leur étude est la suivante : plus il y a de machines virtuelles Spark en parallèle, plus les performances sur les temps de calcul sont améliorées.

3.2.2.4 Étude des changements à partir d'annotations sémantiques d'images [Yao *et al.* 2020]

Une des manières de valoriser les collections d'images est de fournir des services pour les exploiter et mieux les retrouver, mais aussi d'en décrire le contenu en vue de retrouver des images pertinentes sur un lieu ou pour étudier un type de phénomène. Afin de mieux caractériser les informations relatives à une zone géogra-

32. `https://opennlp.apache.org/`
33. `https://www.flickr.com/`
34. `http://geotriples.di.uoa.gr/`
35. `https://spark.apache.org/`

phique donnée, l'approche proposée par [Dumitru *et al.* 2019] consiste à fusionner les analyses tirées de deux sources : une image Sentinel-1 (S1), image radar d'une résolution spatiale de 20 m, et une image optique Sentinel-2 (S2), de la même région à la même période. Chaque type d'image est fournie avec un jeu d'indicateurs qui lui sont propres et qui peuvent être exploités. L'intérêt de la fusion est de récupérer des indicateurs complémentaires de chacune des images sources, de bénéficier du fait que les images S1 ne sont pas sensibles à la présence de nuages et de la plus grande précision des images S2. Cette fusion est préalable à une caractérisation sémantique et à un processus de fouille de données qui permet de chercher des régularités sur les images. Ces travaux ont débouché sur une application à l'étude de changements dans le cadre du projet H2020 CANDELA [36] [Dumitru *et al.* 2018a] [Dumitru *et al.* 2018b]. Les changements observés concernaient l'évolution de forêts (croissance, coupes, plantations ou dégâts suite à des événements météorologiques). Les changements peuvent être identifiés en comparant deux jeux d'images annotées, c'est-à-dire des couples (S1,S2) d'une même région à deux dates différentes [Dumitru *et al.* 2019].

Afin de faciliter la fusion, les images sont découpées selon un maillage géographique identique et les annotations sont fournies au niveau de chaque élément de cette grille (appelé patch). L'approche choisie pour caractériser automatiquement des contenus s'appuie sur un processus de classification basé sur un algorithme d'apprentissage supervisé. Cet algorithme nécessite donc l'annotation manuelle d'une partie de l'image. Cette annotation est dite *sémantique* car elle s'appuie sur un vocabulaire d'entités pouvant être reconnues sur les images. Ce vocabulaire est une liste à plat de termes désignant des objets naturels ou artificiels pouvant être reconnus à la surface de la Terre.

Dans le cadre du projet CANDELA, plusieurs paramètres sont retenus dont le descripteur de Weber adapté pour les images S1 car il caractérise des structures en minimisant le bruit, et pour les images S2, l'histogramme multi-spectral car il contient une information physique qui caractérise toute l'image. Ces deux paramètres sont calculés sur chaque patch et concaténés pour faire un descripteur [Dumitru *et al.* 2018b] [Dumitru *et al.* 2020]. L'algorithme de classification choisi est un algorithme d'apprentissage actif basé sur SVM : l'utilisateur sélectionne des exemples, les classe et corrige les erreurs de l'apprentissage. Le résultat de l'apprentissage est enregistré dans une base de données et constitue ce que les auteurs qualifie d'annotation sémantique. Mais pratiquement, aucune technologie sémantique n'est utilisée. Ces annotations peuvent être exportées et intégrées à d'autres données [Yao *et al.* 2021], à condition d'être traduites dans un format standard comme RDF [Rolland *et al.* 2020].

36. `https://candela-h2020.eu/`

3.3 Positionnement et reformulation du sujet

Cet état de l'art nous permet d'identifier les problématiques inhérentes à la question de recherche au centre de notre thèse, à savoir l'apport des technologies du Web sémantique à la valorisation et l'utilisation des images satellitaires, en particulier pour l'étude des changements. Tout d'abord, la plupart des analyses d'images, y compris l'étude des changements, produisent des données au format raster. Les résultats les plus performants pour identifier des changements par comparaison de deux images utilisent des algorithmes d'apprentissage, et les algorithmes non supervisés parviennent à des résultats tout à fait performants. Nous en retenons qu'il est préférable d'utiliser des changements calculés à partir des fichiers raster (image) d'origine plutôt que sur des représentations sémantiques de leur contenu (par exemple la présence/absence d'objets). Nous démarquons donc des approches présentées dans [Yao *et al.* 2020] (§ 3.2.2.4) ou dans [Harbelot *et al.* 2015] (§ 3.2.1.3) qui calculent les changements à partir de représentations en RDF du contenu de l'image. Dans notre approche, les changements sont considérés comme des données particulières à intégrer aux images via à un graphe de connaissances à partir d'un raster de changement daté par un intervalle. La représentation sémantique de ce raster de changement est une des problématiques traitée dans la thèse. Un des avantages de ce choix est de permettre de traiter de la même manière que les changements toute information au format raster, en particulier les indices calculés par analyse d'image, comme l'indice de couvert végétal NDVI ou le le Burned Area Index (BAI).

Nous retenons également de l'état de l'art qu'une problématique commune à ces travaux est d'associer des données ouvertes aux images et de les relier aux résultats de leur analyse pour les documenter. Les graphes de connaissances du Web sémantique correspondent donc à une solution possible pour représenter les résultats de ces analyses et faciliter la mise en relation avec les métadonnées des images d'une part et d'autres types de données ouvertes d'autre part. Les représentations sémantiques viennent enrichir la description des images, leur associent de nouvelles métadonnées qui les rendent plus pertinentes et facilitent leur recherche.

La figure 3.14 montre une synthèse des travaux présentés sur la représentation de changements liés aux images satellitaires et aux données géospatiales grâce au Web sémantique. Nous distinguons deux types d'approches pour ces travaux : la première approche où le phénomène est identifié sur les images avant le processus de leur sémantisation et la seconde approche où le phénomène est identifié à partir des connaissances annotant les images. Ces travaux ont été étudiés selon quatre critères principaux : le type de données utilisées comme sources, l'utilisation de données contextuelles, le type de phénomène étudié et comment est détecté le changement. Ils nous permettent d'expliquer et situer nos choix. Dans nos travaux, le changement est détecté sur des images satellitaires où le phénomène n'est pas connu. Il sera identifié grâce à des valeurs de changement calculées par apprentissages automatique à partir de deux images. Le changement sera ensuite caractérisé à l'aide de différentes données contextuelles et de plusieurs indices calculés depuis les images. On peut

voir sur ce tableau qu'aucune des approches étudiées ne répondait à l'ensemble de nos choix pour ces critères.

Cette analyse nous permet ainsi de préciser le sujet de la thèse, dont l'objectif est donc de *définir comment représenter et exploiter, à l'aide des technologies du Web sémantique, des données géospatiales identifiées par analyse d'images satellitaires, en particulier des données liées à des changements.* L'objectif est très proche de celui de GeoSensor [Pittaras *et al.* 2019] mais nous supposons déjà réalisée la collecte des images et le calcul des changements, pour nous focaliser sur la représentation sémantique et l'ajout de données contextuelles.

3.4 Conclusion

Dans ce chapitre a été présenté l'essentiel des standards pour la représentation des dimensions spatiales et temporelles de données, ainsi que des bases de connaissances du Web fournissant des données géolocalisées. Nous avons aussi proposé un panorama de travaux visant la représentation sémantique de données géolocalisées ou, plus proches de la question traitée dans la thèse, des recherches sur l'étude des changements sur des images de satellites et leur représentation sémantique. Certains de ces travaux réutilisent des vocabulaires standards alors que d'autres font le choix de définir leurs propres vocabulaires. Leurs résultats contribuent à montrer que le Web Sémantique peut être une solution pour la représentation et l'exploitation de données géospatiales ayant des sources et des formats hétérogènes. Cet état de l'art nous a permis d'identifier les problématiques suivantes soulevées par la thèse :

— sur la manière d'articuler le calcul des changements ou d'indices à partir d'images satellitaires et leur représentation sémantique ;

— sur le choix de construire une ontologie modulaire à base de standards mais aussi de concepts spécifiques pour structurer la représentation sémantique ;

— sur la nécessité de pouvoir spécifier un dimensionnement spatial adapté à chaque phénomène étudié, en sélectionnant soit des zones géographiques préexistantes, soit en calculant des régions à partir des données ;

— sur l'architecture et le processus à prévoir depuis la sélection de fichiers résultant d'analyses d'images à la représentation sémantique de l'interprétation de leur contenu en lien avec des données contextuelles.

Nous développerons notre positionnement relativement à ces problématiques dans le chapitre suivant (chapitre 4), où nous présentons en détail les solutions retenues, les modèles et les algorithmes proposés pour implémenter notre approche. Cette contribution a été évaluée sur différents jeux de données réels, et cette évaluation fait l'objet du chapitre 5.

	Données utilisées sources	Données contextuelles	Phénomène étudié	Détection du changement
[Alirezaie et al. 2017]	Images satellitaires haute résolution	Open Street Map	Inondations	• Représentations en classes de land cover • Inférence sur les données sémantisées
[Bernard et al. 2020]	Nomenclature des Unités Territoriales Statistiques (NUTS), Switzerland Administrative Units Classification (SAU), Australian Statistical Geography Standard (ASGS)		Évolution des unités territoriales	• Représentation des unités territoriales • Inférence sur la géométrie des données sémantisées
[Harbelot et al. 2015]	CORINE Land Cover		Intensification urbaine, Inondations, Déforestation	• Représentation en classes de land cover • Inférence sur les données sémantisées
[Kauppinen et al. 2013]	Images Landsat Thematic Mapper	Données statistiques gouvernementales	Déforestation	• Différenciation sur les classifications
[Kyzirakos et al. 2014]	Images satellitaires MSG/SEVIRI	Open Street Map, LinkedGeoData, GeoNames	Incendies	• Sur les données sémantisées via SPARQL
[Pérez Luque et al. 2015]	Images satellitaires MODIS		Santé de la forêt	• Tendances calculées sur les données sémantisées
[Pittaras et al. 2019]	Images satellitaires Sentinel-1	Reuters, Twitter, Flickr, GADM	Migration	• Sur les images satellitaires
[Dorne 2021]	Images satellitaires	Yago2geo, Twitter, Firecast	Incendies, explosions	• Sur les images satellitaires

Phénomène identifié avant sémantisation Phénomène identifié après sémantisation

FIGURE 3.14 – Positionnement par rapport aux travaux existants

Chapitre 4

Proposition

Content

La contribution de cette thèse vise donc à produire une représentation sémantique à partir de données en lien avec des images satellitaires : des données calculées à partir de ces images (indices calculés à partir d'une image ou changements calculés par apprentissage à partir de la comparaison de deux images), métadonnées de ces images et données contextuelles choisies pour décrire ou documenter le contenu des images. Nous avons retenu le positionnement présenté en fin de chapitre 3 pour atteindre cet objectif :

— nous supposons que les traitements sur les images sont réalisés en amont, qu'il produisent des indices ou des indications de changement accessibles sous format matriciel au sein d'un fichier raster

— nous voulons que l'approche permette de relier aux images et aux régions d'intérêt différents types de données ouvertes ou données contextuelles, localisées et datées, qui fourniront des informations utiles pour qualifier, décrire ou valider les zones d'intérêt identifiées ;

— nous construisons une ontologie modulaire et basée sur des standards de représentation sémantique du temps et des données spatiales, ainsi que des travaux sur la représentation des changements, pour organiser les données ainsi définies au sein d'un graphe de connaissances afin de pouvoir rechercher ces images et ces données ;

— nous définissons une architecture et un processus afin de générer ce graphe de connaissances à partir d'un fichier raster d'indices ou de changements associés aux images.

Au sein d'un fichier raster, les données sont associées à chaque pixel, ce qui ne facilite pas l'interprétation par un humain. Une des questions importantes est de définir comment regrouper les pixels dont les valeurs sont proches afin d'identifier des zones pertinentes caractérisées par ces valeurs. Nous avons expérimenté trois manières de définir ces zones :

— s'appuyer sur des zones fixes définies a priori et ne faisant pas nécessairement sens pour un humain, mais toutes de même taille ; dans ce cas, nous avons exploité le tuilage des images ;

— s'appuyer sur des zones fixes définies a priori et faisant sens pour un humain, de tailles et formes différentes ; dans ce cas, nous avons exploité les polygones des villes ou unités administratives localisées sur la zone de l'image ;

— s'appuyer sur des zones calculées à partir des données, ne faisant pas nécessairement sens pour un humain, mais au plus près de la zone touchée par les phénomènes mis en évidence par les indicateurs ou les changements ; pour cela, nous avons défini la notion de Région d'intérêt (ROI) et proposé un algorithme pour la calculer.

Enfin, l'état de l'art a confirmé que les données contextuelles telles que peuvent les fournir les réseaux sociaux, les données publiques des organismes nationaux et finalement une grande variété de données ouvertes constituent des éléments importants pour annoter les images. Elles peuvent documenter les zones ainsi identifiées, donner une interprétation aux changements ou aux indices, ou encore servir à vérifier la qualité des résultats en confirmant que le phénomène étudié a bien été identifié par d'autres sources. Dans cet objectif, les graphes de connaissances assurent la mise en relation de données contextuelles avec les données tirées des images. Les ontologies définies doivent donc aussi assurer cette mise en relation. Une manière simple est de s'appuyer sur la dimension spatiale (les géométries des données) pour sélectionner à la demande des données ouvertes d'un type particulier relatives aux zones d'intérêt.

Ce chapitre présente nos propositions en distinguant trois manières de restituer les agrégations de pixels similaires (tuile, polygones et ROI). Chacune de ces

approches est mieux adaptée à un type de données particulier et le choix de la dimension spatiale impacte aussi la référence temporelle des données :

— **les tuiles** conviennent pour associer les indices de couvert végétal calculés à partir de l'image ou NDVI ; ainsi, on associe une valeur globale (majoritaire ou moyenne) à une zone ou à la totalité de l'image ; la temporalité est alors celle de la prise de vue ;

— **les polygones** ont été retenus pour associer des indices de couvert des sols calculés sur d'autres séries d'images (les données de CORINE Land Cover (CLC) CORINE Land Cover) et disponibles comme données ouvertes, ce qui revient à associer la nature du couvert des sols majoritaire (considéré comme stable pendant plusieurs années) sur une unité administrative ; la temporalité est ici celle des données ;

— enfin, **les ROI** permettent de définir plus précisément les zones de changements à partir des indices de changements calculés, et donc de localiser assez précisément et rapidement les phénomènes ayant produit les changements. La temporalité est alors celle de la prise de vue des images.

Une première partie de ce chapitre (4.1) développe les choix retenus en réponses aux problématiques identifiées à la fin de l'état de l'art, relativement aux ontologies à définir, aux manières de regrouper les données et les pixels ainsi qu'aux processus de génération d'un graphe de connaissances à partir de données raster. La deuxième partie (4.2) présente les trois ontologies que nous proposons pour chacune de ces manières de regrouper les pixels, notamment les standards utilisés pour la construction des modèles. La troisième partie (4.3) développe les processus de transformation des connaissances raster au format Web sémantique dans chacun des trois cas.

4.1 Problématiques et propositions de la thèse

La thèse soulève plusieurs questions de recherche et problématiques dans la continuité de l'état de l'art présenté au chapitre précédent.

— La première problématique abordée est de **définir un modèle ontologique suffisamment complet et générique** pour représenter les différentes données géospatiales à notre disposition (indicateurs ou changements calculés à partir des images et disponibles au format raster d'une part, données ouvertes géolocalisées d'autre part) en réutilisant des vocabulaires standards. Ce vocabulaire doit aussi faciliter la mise en relation des données calculées, des données contextuelles et des métadonnées d'images.

— La seconde problématique traitée est le **dimensionnement des entités spatiales** nécessaires pour représenter les données géolocalisées calculées à partir des images. En effet, les données sont associées aux pixels, mais le niveau pixel peut ne pas être pertinent pour les restituer aux utilisateurs si la surface représentée par un pixel n'est pas suffisante par rapport au phénomène étudié.

De plus, les pixels sont très nombreux sur chaque image, ce qui génère un gros volume de données dont une partie est inutile lorsque l'indice calculé n'a pas une valeur significative. La question est donc de définir une manière d'agréger les pixels proches géographiquement et dont la valeur est similaire pour les associer à des zones géographiques "interprétables", qui fassent sens. Il s'agit aussi de prévoir un algorithme pour calculer cette agrégation.

— La troisième problématique consiste à définir, implémenter et évaluer un processus qui s'appuie sur cette ontologie et intègre le choix d'unités spatiales adaptées pour générer un graphe de connaissances qui permette de relier les données tirées des images, les données contextuelles et les images.

Nous revenons sur chacune de ces problématiques.

4.1.1 Un modèle ontologique pour représenter des données géolocalisées associées à des images satellitaires

Dans les parties 3.1.1 et 3.1.2, nous avons présenté plusieurs standards et vocabulaires disponibles pour la représentation des dimensions spatiales et temporelles de données. Cependant, il n'existe pas d'ontologie standard permettant la représentation de connaissances spécifiques aux résultats d'analyse d'images comme l'indice NDVI dans [Pérez Luque *et al.* 2015] ou la couverture des sols dans [Harbelot *et al.* 2015]. De même, il n'existe aujourd'hui aucune ontologie standard pour représenter tous les types d'images satellitaires, ni même pour représenter en RDF les métadonnées associées (cf section 2.1.3.1) qui dépendent des choix du constructeur du satellite. Toutefois, plusieurs approches, comme [Kyzirakos *et al.* 2014a], convergent en représentant les images satellitaires comme un phénomène observé via un capteur grâce à l'ontologie SWEET. Ce type d'approche requiert de définir dans le modèle une correspondance entre tous les types de métadonnées existants pour pouvoir les aligner.

Nous avons choisi de créer une ontologie spécifique, qui puisse être réutilisée pour des travaux similaires, tout en répondant au mieux aux besoins de modélisation des trois types de données géolocalisées qui nous intéressent : les indices et changements calculés à partir des images, les métadonnées d'images et des données contextuelles. Pour construire cette ontologie, nous avons réutilisé partiellement des ontologies et des vocabulaires existants, si possible des standards recommandés par l'OGC et/ou le W3C, afin de favoriser la potentielle réutilisation de ce modèle pour d'autres applications ou dans le LOD. Cette ontologie est modulaire, au sens où chaque type de données (indices et changements, métadonnées d'images et donnés contextuelles) possède son modèle qui comporte une dimension spatiale et une dimension temporelle. Ces deux dimensions sont modélisées grâces aux ontologies standards GeoSPARQL et OWL-Time. Nous partons du principe que nous ne connaissons pas toutes les données qui seront ajoutées au graphe de connaissances selon ce modèle mais que chaque donnée doit pouvoir être retrouvée via ses dimensions spatiale et temporelle.

4.1.2 Dimensionnement spatial

Les algorithmes de comparaison d'images traitent les données spectrales ou d'autres descripteurs des images. Ils fournissent en sortie une image de changement, c'est-à-dire un fichier au format raster couvrant la zone comparée et pour lequel, à chaque pixel, est associé un taux de changement (soit une valeur numérique, soit une valeur symbolique calculée à partir de valeurs numériques et de seuils). La question soulevée ici est se savoir comment représenter le contenu d'un raster en fonction d'un type de changement étudié de manière automatique. Il existe plusieurs méthodes pour la représentation de ces fichiers raster à l'aide des technologies du Web sémantique. Une des approches réalisée par [Pérez Luque *et al.* 2015] est de représenter chaque pixel de l'image comme une entité au sein d'un graphe RDF. Cette approche de modélisation garantit que l'ensemble du contenu du raster soit accessible en RDF, mais peut présenter un problème de passage à l'échelle. De plus, elle ne répond pas au critère d'interprétation du phénomène associé au changement si l'échelle du pixel n'est pas adaptée.

De ce fait, nous avons choisi une approche alternative, où l'on ne représente pas chaque pixel, mais des groupes de pixels comme dans [Pittaras *et al.* 2019]. En effet, pour les images des satellites Sentinel-2 que nous utilisons, une grille contient plus d'un million de pixels pour un seul raster. De plus, tous les pixels ne sont pas intéressants au sens de l'étude des changements : seuls ceux présentant un *changement significatif* et sur une *surface suffisamment grande* seront retenus. Ces deux paramètres (valeur du changement et surface minimale) dépendent du type de changement étudié, mais aussi de la résolution spatiale des images satellitaires. Finalement, seules les parties d'une image concernées par un changement "significatif" en surface et en importance sont représentées sous forme sémantique.

L'approche que nous avons retenue est de rassembler les pixels du raster ayant une valeur proche sous forme de groupes appelés Régions d'Intérêt (Region of Interest ou ROI par la suite). Contrairement aux travaux de [Pittaras *et al.* 2019], qui ont utilisé l'algorithme DBSCAN pour regrouper les pixels, nous avons développé notre propre algorithme pour la création de ces groupes. Cet algorithme est moins précis que DBSCAN dans la forme des groupes car il ne génère que des polygones rectangulaires (qui engloberaient le résultat de DBSCAN). En revanche, il permet d'obtenir de meilleures performances sur l'exploitation d'un nombre important de polygones, assurant un traitement plus rapide des données et un meilleur passage à l'échelle. Ces ROI sont des polygones géo-référencées et datés avec les dates fournies par le raster de changement. Cette approche est innovante car elle permet de sauvegarder, de manière automatique, uniquement les zones de l'image dans lesquelles se trouve une forte probabilité de changement, et non l'ensemble des pixels. De plus, une seule valeur est associée à chaque ROI, ce qui réduit fortement le volume de données à représenter en RDF et facilite grandement l'exploitation.

Pratiquement, pour que les données puissent être associées à une zone géographique, nous avons envisagé et expérimenté plusieurs propositions de représentation, qui correspondent à plusieurs manières d'agréger les pixels mais aussi les données

associées. Dans tous les cas, il faut aussi définir un algorithme pour associer une valeur à cette zone à partir de la valeur des pixels qui la composent : moyenne, maximum ou minimum, etc. Ces zones géographiques sont utilisées comme un outil permettant d'associer de la connaissances liée à un changement à une entité géolocalisée.

— Le premier choix a été de représenter les données en les associant à une tuile de la grille Sentinel 2 définie par l'ESA. Ces tuiles sont des polygones recouvrant une surface allant jusqu'à 110km^2. L'avantage de ce choix est que toutes les zones géographiques sont de même taille et de même forme. Ces zones sont peu nombreuses et réduisent la quantité de données à stocker.

— La seconde manière de définir des zones spatiales testée dans nos travaux est la représentation à l'aide de polygones complexes. Ces polygones peuvent être prédéfinis, et par exemple correspondre à la surface au sol d'une unité administrative. Il peuvent aussi être calculés, en fonction de la valeur des pixels, et correspondre à la surface générée par un ensemble de pixels contigus ou voisins. Cette représentation à l'avantage de faire sens pour les utilisateur et d'être très précise sur la zone à étudier. Cependant, elle génère des entités dont la taille de stockage est plus importante. De plus, cette représentation peut engendrer des temps de réponse conséquents lorsqu'elle est utilisée dans des opérations de comparaisons spatiales.

— La dernière unité spatiale sur laquelle nous avons travaillé est la région d'intérêt ou ROI. Ces polygones sont générés au moment de l'intégration des données. Ils correspondant à une simplification par un rectangle englobant, de la surface correspondant à l'agrégation de pixels voisins de valeur similaire. Pratiquement, ces zones couvrent approximativement la zone à étudier. Mais leur avantage majeur est leur faible taille, et donc leur stockage moins volumineux, qui permet d'en gérer une grande quantité et de faciliter leur exploitation.

Dans la suite, nous présentons une expérience menée avec chacun de ces trois types de zone regroupant des données associées à des pixels. Nous montrons comment les données sont gérées dans chacun des cas, l'ontologie qui a été construire pour organiser le graphe de connaissances, et dans la partie suivante, comment un graphe de connaissances est généré à partir des données de raster et comment il est relié à des données ouvertes.

4.1.3 Processus de génération d'un graphe de connaissances géolocalisées à partir de raster

La troisième problématique consiste à définir, implémenter et évaluer un processus qui s'appuie sur l'ontologie et intègre le choix d'unités spatiales adaptées pour générer un graphe de connaissances qui permette de relier les données tirées des images, les données contextuelles et les images. La difficulté ici est de rendre l'approche suffisamment générale pour prendre en compte différents types d'images

avec des métadonnées spécifiques, plusieurs types de données calculées à partir d'images, des critères de surface et d'importance des indices pour sélectionner des régions d'intérêt.

Dans nos travaux, les images satellitaires servent à calculer des indices comme le NDVI ou le BAI. Une fois ces indices calculés, nous regroupons les valeurs via un algorithme de clusterisation et nous les représentons au format RDF. Nous avons fait le choix de ne représenter que les métadonnées les plus importantes comme l'identifiant de l'image ou sa date afin de retrouver la ou les images utilisées dans les différents calculs et traitements. Cette approche nous permet une grande flexibilité sur la modélisation de différentes types de métadonnées et donc d'intégrer plus facilement des images issues de nouveaux satellites.

4.2 Ontologies pour représenter des connaissances géolocalisées et des changements

Nous nous intéressons à la représentation sémantique de données extraites des images et des données contextuelles géo-localisées qui évoluent dans le temps. Nous nous sommes focalisés successivement sur plusieurs types de données disponibles au format raster, et avons proposé différentes manières de les regrouper en zones géographiques. Ceci nous a conduits à définir une ontologie spécifique pour chacune de ces modélisations. Dans la pratique, ces ontologies ont une structure commune et peuvent être adaptées afin de représenter d'autres types de données soit tirées de l'analyse d'images, soit extraites de bases de données ou du web.

Chacune de ces ontologies est construite de manière modulaire, chaque module correspondant à un vocabulaire réutilisé ou nouveau, ayant un espace de nom spécifique. Le nom du module est exploité dans ce qui suit pour préfixer le vocabulaire qu'il définit. La noyau de ces ontologies réutilise l'ontologie présentée dans [Arenas *et al.* 2018], définie pour intégrer des données géolocalisées et datées en lien avec des images satellitaires. Ce noyau intègre plusieurs vocabulaires du LOD reconnus comme des standards pour faciliter la réutilisation des données ainsi représentées, dont GeoSPARQL pour la dimension spatiale, l'ontologie Sensor, Observation, Sample, and Actuator (SOSA) [1] pour représenter une partie des méta-données d'images satellitaires considérées ici comme données de capteurs, et OWL-Time pour la dimension temporelle. D'autres modules sont spécifiques à notre démarche, nous les présentons dans la suite.

La représentation du land cover ainsi que la représentation du NDVI reposent sur une approche par pourcentage de surface. Cela revient à associer à une surface une représentation de la valeur de toutes les classes de land cover. Cette valeur correspond à un pourcentage : il indique le pourcentage de la surface couvert par ce type de sol. Cette approche générique peut s'appliquer à tout indice numérique pouvant être classé en plusieurs catégories étiquetées symboliquement (ou classes) selon plusieurs valeurs de seuil. On passe ainsi d'une représentation numérique par

1. `https://www.w3.org/2015/spatial/wiki/SOSA_Ontology`

pixel à une représentation de pourcentages par valeur symbolique associée. La représentation du changement repose sur une approche par collections de polygones. Cette approche peut être utilisée pour représenter tous les indices numériques qui vont donner lieu à une classification en fonction d'une seule valeur de seuil. Les pixels dont la valeur est supérieure à ce seuil sont retenus et regroupés lorsqu'ils sont adjacents. Ils forment ainsi des ROI. Cette approche peut être utilisée pour des indices tels que le BAI pour identifier des zones brûlées ou le NDSI pour identifier des zones enneigées. Ces approches sont complémentaires et doivent être utilisées en fonction du cas d'utilisation et de l'indice à représenter. L'approche par pourcentage est utile pour associer des valeurs numériques à des zones géographiques et l'approche par collections de polygones sert à identifier les zones impactées par un type de phénomène.

4.2.1 L'ontologie landcover : associer des données contextuelles à des unités administratives via leurs polygones

4.2.1.1 Étude de la couverture du sol grâce aux données de Land Cover

Un des types de données fondamentaux pour de nombreuses applications utilisant les images de satellites d'observation de la Terre est la caractérisation de la nature biophysique de la couverture du sol terrestre (eau, cultures, zone urbaine, etc.), ou Land Cover (cf section 2.1.4.3). Il s'agit d'un indice calculé à partir des données spectrales d'images comme celles de la collection Sentinel-2. Cet indice est disponible sous forme de cartes produites par différents organismes après un traitement massif des images à différentes résolutions et une période donnée. Des exemples de bases de données de couverture du sol sont le Global Land Cover Share ; Corine Land Cover qui est fourni pour des années particulières (1990, 2000, 2006 et 2012), et les données française du Land Cover fournies par le CESBIO, qui sont publiées annuellement.

Une manière d'exploiter ces données est de calculer l'indice rattaché à une zone géographique particulière à une période donnée, comme par exemple une ville, de manière à identifier l'urbanisation. La comparaison des valeurs de cet indice d'année en année permet alors de repérer les changements dans la manière d'exploiter les sols, et de caractériser les images de cette zone.

Dans un premier temps, nous nous intéressons au calcul d'une couverture de sols d'une zone géographique donnée comme une ville. Il s'agit donc là d'un premier exemple d'exploitation de données calculées à partir d'images, disponibles dans des bases de données existantes, et utiles pour évaluer caractériser la couverture de sols d'une zone géographique , ou de le comparer entre deux dates données. Plus précisément, nous avons proposé un service qui calcule à la demande cette couverture sur une zone passée en paramètre pour un Land Cover daté.

4.2.1.2 Une ontologie pour les données Land Cover associées à une zone géographique

Pour représenter les différentes classes de landcover associées à une zone géographique, avec ses dimensions spatiale et temporelle, nous avons élaboré une ontologie modulaire à deux niveaux représentée sur la figure 4.1. Le premier niveau de cette ontologie contient uniquement des concepts de haut niveau et a pour but d'être spécialisée par des concepts de plus bas niveau. Le premier niveau fournit les concepts nécessaires pour représenter un jeu de données géolocalisées calculées pour une zone géographique, quelle que soit la nature de ces données. Ce jeu de données est enregistré comme tel grâce à la notion de dataSet, ce qui permet de dater et localiser le moment où le jeu de données a été constitué. Pour la représentation des dimensions spatiales et temporelles des données, l'ontologie fait appel à deux ontologies standards : OWL-Time et GeoSPARQL. En particulier, la zone géographique est représentée grâce à la classe `EOFeature`, elle même sous-classe de `geo:SpatialObject`, ce qui permet de lui associer une localisation sous forme de `geo:Geometry` par la propriété `geo:hasGeometry`. Cette localisation est de type geo :Polygon, qui est le format le plus répandu pour les unités administratives dans les bases de données géolocalisées.

Le second niveau de cette ontologie spécialise le premier niveau et permet de décrire un jeu de données correspondant à un landcover : c'est un cas particulier de données calculées, associées à une zone géographique et qui possèdent une dimension spatio-temporelle.

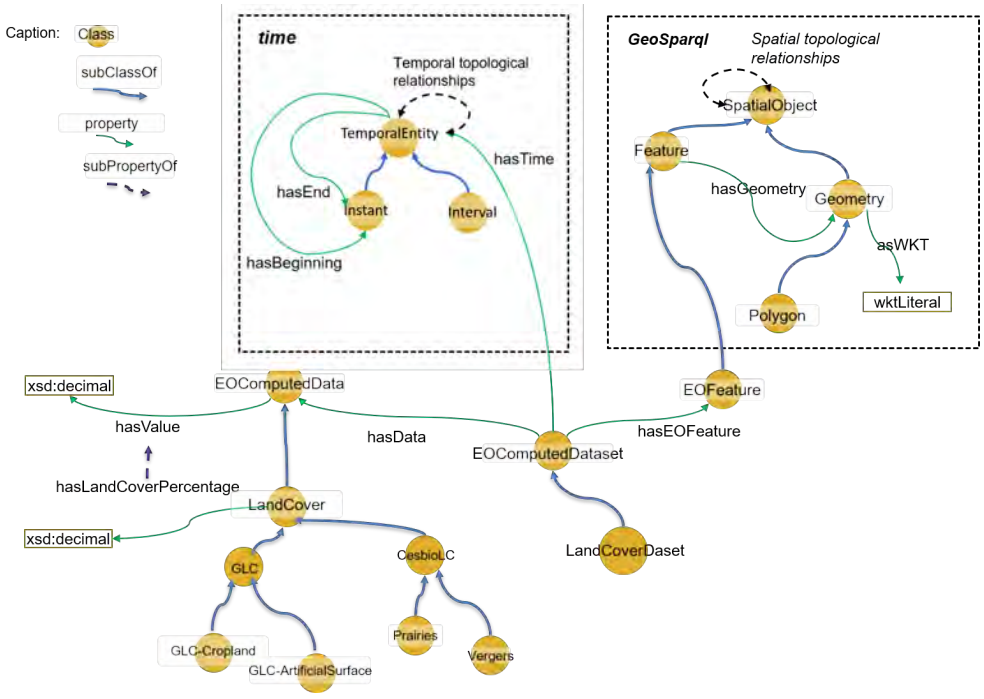


FIGURE 4.1 – Ontologie modulaire proposée pour la représentation du landcover

La classe `EOFeature` représente une entité géographique associée à une géométrie comme un point ou un polygone. Pour exprimer la localisation du jeu de données, une instance de type `EOFeature` est liée à une instance de la classe `EOComputedDataset` via la proprieté `hasEOFeature`. La classe `EOComputedDataset` est liée à toutes les informations concernant le jeu de données calculées que l'on veut représenter. Parmi ces propriétés, il y a la date de validité des données traduite par la propriété `hasTime` de l'ontologie OWL-Time ou les données utilisées en entrée pour le calcul. Chaque donnée calculée est représentée via la classe `EOCmputedData` est associée au dataset via la propriété `hasData`.

Une autre partie de l'ontologie permet la description d'un jeu de données land cover qui va être la source des données associées aux zones géographiques. Il faut bien noter que le graphe RDF ne représentera pas le contenu de la base de données de land cover en RDF, mais bien une utilisation de ces données pour définir un land cover associé à des zones géographiques fournies à l'application.

Plusieurs travaux ont proposé des vocabulaires pour représenter le land cover. Dans [Espinoza-Molina *et al.* 2015], les classes du land cover CORINE sont représentées comme des concepts de l'ontologie au sein d'une taxonomie. Le projet européen INSPIRE a également mis en place une représentation du land cover CORINE [2] avec le vocabulaire Simple Knowledge Organization System (SKOS).

Pour notre étude, nous avons considéré deux types de land cover avec des résolutions différentes :

— Le land cover glsglc-share [3] couvre l'ensemble de la planète avec une résolution spatiale de 1km. Ce land conver est produit par la FAO et la dernière version disponible date de 2014. Il possède 11 classes comme `Grassland`, `Waterbodies`, ou `Mangroves`.

— Les land cover produits par le laboratoire CESBIO couvrent la France uniquement avec une résolution spatiale de 10m. Chaque land cover couvre une période d'une année calendaire. Ils sont disponibles pour toutes les années entre 2016 et 2019. Ces land cover possèdent plus de 17 classes comme `Vineyards`, `Road surfaces` ou `Sunflower`.

Pour l'intégration des land cover dans notre ontologie, nous avons choisi qu'une valeur de land cover d'une certaine classe représente le pourcentage de pixels de la zone géographique considérée dont la valeur est cette classe.

Dans notre ontologie, la classe `lci:LandCover` représente donc un pourcentage de pixels ayant une certaine une valeur de land cover (`lci` pour land cover information). Un type de land cover (comme `GLC` ou `CesbioC`) est une sous-classe de la classe `LandCover`, et les natures possibles de couvert végétale d'un type de land cover sont définies comme leurs sous-classes. Autrement dit, la classe `lci:LandCover` de cette ontologie est spécialisée en fonction des classes du land cover considéré. La valeur calculée du pourcentage est liée à une instance de

2. `https://www.w3.org/2015/03/corine`

3. `http://www.fao.org/land-water/land/land-governance/land-resources-planning-toolbox/category/details/en/c/1036355/`

cette classe via la propriété `lci:hasLandCoverPercentage` qui est une spécialisation de la propriété `owl:hasValue`. Les deux types de land cover sont représentés avec les classes `lci:LandCoverDataset` qui est une spécialisation de la classe `lci:EOComputedDataset` présente dans le premier niveau de cette ontologie. Avec ce modèle, toute entité géographique représentée avec la classe `lci:EOFeature` peut avoir une valeur pour chacun des types de couverture du sol du land cover CESBIO ou GLC-SHARE.

Cette ontologie a été présentée dans l'article [Dorne *et al.* 2020]. Elle est accessible en ligne en suivant ce lien : `http://melodi.irit.fr/ontologies/lci.owl`.

4.2.2 L'ontologie NDVI : associer des données extraites d'images à des tuiles d'images

4.2.2.1 Associer plusieurs types de données à des images d'observation de la Terre

La grande majorité des données du programme Copernicus (images des satellites Sentinel-1 et Sentinel-2 entre autres) sont disponibles et accessibles gratuitement par tous [4]. Leur disponibilité ainsi que les nombreuses autres sources de données géolocalisées accessibles via le web ouvre de nombreuses perspectives d'applications dans des domaines variés. Ces applications peuvent en effet bénéficier à la fois des méta-données des images (telles que la couverture nuageuse), d'une analyse automatique du contenu des images (par exemple pour y repérer la végétation, et calculer des indices comme le land cover) et de données ouvertes géo-localisées et datées pouvant être "situées" sur ces images (données territoriales ou météorologiques, lieux, etc.).

Pour contribuer à associer aux images ces trois types de données (fournies et extraites des images ou collectées sur le web) à l'aide des technologies sémantiques, nous avons défini une ontologie plus complète et plus générique que celle développée pour représenter le land cover. Ce vocabulaire permet de représenter les données des différentes sources envisagées et d'y accéder de façon homogène. Nous nous sommes intéressés à un indice particulier, le NDVI, qui reflète la couverture végétale d'une surface à l'échelle du pixel. Comme pour le land cover, la représentation sémantique ne reflète pas le NDVI de chaque pixel, mais un indice recalculé pour des zones géographiques plus grandes. Dans le cas de cette étude, nous avons associé l'indice à des tuiles d'images, c'est-à-dire des zones d'une grille qui découpe la surface terrestre en rectangles de taille identique. Les tuiles des images Sentinel-2 correspondent à la surface d'une image entière.

4.2.2.2 Une ontologie pour représenter l'indice NDVI associé à des tuiles d'images

L'ontologie que nous avons développée réutilise des modules de l'ontologie de [Arenas *et al.* 2018], c'est-à-dire les modules *sosa*, *geo*, *grid*, *eom*. Nous les présen-

4. `http://copernicus.eu/data-access`

tons dans la suite, et dans un deuxième temps nous détaillerons le module *ndvi* que nous avons ajouté.

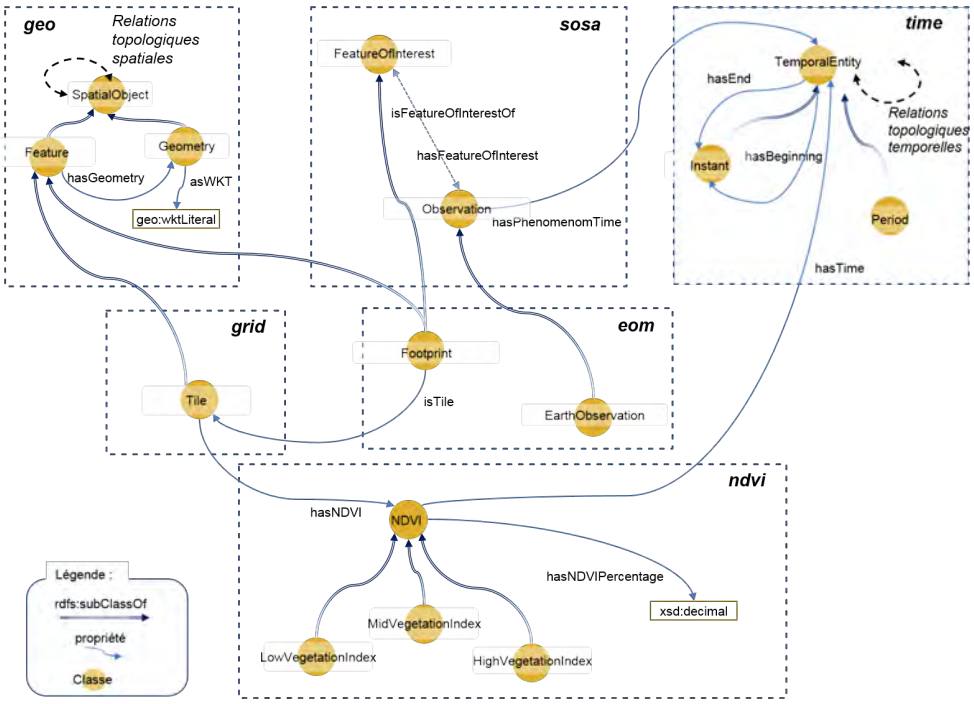


FIGURE 4.2 – Ontologie modulaire proposée pour la représentation du NDVI

Comme dans l'ontologie *LandCover* (partie 4.2.1), le module *geo* de la figure 4.2 sert à représenter la dimension spatiale des données à intégrer et réutilise GeoSPARQL présenté dans la section 3.1.1.4. Il en reprend la classe `geo:Feature` pour représenter toute entité ou donnée géo-localisée à laquelle est associée une `geo:Geometry`, et des relations spatiales entre les géométries définies par RCC8 présentées dans la section 3.1.1.1. Nous utilisons ces relations pour lier deux ressources géo-localisées représentées comme instances de `geo:Feature`. Les zones sur lesquelles sont calculés un indice comme le NDVI sont des instances de cette classe.

Afin de représenter le concept d'image satellitaire, nous réutilisons une partie de l'ontologie standard SOSA. Le module SOSA développé par [Janowicz *et al.* 2019] permet de représenter toute observation (`sosa:Observation`) réalisée par un capteur sur une "entité" (`sosa:FeatureOfInterest`), comme une activité datée (`sosa:phenomenomTime`), et produisant un résultat (propriété `sosa:hasResult`). Cette ontologie est une simplification de l'ontologie Semantic Sensor Network (SSN) qui permet aussi de décrire une observation scientifique issue de capteurs mais celle-ci possède de nombreuses dépendances. Les avantages majeurs de SOSA sont sa légèreté et sa modularité. Elle comprend 6 modules permettant de lier ses concepts à d'autres standards comme PROV Ontology (PROV-O) [5] ou Basic Formal Ontology

5. `https://www.w3.org/TR/prov-o/`

(BFO) [6].

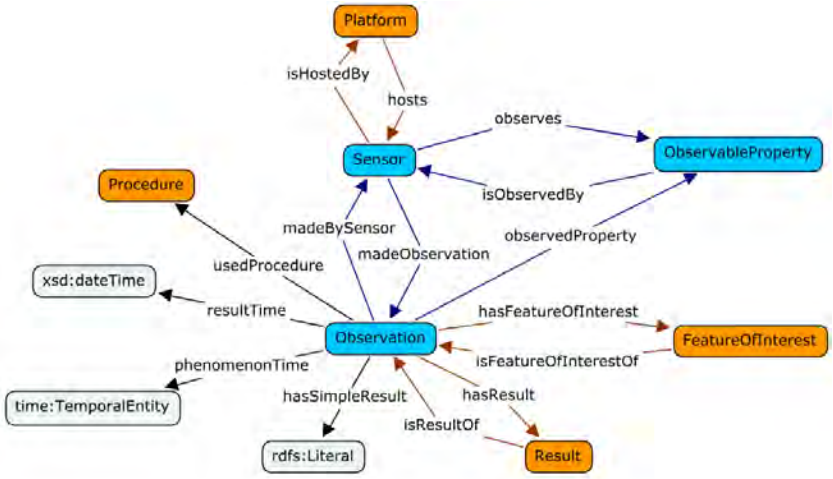


FIGURE 4.3 – Principaux concepts de l'ontologie SOSA par [Janowicz *et al.* 2019]

Notre ontologie NDVI réutilise le module SOSA-core qui contient les classes et propriétés présentées dans la figure 4.3, en particulier les concepts `FeatureOfInterest` et `Observation`. Une image satellitaire est considérée comme le résultat (propriété `sosa:hasResult`) d'une `Observation` réalisée par un capteur situé sur un satellite ; l'emprise au sol de l'image est représentée comme la `FeatureOfInterest` de l'observation.

Une des particularités de cette ontologie est le concept de grille dont chaque tuile (`grid:Tile`) permet d'agréger toutes les données se rapportant à une même zone, notamment les données extraites/calculées à partir de différentes images de la même tuile. Le module *grid* décrit les tuiles définies par le système de grille de l'ESA [7]. Chaque tuile est représentée comme une instance de la classe `grid:Tile`, spécialisation de `geo:Feature` dont la propriété `geo:hasGeometry` fournit l'emprise au sol sous forme d'un polygone fermé défini par des coordonnées au format EPSG :4326.

Le module *eom* sert à représenter les méta-données les plus importantes des images (module simplifié à deux classes sur la Figure 4.2). La propriété `eom:isTile` permet d'associer une tuile au `eom:Footprint` de chacune des images Sentinel 2 : un footprint représente la zone géographique couverte par une image sous forme d'un polygone fermé. Ce footprint est considéré par ailleurs comme l'élément d'intérêt d'observations réalisées par des capteurs du satellite ayant pris l'image ; la classe `eom:Footprint` spécialise ainsi les classes `geo:Feature` et `sosa:FeatureOfInterest`. Le `eom:Footprint` d'une image est ainsi daté via la propriété `sosa:phenomenomTime` qui identifie le moment où l'image a été prise.

Adapter l'ontologie à la représentation d'un nouveau type de données (comme les

6. `http://www.obofoundry.org/ontology/bfo.html`
7. `https://sentinel.esa.int/web/sentinel/missions/sentinel-2/data-products`

indices de végétation, ou ndvi) revient à ajouter à ces modules un nouveau module pour ces données, tel que le module *ndvi*, et à le lier aux modules existants. Chaque cadre de la Figure 4.2 présente un extrait du vocabulaire d'un module exploité pour associer les indices NDVI aux images Sentinel-2.

Au sein du module *ndvi*, l'indice de végétation est représenté par la classe `ndvi:NDVI`. Les propriétés spatiale et temporelle d'un indice de végétation sont fournies respectivement par la tuile à laquelle il est associé par la propriété `ndvi:hasNDVI` et par la cible de la relation `time:hasTime`.

Les algorithmes de calcul du NDVI [8] produisent des valeurs comprises entre -1 et 1 au niveau des pixels. Les valeurs comprises entre -1 et 0 signalent principalement des éléments constitués d'eau comme les lacs, rivières ou nuages. Les valeurs comprises entre 0 et 0.25 représentent principalement les éléments constitués de roche ou de terre. Les valeurs comprises entre 0.25 et 1 correspondent en grande partie à la présence d'éléments végétaux.

La figure 4.4 montre la classification des différentes valeurs de NDVI que nous avons définies de manière à passer d'un indice NDVI numérique connu pour chaque pixel d'une image à une indication globale pour l'ensemble de la zone concernée, à savoir une tuile. Notre modélisation vise à ne représenter que les indications relatives à la végétation, ce qui correspond à des valeurs de NDVI entre 0,25 et 1, appartenant aux trois intervalles colorés en vert sur la figure 4.4). Cette représentation est réalisée à l'aide de trois propriétés indiquant, pour une zone géographique donnée, le pourcentage de pixels correspondant aux classes représentant ces 3 intervalles : `LowVegetationIndex`, `MidVegetationIndex` et `HighVegetationIndex`.

Ce processus se déroule en deux temps : tout d'abord, en fonctione de la valeur de son indice NDVI, chaque pixel est classé dans une des trois catégories `LowVegetationIndex`, `MidVegtationIndex` et `HighVegetationIndex`. Ensuite, pour la zone donnée, on calcule le pourcentage de pixels de chacune de ces trois catégories. La classification d'un pixel P ayant une valeur de NDVI V s'effectue selon le principe suivant :

— Si V est comprise entre 0.25 et 0.50, P est catégorisé de `LowVegetationIndex`

— Si V est comprise entre 0.50 et 0.75, P est catégorisé de `MidVegetationIndex`

— Si V est comprise entre 0.75 et 1, P est catégorisé de `HighVegetationIndex`

Donc pour une tuile donnée, ses propriétés `LowVegetationIndex`, `MidVegtationIndex` et `HighVegetationIndex` contiennent respectivement le pourcentage de pixels de la tuile classés `LowVegetationIndex`, `MidVegetationIndex` ou `HighVegetationIndex`.

Cette ontologie est présentée dans l'article [Dorne *et al.* 2018].

8. `https://desktop.arcgis.com/fr/arcmap/latest/manage-data/raster-and-images/ndvi-function.htm`

FIGURE 4.4 – Classification des différentes valeurs de NDVI

4.2.3 Modélisation des rasters de changements et leurs données contextuelles

Dans cette troisième étude, la problématique est différente : il ne s'agit plus de caractériser globalement une zone géographique définie a priori à l'aide d'un indicateur donné à l'échelle du pixel et recalculé pour cette zone. Il s'agit de se servir des valeurs de cet indicateur par pixel pour identifier des régions formées de pixels voisins ayant des valeurs identiques ou proches et significatives par rapport à un phénomène à observer. Ces régions sont appelées *Régions d'Intérêt* (ROI). Plus précisément, nous définissons la notion de ROI comme une région de la Terre identifiée grâce à une analyse d'image à partir d'un fichier raster relatif à un indice ou un indicateur (de changement par exemple). Le principe de calcul d'une région d'intérêt est qu'elle est l'enveloppe rectangulaire d'une surface composée de pixels dont la valeur est supérieure à un certain seuil et formant une surface totale supérieure à un minimum. Le seuil et la surface minimum sont fixés en fonction du phénomène étudié à l'aide des changements.

Comme dans les expériences précédentes, les informations relatives aux indicateurs sont disponibles au format raster initialement et des données ouvertes peuvent venir enrichir la description de l'image. Une ontologie est nécessaire pour représenter ces informations sous forme de graphe de connaissances. Les deux premières ontologies *LandCover* et *Ndvi* ont inspiré la construction de cette troisième ontologie : on y retrouve la même structure modulaire, la même représentation des images et des données contextuelles. Mais l'ontologie se démarque par la présence de la notion de ROI.

Ce vocabulaire possède un module pour chaque type de données à représenter : données de raster correspondant à un indice ou indicateur de changement ; données ouvertes formant des données contextuelles (issues de bases de données ou calculées à partir d'images ou extraites de réseaux sociaux). Nous avons prévu de définir une classe par type de données contextuelles prises en compte. Cette ontologie peut être adaptée à de nouveaux types de données contextuelles : il suffit de modifier les classes de l'ontologie selon l'indice ou les données ouvertes utilisées. Dans le cadre de cette étude, nous nous sommes intéressés à l'identification de feux à partir de changements observés entre deux images. Afin de confirmer que le changement correspond bien à un feu, nous utilisons plusieurs types de données contextuelles : d'une part, un indice calculé à partir de l'image et qui estime une probabilité de feu, le BAI (pour Burned Area Indice) ; d'autre part, des données ouvertes, comme des nomes de lieux tirés de OSM ou les firepoints, qui sont des points de feu répertoriés

dans une base de données publique ; et enfin, des données de réseaux sociaux servent à repérer des termes confirmant qu'un feu a été identifié dans un des lieux reconnus sur la zone concernée.

Sur la figure disponible dans l'annexe A.1 sont représentés 5 modules différents (en vert) correspondants aux données du raster de changements et aux données contextuelles retenues pour évaluer l'étude : firepoints, indice BAI, unités administratives et données issues des réseaux sociaux. Ces modules réutilisent les ontologies standards OWL-Time et GeoSPARQL (en blanc). L'ensemble de ces données seront décrites dans le chapitre 5. Comme dans les ontologies *LandCover* et *Ndvi*, les concepts réutilisés de l'ontologie GeoSPARQL sont `geo:Feature` pour représenter les entités géolocalisées comme les tuiles, les ROI et les unités administratives. Le concept `geo:Geometry` permet de stocker la géométrie d'un concept au format WKT. L'ontologie OWL-Time est ici réutilisée pour la dimension temporelle des concepts. Le concept `ChangeRaster` sert à représenter un raster de changement. On lui associe un intervalle de temps correspondant aux deux dates des images utilisées pour générer ce même raster. Les concepts `SocialData` et `FirePoint` sont liés à un instant via la classe `time:Instant`.

De la même manière que pour le land cover ou le ndvi, les changements sont à l'origine des valeurs associés à chaque pixel du raster de changement, qui traduisent la probabilité que ce pixel ait changé entre deux images de la même région. Pour passer du format raster à une représentation sémantique, ces informations sont agrégées au niveau d'une région, qui dans cette ontologie est une ROI calculée en fonction de la surface et de l'importance du changement. On passe donc d'une valeur numérique du changement à une valeur symbolique, et seuls les changements considérés comme "high change" sont stockés en lien avec une ROI. c'est la propriété `o-change:hasHighCHange` qui associe une ROI à un raster de changement `ChangeRaster`.

Cette ontologie a été présentée dans l'article [Dorne *et al.* 2021].

4.3 Processus de transformation des données d'images en graphes de connaissances

Afin de représenter des données raster dans un format compatible avec le Web sémantique, nous avons mis au point une architecture et un processus méthodologique représenté dans la figure 4.5. Ce processus est constitué d'un ensemble de scripts qui s'appuient sur les ontologies définies dans la section précédente. Il prend en entrée un raster qui peut être un raster de NDVI, de land cover ou de changement. D'autres paramètres peuvent être nécessaires notamment pour calculer les régions d'intérêt à partir des données de raster. Enfin, des données contextuelles géolocalisées peuvent ensuite être reliées aux données RDF obtenues : données de réseaux sociaux, données disponibles sous forme de données liées ou encore d'autres indices calculés à partir des images concernées. Cette méthodologie est générique et peut s'appliquer à plusieurs types de données géospatiales du moment qu'elles sont

disponibles au format raster. En effet, les traitements utilisant les données d'entrée du processus pour générer des instances des classes de l'ontologie ou de relations entre ces instances sont exprimés de manière déclaratives sous forme de règles de création de données en RDF au sein de fichiers appelés *templates de triplification*.

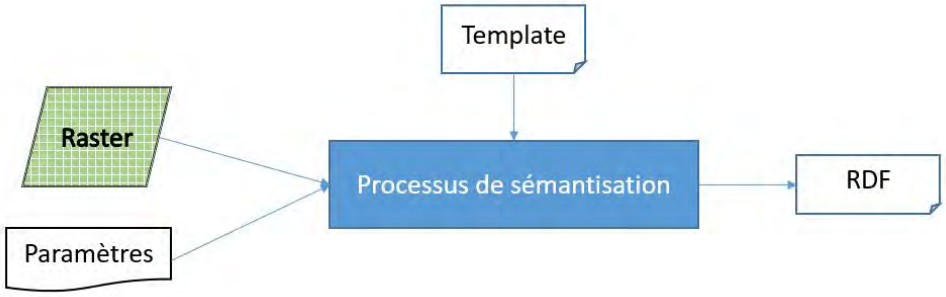


FIGURE 4.5 – Représentation graphique du processus de sémantisation des rasters

Pour adapter le processus à un nouveau jeu de données, il suffit de préciser ces données au sein de l'ontologie en spécialisant des classes ou en créant de nouvelles classes. Mais surtout, il faut adapter le template de triplification au modèle de données défini auparavant.

L'utilisation de ces templates pour le processus de sémantisation s'inscrit dans la continuité des travaux réalisés par [Arenas *et al.* 2016]. Pour chaque type de données, le template explicite la correspondance entre la ou les données et sa représentation sémantique selon des concepts et propriétés de l'ontologie concernée. En choisissant de développer un script par type de données à intégrer, nous possédons une architecture modulaire qui peut s'adapter aux données que l'on souhaite prendre en compte pour décrire les contenus d'images. Dans [Arenas *et al.* 2016], les données sont représentées au format JSON dans un premier temps puis au format RDF. Nous avons choisi de créer directement un fichier au format standard RDF qui sera ensuite intégré dans un triple store ou pourra être exploité tel quel.

4.3.1 Représentation sémantique du land cover par pourcentage de polygone

Pour cette première approche, nous avons défini un processus de sémantisation qui exploite un raster de land cover, des données dont on connaît la géométrie sous forme de polygone, et qui génère en sortie un graphe RDF associant une valeur de land cover à cette géométrie. Ce processus, présenté dans la figure 4.6, exploite l'ontologie land cover définie dans la partie 4.2.1. La valeur ajoutée est ici la mise en place d'une API permettant de calculer le land cover d'un polygone à la demande au format RDF. Cette API peut être exploitée par d'autres logiciels et possède une interface Web.

Les land cover sont des fichiers rasters accompagnés de fichier textes contenant les noms des classes ainsi que les valeurs des pixels associées à ces classes. Ces

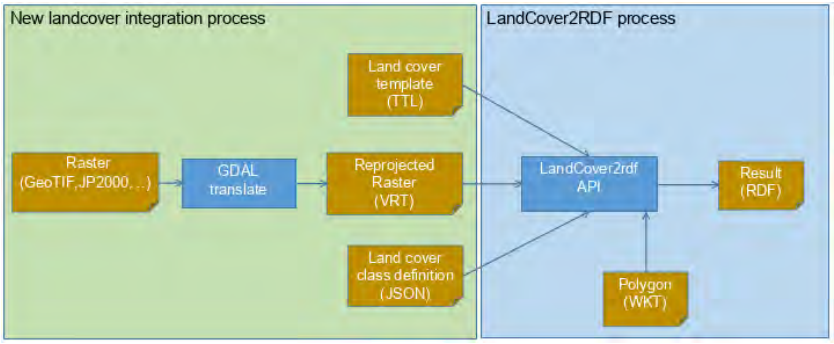


FIGURE 4.6 – Processus d'agrégation du land cover selon des polygones et de génération d'un graphe RDF

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix time: <http://www.w3.org/2006/time#> .
@prefix geo: <http://www.opengis.net/ont/geosparql#> .
@prefix lci: <http://melodi.irit.fr/ontologies/lci.owl#> .
@prefix g-lci: <http://melodi.irit.fr/lod/lci/> .
# ****************************************
dummy_cultureEte a lci:CESBIO-CultureEte .
dummy_cultureHiver a lci:CESBIO-CultureHiver .
dummy_eau a lci:CESBIO-Eau .
dummy_foretConiferes a lci:CESBIO-ForetConiferes .
....
# ****************************************
dummy_dataset a lci:LandCoverDataset .
dummy_dataset time:hasTime dummy_dataset_interval .
dummy_dataset_interval a time:Interval .
dummy_dataset_interval time:hasBeginning stringValueToTimeInstant($.↩
    startDate) .
dummy_dataset_interval time:hasEnd stringValueToTimeInstant($.endDate) .
# ****************************************
dummy_cultureEte lci:hasLandCoverPercentage valueToDecimalLiteral($.values.↩
    CESBIO-CultureEte) .
dummy_cultureHiver lci:hasLandCoverPercentage valueToDecimalLiteral($.↩
    values.CESBIO-CultureHiver) .
dummy_eau lci:hasLandCoverPercentage valueToDecimalLiteral($.values.CESBIO-↩
    Eau) .
dummy_foretConifere lci:hasLandCoverPercentage valueToDecimalLiteral($.↩
    values.CESBIO-ForetConifere) .
...
# ****************************************
dummy_dataset lci:hasData dummy_cultureEte .
dummy_dataset lci:hasData dummy_cultureHiver .
dummy_dataset lci:hasData dummy_eau .
dummy_dataset lci:hasData dummy_foretConiferes .
dummy_dataset lci:hasData dummy_foretFeuillus .
...
```

FIGURE 4.7 – Extrait du fichier template permettant la représentation RDF du land cover CESBIO

fichiers rasters peuvent avoir plusieurs formats en fonction de l'organisme qui les produit. Le land cover GLC-SHARE est au format GeoTIFF et utilise un système de coordonnées EPSG-4326 alors que le land cover produit par le CESBIO utilise un format de raster JPEG2000 et un système de coordonnées Lambert-93. Afin d'homogénéiser les systèmes de coordonnées, nous avons choisi de n'utiliser que le système EPSG-4326 qui est le plus répandu. Pour effectuer cette reprojection, nous avons utilisé la librairie GDAL qui permet de convertir un raster en Virtual Raster (VRT) [9]. Ce format de raster permet de ne pas recréer un autre raster pour la reprojection de coordonnées mais un fichier au format XML contenant uniquement les métadonnées modifiées et un lien vers le fichier raster original. Une fois le raster reprojeté, nous avons créé un fichier template au format Turtle dont un extrait est présenté sur la figure 4.7 pour le land cover CESBIO. Le mot-clé "dummy" sera ici remplacé au moment de la génération du fichier RDF par un URI créé dynamiquement grâce à la fonction md5 [10] appliquée sur le polygone.

Cette API est implémentée en Python grâce au framework Web Django [11]. Une interface Web [12] (figure 4.8) permet aux utilisateurs de choisir le land cover et les polygones plus facilement. Le polygone renseigné doit être au format WKT et dans le système de coordonnées WGS84/EPSG-4326. Le service va "découper" dans le fichier land cover choisi, et récupérer les valeurs du land cover pour les pixels situés dans le polygone renseigné. Il calcule ensuite le pourcentage de pixels de chaque classe pour l'ensemble du polygone avec la libraire GDAL. Le résultat est, pour chaque classe du land cover, une valeur en pourcentage de la zone. Le choix du land cover utilisé ici est d'une grande importance car ceux-ci n'ont pas tous la même résolution spatiale et les pourcentages peuvent ne pas être significatifs si le polygone choisi est mal dimensionné. Le land cover GLC-SHARE possède une résolution spatiale de 1km par pixel. Il ne peut donc pas être représentatif pour l'étude d'une petite ville par exemple.

Par cette approche, nous avons proposé un processus de génération de fichier RDF à partir de fichiers raster et de templates qui peut s'appliquer à d'autres fonctions. L'utilisation d'une API permet de ne pas stocker l'ensemble des données au format RDF et de pouvoir personnaliser la zone d'étude. Ce processus peut être réutilisé dans une chaîne de transformation de données automatique. Les connaissances ainsi générées peuvent servir à la détection de changements. La limite dans l'automatisation du processus de sémantisation reste le choix du land cover le plus approprié pour la zone à étudier.

9. https://gdal.org/drivers/raster/vrt.html
10. https://www.ietf.org/rfc/rfc1321.txt
11. https://www.djangoproject.com/
12. http://melodi.irit.fr/rasterStats

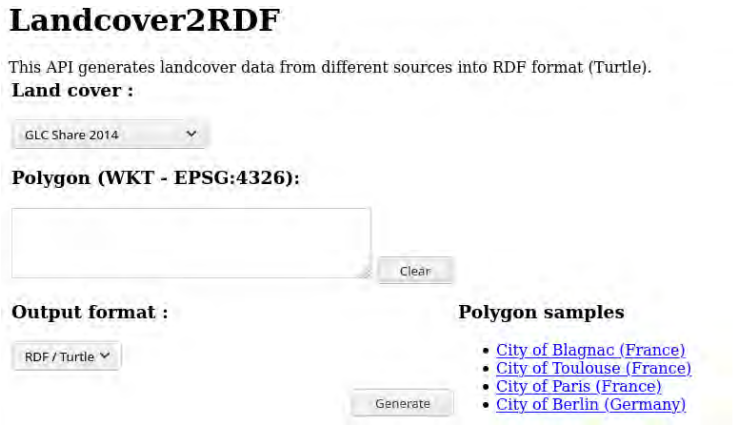


FIGURE 4.8 – Aperçu de l'interface de l'API Landcover2RDF

4.3.2 Représentation sémantique du NDVI par pourcentage de tuile

Afin de représenter au format RDF le NDVI d'une tuile calculé à partir d'une image Sentinel 2, nous avons développé le processus de sémantisation qui s'appuie sur l'ontologie présentée dans la section 4.2.2. Il utilise la tuile Sentinel 2 comme unité géographique de base car une image Sentinel 2 correspond à une tuile. Le fichier contenant l'ensemble des polygones des tuiles de la surface de la Terre est disponible sur le site de l'ESA [13]. Le NDVI est calculé avec un script Python pour l'ensemble de l'image grâce à la libraire GDAL. Ce script va ensuite calculer le nombre total de pixels dans l'image et compter le pourcentage de pixels dont les valeurs correspondent à une des trois classes de NDVI que nous avons définies dans l'ontologie. Chaque image possède 3 valeurs de pourcentage correspondant aux 3 classes de NDVI.

La figure 4.9 montre le fichier template que nous avons défini pour transformer les données au format JSON en un graphe RDF utilisant les classes et propriétés de l'ontologie *ndvi* (partie 4.2.2). Les valeurs `dummy_low`, `dummy_mid` et `dummy_high` sont remplacées par les URIs des instances à représenter. La fonction `getTileUrl($.tileId)` est une fonction du script Python qui construit l'URI de la tuile à partir de son identifiant. Ces URIs sont construites avec le préfixe "http ://-melodi.irit.fr/lod/" et l'identifiant de la tuile constitué de 2 chiffres et 3 lettres. Les fonctions `valueToDecimalLiteral` et `stringValueToTimeInstant` permettent de formaliser les valeurs décimales et la date avec la syntaxe RDF. Le fichier RDF obtenu est ensuite inséré dans le triplestore Virtuoso pour pouvoir l'exploiter à l'aide de requêtes SPARQL et notamment effectuer une détection de changement entre les différentes valeurs de NDVI.

13. `https://sentinel.esa.int/documents/247904/1955685/S2A_OPER_GIP_TILPAR_MPC_20151209T095117_V20150622T000000_21000101T000000_B00.kml`

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix ndvi: <http://melodi.irit.fr/ontologies/ndvi.owl#> .
@prefix time: <http://www.w3.org/2006/time#> .
# ****************************************
# this template to define the NDVI levels and associate it to a tile
dummy_low a ndvi:LowVegetation .
dummy_mid a ndvi:MidVegetation .
dummy_high a ndvi:HighVegetation .
# ****************************************
#Link NDVI to tile
getTileUrl($.tileId) ndvi:hasNdvi dummy_low .
getTileUrl($.tileId) ndvi:hasNdvi dummy_mid .
getTileUrl($.tileId) ndvi:hasNdvi dummy_high .
# ****************************************
#Get the ndvi percentage for this tile
dummy_low ndvi:hasNdviPercentage valueToDecimalLiteral($.lowVegetation) .
dummy_mid ndvi:hasNdviPercentage valueToDecimalLiteral($.midVegetation) .
dummy_high ndvi:hasNdviPercentage valueToDecimalLiteral($.highVegetation) .
# ****************************************
#Get the ndvi time for this tile
dummy_low time:hasTime stringValueToTimeInstant($.date) .
dummy_mid time:hasTime stringValueToTimeInstant($.date) .
dummy_high time:hasTime stringValueToTimeInstant($.date) .
```

FIGURE 4.9 – Fichier template utilisé pour la représentation sémantique du NDVI

4.3.3 Représentation sémantique du changement et données contextuelles par collections de ROI

4.3.3.1 Description du processus

Pour cette partie de nos travaux, nous utilisons un raster de changement issu de la détection de changement par un algorithme d'apprentissage non-supervisé décrit dans [Aubrun *et al.* 2020]. Ce raster débouche sur une représentation sémantique associée à l'image en exploitant les régions d'intérêt (ROIs). Nous avons choisi de ne garder que les valeurs les plus élevées de changement dans le but de pouvoir identifier un événement significatif, comme un feu de forêt ou une inondation. Cette identification est possible grâce à un ensemble de données contextuelles (indices et données ouvertes) qui permettent d'ajouter de la connaissance au raster. Ces données sont issues du traitement d'images ou de l'open data. Ce processus, présenté en figure, se base sur le processus présenté dans la figure 4.5. Il comporte 4 étapes :

— Identifier les ROIs

— Calculer les indices sur les images

— Récupérer des données ouvertes pertinentes

— Générer les graphes de connaissances

On peut voir sur la figure 4.10 que les données en entrée sont de sources hétérogènes et de formats différents. Les résultats obtenus sont des graphes de connaissances au format RDF exploitables avec les technologies du Web sémantique comme SPARQL.

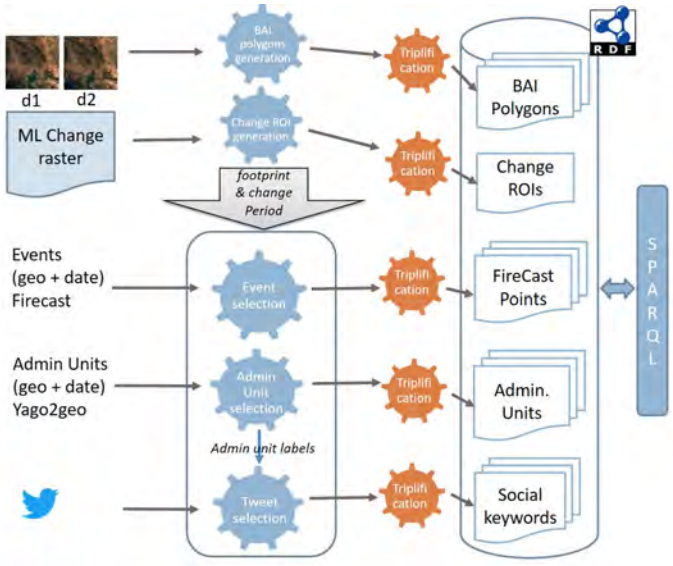


FIGURE 4.10 – Processus de sémantisation du changement et données contextuelles

4.3.3.2 Extraction des ROIs

Un raster de changement est produit à partir de deux images satellitaires partageant la même emprise au sol et acquises à des dates différentes. La surface couverte et les dates des images font partie des métadonnées des images que nous sauvegardons. Chaque pixel du raster possède une valeur correspondant à une probabilité de changement. L'image 4.11 représente un exemple de fichier raster issu de la détection de changement non supervisée. Les valeurs contenues dans ce raster sont comprises entre 0 et 1. Plus la valeur est proche de 1, plus la probabilité d'un changement est importante. A partir de ce fichier raster, l'algorithme produit un ensemble de ROIs représentant uniquement les plus fortes probabilités de changements. Le but de ces polygones est de pouvoir identifier automatiquement dans une image la position d'un ou plusieurs évènements potentiels. Pour des raisons de performance de calcul et d'interrogation, nous avons choisi de représenter ces ROIs sous forme de rectangles. Une fois ces ROIs identifiées, le processus les relie à des données contextuelles.

Pour générer les ROIs à partir du raster, nous avons développé l'algorithme 1 constitué de deux parties.

Du raster aux polygones Cette première étape consiste à générer un masque du raster ne contenant que les pixels ayant une valeur supérieure au paramètre `threshold` définit par l'utilisateur (ligne 1). Ce masque est ensuite transposé au format shapefile grâce à la fonction `Polygonize` de la librairie GDAL (ligne 2). Cette fonction permet de transposer des données depuis le format raster vers le format vectoriel. Ce fichier shapefile contient une quantité importante de polygones pouvant avoir la taille d'un seul pixel.

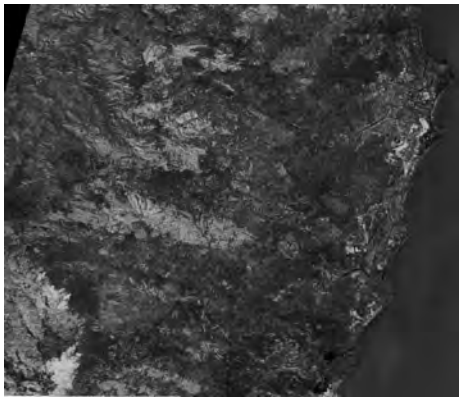


FIGURE 4.11 – Exemple d'un fichier raster issu de la détection de changement non supervisée

Construction des ROIs A partir des polygones contenus dans le shapefile, l'algorithme exclut les polygones en fonction de leur taille : il élimine les plus petits polygones (lignes 3 à 8), inférieur à une surface minimale fournie en paramètre. L'algorithme compare la surface de chaque polygone à un paramètre `minArea` défini par l'utilisateur. Cette surface est exprimée en mètres carrés. L'algorithme exécute ensuite un processus de simplification et agrégation de ces polygones (lignes 9 à 14). La simplification utilise la fonction `Envelope()` (ligne 12) pour obtenir l'emprise géométrique minimale de chaque polygone sous forme de rectangles. La seconde étape exécute la fonction `CascadedUnion()` qui permet d'agréger plusieurs polygones si leurs géométries s'intersectent. Ces deux opérations sont répétées jusqu'à ce que le nombre de polygones de la liste ne change plus, ce qui signifie que les polygones ne peuvent plus êtres simplifiés. Le résultat est donc une liste de polygones disjoints avec des géométries rectangulaires.

Algorithm 1 Construction des ROIs à partir d'un raster

1: $highChangeRaster \leftarrow FilterHighChange(raster, threshold)$
2: $listPolygons \leftarrow Polygonize(highChangeRaster)$
3: $listROIs \leftarrow \emptyset$
4: **for** each *polygon* in $listPolygons$ **do**
5: **if** $polygon.area >= minArea$ **then**
6: $listROIs.add(polygon)$
7: **end if**
8: **end for**
9: $nbPolygons \leftarrow 0$
10: **while** $len(listROIs) \neq nbPolygons$ **do**
11: $nbPolygons \leftarrow len(listROIs)$
12: $listROIs \leftarrow Envelope(listROIs)$
13: $listROIs \leftarrow CascadedUnion(listROIs)$
14: **end while**
return $listROIs$

La figure 4.12 illustre sur un exemple les résultats des différentes étapes de l'algorithme. L'image (a) représente le shapefile résultant de la fonction `Polygonize` appliquée au fichier raster de changements. Ce fichier contient plusieurs milliers de polygones disjoints et de très petite taille. Les images (b) à (g) sont les résultats des instructions de simplification et d'agrégation exécutées en 6 itérations. On observe que pour chaque itération, le nombre de polygones diminue, que leur surface augmente et qu'un polygone se démarque par sa taille. Le résultat de l'algorithme représenté dans l'image (g) montre que tous les polygones sont disjoints et de forme rectangulaire et en quantité raisonnable pour l'exploitation. Ces polygones sont les ROI identifiés à partir du fichier d'origine.

FIGURE 4.12 – Évolution des ROIs pour chaque itération du processus de simplification

4.3.3.3 Génération des graphes de connaissances

Le processus de génération des graphes des connaissances est similaire pour toutes les sources de données. La première étape consiste à extraire, pour chaque donnée, sa dimension spatiale et sa dimension temporelle. Ces données sont enregistrées dans une base de données NoSQL au format JSON. Nous avons défini pour chaque type de données un fichier template au format Turtle comme dans les approches présentées précédemment. Un script python explicite la correspondance entre les données stockées au format JSON et leur représentation au format RDF. Grâce à cette approche, il est facile d'ajouter de nouvelles données dans cette chaîne en définissant un nouveau template au préalable.

La figure 4.13 montre le fichier template utilisé pour la représentation d'un firepoint au format RDF. Dans ce fichier, on retrouve les propriétés `hasGeometry` de l'ontologie GeoSPARQL pour la dimension spatiale et la propriété `hasTime` de l'ontologie OWL-Time pour la dimension temporelle. Les propriétés `hasConfidence` et `hasType` sont propres au type de données et ne font pas partie d'une ontologie stan-

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix o-firecast: <http://melodi.irit.fr/ontologies/firecast.owl#> .
@prefix time: <http://www.w3.org/2006/time#> .
@prefix geo: <http://www.opengis.net/ont/geosparql#> .
# ********************************************************************
dummy a o-firecast:FirePoint .
# ********************************************************************
dummy geo:hasGeometry dummy_geo .
dummy_geo a geo:Geometry .
dummy_geo geo:asWKT valueToWktLiteral($.geometry) .
# ********************************************************************
dummy time:hasTime stringValueToTimeInstant($.date) .
# ********************************************************************
dummy o-firecast:hasConfidence valueToDecimalLiteral($.confidence) .
dummy o-firecast:hasType stringToLiteral($.type) .
```

FIGURE 4.13 – Fichier template permettant la représentation RDF d'un firepoint

dard. Tout comme dans les approches précédentes, le mot-clef `dummy` sera remplacé par une URI par le script python chargé de faire le lien.

4.4 Conclusions

Dans ce chapitre, nous avons montré plusieurs contributions concernant la traduction, sous forme de graphes de connaissances, de données tirées de l'analyse d'images d'observation de la Terre et regroupées en fonction de zones géographiques. Ces données sont décrites à l'aide d'une ontologie facilement adaptable à chaque cas. Pour montrer la diversité des problématiques soulevées, nous avons traité trois situations différentes : (i) représenter des indices calculés sur l'image en les associant à une tuile couvrant la surface de l'image ; (ii) représenter des indices calculés sur d'autres images et associés à des zones de l'image définies par des polygones ; (iii) repérer des zones d'intérêt ayant été le lieu de changements calculés entre deux images, qualifier leur nature et les lieux concernés à l'aide d'autres indices et de données ouvertes. Nous avons proposé pour chacune une ontologie modulaire définissant les classes et propriétés requises. Nous avons également défini et implémenté un processus de génération de graphes de connaissances en vue de l'interrogation de ces graphes.

Grâce à l'utilisation de templates pour la sémantisation des données, le processus d'ajout de nouvelles sources de données contextuelles est plus flexible. L'essentiel est de pouvoir retrouver les dimensions spatiales et temporelles sur chaque entité représentée sémantiquement. Notre objectif est de mettre en application ces différentes propositions sur d'autres jeux de données afin de les valider et de connaître leurs limites.

Chapitre 5

Expérimentations

Content

5.1 Cadre d'évaluation

L'objectif de ce chapitre est de présenter les différentes mises en application des travaux présentés dans le Chapitre 4. Ces expérimentations ont pour but d'apporter une évaluation aux différentes contributions faites dans cette thèse. Elles vont de la modélisation jusqu'à l'interrogation de données sémantiques en passant par la construction des graphes de connaissances. Le but est ici de montrer que les contributions sont valides et peuvent répondre à un besoin concret dans le cadre de différents scénarios. La notion de changement étant un élément clé de ces travaux, les scénarios choisis pour la mise en application mettent tous en avant un changement ou une évolution des données dans le temps. Le but est ici de pouvoir constater ce changement à partir des données intégrées et d'apporter du contexte à celui-ci, mais aussi de faciliter ainsi la recherche d'images ou de données sur ces changements. Les contributions évaluées s'appliquent aux trois cas d'études décrits dans le Chapitre 4 et portent sur trois aspects de la contribution :

— Les modèles et ontologies pour la représentation des données et leur intégration dans un système Web sémantique ;

— Le processus de sémantisation de données géolocalisées en réutilisant ces modèles ;

— Le processus d'interrogation des données via un environnement SPARQL.

5.1.1 Objectifs d'évaluation

L'évaluation porte sur les différentes approches proposées pour la représentation sémantique de données géospatiales : pour chacune, elle englobe la validation de l'ontologie, du processus de génération d'un graphe de connaissances à partir de données raster et des possibilités d'interrogation de ce graphe. Les trois approches proposées sont : la représentation par polygones complexes de données préexistantes (land Cover), la représentation par tuile de données calculées sur l'image (NDVI) et la représentation par ROI. Un autre objectif de cette évaluation par cas d'utilisation est de montrer la capacité de traiter des données géolocalisées de formats et sources hétérogènes. Ces données peuvent être des données ouvertes sémantiques (du LOD) ou non, ou bien extraites à partir des images satellitaires (indices NDVI). Enfin l'ensemble du processus doit aider les utilisateurs à obtenir des informations contextuelles sur un changement ou une évolution.

5.1.2 Méthodes retenues pour l'évaluation de chaque contribution

Pour évaluer ces contributions, nous avons choisi de réaliser plusieurs approches expérimentales avec des données in situ pour répondre à des besoins applicatifs. Ces expérimentations nous ont permis de valider les modèles de données que nous avons proposées précédemment notamment dans leur cohérence. Ces expérimentations permettent également de trouver les limites des approches proposées. L'objectif est d'obtenir une représentation sémantique de ces données in situ et de pouvoir les exploiter grâce aux technologies du Web sémantique. La validation se fera donc par l'exploitation de ces graphes de connaissances à l'aide de requêtes SPARQL qui représentent une exploitation des données par un utilisateur final. Les cas d'utilisation portent, dans un premier temps, sur l'évolution de la végétation sur une zone établie. Dans un second temps, une étude de l'évolution des sols au niveau d'une ville. Enfin, plusieurs scénarios permettent d'ajouter du contexte à des images portant sur des feux de forets.

5.2 Étude de l'évolution du NDVI par tuile

Pour cette étude, nous avons évalué le modèle proposé dans la section 4.2.2 pour la représentation du NDVI grâce à trois classes et par tuile Sentinel-2. Le but est de valider l'approche présentée dans la section 4.3.2 avec un cas d'utilisation. Nous avons choisi d'étudier l'évolution du NDVI pour la tuile 31TCJ sur une année. Cette tuile se situe dans le sud-ouest de la France et couvre une surface de 110 km^2. La période choisie se situe entre le 1er Mai 2017 et le 1er Mai 2018. Pour que le calcul du NDVI soit correct, nous n'utilisons que des images Sentinel-2 ayant une couverture nuageuse inférieure à 5%. Nous avons développé un script Python reposant sur l'API Copernicus pour télécharger automatiquement les images répondant à ces critères :

localisées sur la tuile choisie, dans la période définie et avec une couverture nuageuse inferieure à 5%. Nous avons obtenu 12 images pour la période. Pour chacune d'elles, nous avons calculé le NDVI et exécuté le processus de sémantisation 4.3.2. Le graphe de connaissances généré a ensuite été inséré dans l'entrepôt de données Virtuoso pour l'interroger via des requêtes SPARQL. La requête de la figure 5.1 permet de récupérer l'ensemble des données collectées sur la période.

```
select distinct ?time ?tileId ?ndviHighPercent ?ndviMidPercent ?ndviLowPercent
WHERE{
 ?ndviHigh a ndvi:HighVegetation .
 ?tileId ndvi:hasNdvi ?ndviHigh .
 ?ndviHigh time:hasTime ?ndviTimeInstant .
 ?ndviTimeInstant time:inXSDDateTime ?time.
 ?ndviHigh ndvi:hasNdviPercentage ?ndviHighPercent .
 ?ndviMid a ndvi:MidVegetation .
 ?ndviMid ndvi:hasNdviPercentage ?ndviMidPercent .
 ?ndviMid time:hasTime ?ndviTimeInstant .
 ?ndviLow a ndvi:LowVegetation .
 ?ndviLow time:hasTime ?ndviTimeInstant.
 ?ndviLow ndvi:hasNdviPercentage ?ndviLowPercent.
 FILTER (?time > "2017-05-01T00:00:00.00"^^xsd:dateTime AND ?time <
"2018-05-01T00:00:00.00"^^xsd:dateTime)
}
ORDER BY ?time
```

FIGURE 5.1 – Requête SPARQL permettant d'obtenir l'évolution du NDVI pour la tuile 31TCJ sur une année

La figure 5.2 montre les résultats retournés par cette requête. Les valeur de la colonne `time` correspondent aux dates de prise de vue des images Sentinel-2 utilisées pour le calcul du NDVI. On remarque qu'il n'y a pas d'intervalle de temps régulier entre les images disponibles sur une année si on se limite aux images avec une couverture nuageuse inférieure à 5% . L'URI des tuiles (colonne `tileId`) a été générée automatiquement par le script Python de conversion du format JSON en RDF. Nous retrouvons également les 3 classes de l'ontologie définies dans la section 4.2.2 ainsi que les valeurs calculées à partir des images.

Le graphique présenté dans l'image 5.3 fournit une représentation visuelle de l'évolution du NDVI pour la période. On note ici une baisse de la végétation dense, représentée par la classe `HighVegetationIndex`, entre les mois Mai et Novembre 2017. On remarque également une augmentation de cette même végétation entre Février et Mai 2018. Cette variation de végétation pourrait s'expliquer par le changement de saisons.

Grâce à ce cas d'utilisation, nous avons montré que l'ontologie NDVI pouvait être exploitée afin d'étudier une évolution de la végétation. Il s'agit ici de la première étape avant de pouvoir étudier un changement ou une tendance. Le processus de sémantisation permet d'exploiter directement des données issues des images sans être un expert en télédétection. De plus, le format standard RDF utilisé pour représenter ce graphe de connaissances permet son exploitation via le LOD.

L'inconvénient majeur de cette approche est le dimensionnement spatial qui restreint le choix du cas d'utilisation. L'ensemble des données est calculé pour une tuile d'une superficie de 110 km^2. Cette surface est pertinente uniquement si l'on souhaite étudier des évolutions sur des territoires de la taille d'une région ou d'un départe-

time	tileId	ndviHighPercent	ndviMidPercent	ndviLowPercent
2017-05-26T10:50:31.026	http://melodi.irit.fr/lod/grid_S2ST/31TCJ	18.22	47.09	16.91
2017-07-05T10:50:31.026	http://melodi.irit.fr/lod/grid_S2ST/31TCJ	16.11	35.22	23.81
2017-08-14T10:50:31.026	http://melodi.irit.fr/lod/grid_S2ST/31TCJ	8.19	36.38	31.73
2017-10-08T10:50:09.027	http://melodi.irit.fr/lod/grid_S2ST/31TCJ	2.66	37.25	23.83
2017-10-13T10:50:31.026	http://melodi.irit.fr/lod/grid_S2ST/31TCJ	0.45	35	26.2
2017-10-28T10:51:29.027	http://melodi.irit.fr/lod/grid_S2ST/31TCJ	0.38	31.12	26.38
2017-11-07T10:52:29.027	http://melodi.irit.fr/lod/grid_S2ST/31TCJ	0.33	29.3	28.77
2017-11-22T10:53:41.026	http://melodi.irit.fr/lod/grid_S2ST/31TCJ	0.16	14.7	38.44
2017-11-27T10:53:59.027	http://melodi.irit.fr/lod/grid_S2ST/31TCJ	0.23	10.77	43.09
2018-02-10T10:52:01.026	http://melodi.irit.fr/lod/grid_S2ST/31TCJ	0.44	18.17	55.53
2018-02-25T10:50:19.027	http://melodi.irit.fr/lod/grid_S2ST/31TCJ	0.24	15.84	58.84
2018-04-21T10:50:31.026	http://melodi.irit.fr/lod/grid_S2ST/31TCJ	12.48	54.44	15.11

FIGURE 5.2 – Résultat de la requête SPARQL sur l'évolution du NDVI pour la tuile 31TCJ sur une année

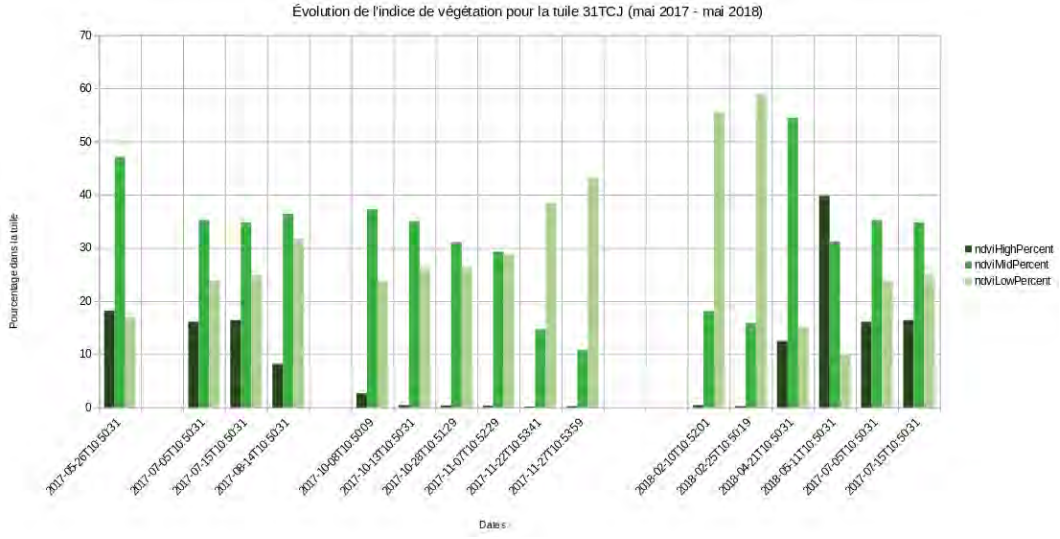


FIGURE 5.3 – Résultats graphiques de l'évolution du NDVI pour la tuile 31TCJ sur une année

ment. La résolution spatiale de 10 m peut poser problème pour l'étude d'une ville de petite taille par exemple. Pour réaliser une telle étude, il faut obligatoirement découper l'image ce qui introduit une étape supplémentaire. Un second inconvénient, relevé avec cette approche par tuile, est la difficulté d'ajouter des données provenant d'autres satellites comme Landsat-8 ou Sentinel-1 qui n'utilisent pas le

même référentiel spatial en tuile que Sentinel-2.

5.3 Étude de l'évolution du land cover dans le temps

Afin de valider le modèle de données proposé dans la section 4.2.1 ainsi que l'API présentée dans la section 4.3.1, nous étudions l'évolution du land cover au niveau d'une ville ou d'une région. Nous nous sommes intéressés à la ville de Blagnac dans le sud-ouest de la France. L'API que nous avons développée nécessite un polygone au format WKT et suppose de choisir un land cover pour l'année considérée. Nous avons récupéré le fichier WKT avec le polygone de la ville de Blagnac via le site Web de l'Institut National de la Statistique et des Études Économiques (INSEE). Pour ce cas d'utilisation, nous étudions l'évolution de la végétation dans la ville grâce aux classes des landcover CESBIO 2016 et 2017. L'image 5.4 représente un extrait de la commande exécutée pour obtenir le land cover de la ville de Blagnac pour l'année 2016 avec en paramètre le polygone au format WKT ainsi que le nom du land cover. La commande complète est disponible à cette adresse [1].

```
curl --data "datasetId=land_cover_cesbio_2016" --data "wkt=POLYGON↩
    ((1.3473654876713013 43.63352225951501, ...))" http://melodi.irit.fr/↩
    rasterStats/
```

FIGURE 5.4 – Extrait de la commande pour l'interrogation de l'API Landcover2RDF

Le résultat retourné par l'API est un graphe RDF qui contient pour chaque classe du land cover une valeur en pourcentage de la zone. Nous avons généré deux fichiers RDF, un premier pour l'année 2016 et un second pour l'année 2017.

Les raster land cover utilisés (du CESBIO) sont constitués à partir d'images satellitaires sur une année. Afin d'étudier une évolution temporelle, il faut utiliser le même land cover sur au moins deux années différentes. Un des problèmes est qu'il n'existe pas d'alignement entre les classes de land cover de sources différentes.

Afin d'exploiter ces deux graphes, nous avons choisi de les insérer dans l'entrepôt Virtuoso. Pour étudier l'évolution de la végétation et de l'urbanisation, nous avons additionné les valeurs des classes liées à la végétation comme `lci:CESBIO-CultureEte`, `lci:CESBIO-ForetConifere` ou `lci:CESBIO-ForetFeuillus`. De la même manière, avons additionné les valeurs des classes liées à l'urbanisation `lci:CESBIO-UrbainDense` ou `lci:CESBIO-UrbainDiffus`. Ces opérations sont réalisées par une requête SPARQL disponible dans l'annexe A.2.

Les résultats de cette requête sont présentés sur la figure 5.5. On remarque ici que le pourcentage de végétation diminue entre le 1er Janvier 2016 et le 31 décembre 2017 alors que le pourcentage d'urbanisation augmente dans la ville de Blagnac.

1. `http://melodi.irit.fr/rasterStats?query-blagnac`

landcoverStart	landcoverEnd	totalVegetationCESBIO	totalUrbainCESBIO
2016-01-01T00:00:01	2016-12-31T23:59:59.999	33.0485	59.4996
2017-01-01T00:00:01	2017-12-31T23:59:59.999	31.3683	62.988

FIGURE 5.5 – Résultat de la requête SPARQL permettant l'étude de l'évolution du land cover entre 2 années sur la ville de Blagnac

Les résultats de cette étude montrent une exploitation possible de l'API Landcover2RDF pour l'étude de changements dans la couverture végétale terrestre. La requête de ce cas d'utilisation démontre qu'il est possible de créer des indicateurs à partir des données collectées au format Web Sémantique. Ces indicateurs peuvent servir à élaborer des tendances afin de qualifier une évolution anormale détectée à partir des connaissances existantes. La principale limite observée réside dans le choix du polygone à étudier qui doit être connu et fourni par l'utilisateur. Une perspective envisagée est de créer un alignement entre les différentes classes de plusieurs land cover et de pouvoir comparer les évolutions à partir de ces différentes sources.

5.4 Mise en application du processus de sémantisation pour la détection d'évènements par ROI

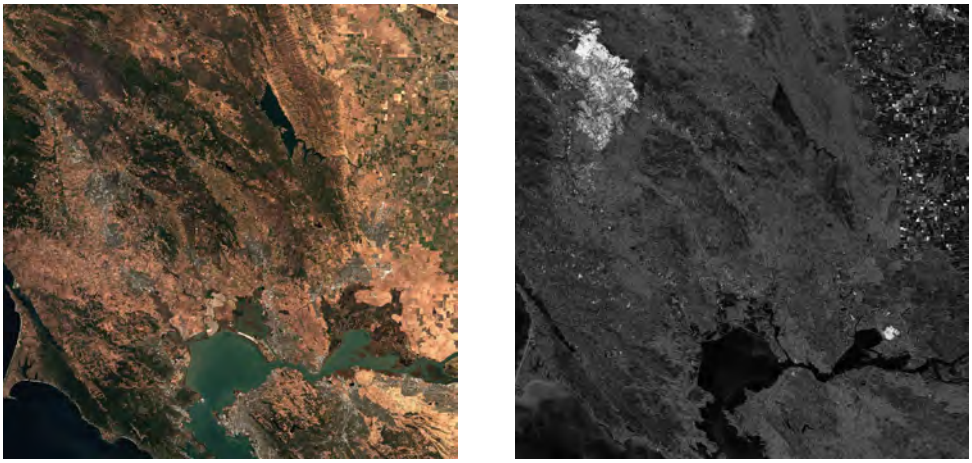
Cas d'étude	Image	Écart temporel	Phénomène étudié	Données contextuelles	Validation	Objectif
Kindcade Fire, CA	S2	10j	Feu	Tweets, Yago2Geo, BAI	Firecast, Manuelle	Mise au point paramètres ROI
Camp Fire, CA	S2	70j	Feu	Tweets, Yago2Geo, BAI	Firecast, Manuelle	Validation choix paramètres
Feu de forêt Turquie	S2	10j	Feu	BAI	Firecast, Manuelle	Validation choix paramètres
Bush fire Australie	S2	85j	Feu	Tweets, Yago2Geo, BAI	Firecast, Manuelle	Validation choix paramètres
Explosion Beyrouth	WorldView-2	5j	Explosion	Tweets	Manuelle	Étude d'un autre phénomène avec un autre type d'image

FIGURE 5.6 – Ensemble des cas d'utilisation pour l'expérimentation du processus d'aide à la qualification d'évènements

Cette section présente les différents cas d'utilisation mis en place afin de valider l'ensemble de la chaîne de traitement présentée dans la section 4.3.3 et de vérifier la faisabilité technique de cette approche. Nous avons testé l'ensemble de cette chaîne sur plusieurs cas d'utilisation listés dans la figure 5.6. Le premier cas d'utilisation nous a permis de déterminer les différentes valeurs de paramètres notamment pour

l'algorithme de création de ROI. Les autres cas d'utilisation nous ont permis de valider ces paramètres en les appliquant sur des données différentes. Nous avons pu ainsi comparer les résultats avec le premier cas. Le dernier cas d'utilisation nous a permis de tester la faisabilité de notre chaîne de traitement sur un autre évènement que les incendies et en utilisant un autre type d'image satellitaire. Pour chaque cas, nous avons utilisé un couple d'images satellitaires disponibles en ligne. Pour chaque couple, la première image est prise avant la date de début du phénomène et la seconde après la date de fin du phénomène. Dans certains cas, aucune image avec une couverture nuageuse faible n'était disponible proche de la date de début ou de fin du phénomène, ce qui explique les écarts temporels plus ou moins importants.

Nous avons étudié le feu de forêt *Kincade Fire* ayant touché la Californie en 2019. Cet incendie a eu lieu du 23 Octobre 2019 au 6 novembre 2019 dans le sud de la Californie et a brûlé plus de 30000 ha de forêt. L'algorithme de calcul du raster de changement utilisé est présenté dans [Aubrun *et al.* 2020]. Il a utilisé deux images Sentinel-2 : la première image date du 22 Octobre 2019, un jour avant l'incendie, et la seconde image date du 1er Novembre 2019, quelques jours avant la fin de l'incendie. La figure 5.7(a) correspond à cette 2ème image. Le raster de changement résultant est celui de la figure 5.7(b).



(a) (b)

FIGURE 5.7 – (a) Image Sentinel-2 du 01/11/2019 (b) Raster de changement résultant de l'algorithme de détection de changement non supervisé.

5.4.1 Construction des Régions d'intérêts

La première étape de la chaîne de traitement, une fois le raster de changement obtenu, est de construire les régions d'intérêts avec l'algorithme décrit dans la section 4.3.3.2. Cet algorithme prend en paramètre la valeur minimale d'un pixel pour que celui-ci soit considéré dans un polygone ainsi que la taille minimale qu'un polygone doit avoir pour être retenu comme région d'intérêt. Nous avons essayé différentes valeurs pour ces deux paramètres afin d'optimiser les résultats, le but étant d'obtenir un fichier shapefile contenant les ROIs du raster de changement. Il faut

que les polygones couvrent bien l'incendie et qu'il n'y ait pas de polygones représentant des changements considérés anodins comme un groupe de véhicules ayant été déplacés par exemple. Les résultats obtenus sont répertoriés dans le tableau de la figure 5.8.

Surface Min (m2) / high change threshold	0	200	400	600	800	1000	1200	1400	1600	1800	2000	5000	10000	15000
0,5	73942	37427	23246	16932	13224	9859	8358	7568	6942	6459	5726	2759	1478	1018
0,55	28607	16334	11246	8802	7232	5768	4949	4536	4177	3910	3578	1759	945	631
0,6	13197	8447	6276	5180	4411	3668	3228	2979	2770	2593	2367	1192	647	468
0,65	7047	4818	3694	3086	2689	2261	1973	1838	1708	1611	1462	781	484	332
0,7	4049	2862	2252	1903	1680	1419	1261	1183	1109	1042	959	516	391	291
0,75	2192	1580	1264	1102	993	867	766	726	696	652	606	469	290	226
0,8	1849	1321	1061	924	831	729	644	609	581	555	530	290	255	203
0,85	1735	1242	1011	883	873	770	692	646	610	572	537	285	180	165
0,9	1730	1081	1199	960	809	695	550	492	450	408	379	222	124	88

FIGURE 5.8 – Nombre de polygones obtenus par l'algorithme de création des ROIs en fonction des paramètres

De manière non surprenante, plus la valeur du paramètre `threshold` est faible, plus le nombre de polygones générés augmente. De la même manière, plus la surface minimale est petite, plus le nombre de polygones augmente. Lorsque les deux paramètres ont des valeurs trop basses, les temps de calcul s'allongent et atteignent plusieurs heures. Nous avons observé que lorsque la valeur du paramètre `threshold` dépassait 0.7, les polygones générés ne couvraient pas l'ensemble de la surface de l'incendie. Nous avons donc choisi une valeur de `threshold` de 0.65 pour couvrir l'ensemble de la surface de l'incendie. Concernant la valeur du paramètre de surface minimale de chaque polygone, nous avons fixé une valeur de 10000 m^2, ce qui correspond à 1 ha. Cette valeur signifie que tous les changements inférieurs à cette surface ne feront pas partie de la collection de ROIs générées. Nous avons fait ce choix en accord avec des experts en télédétection qui estiment que cette valeur est représentative pour des images Sentinel-2 avec une résolution spatiale de 10 m.

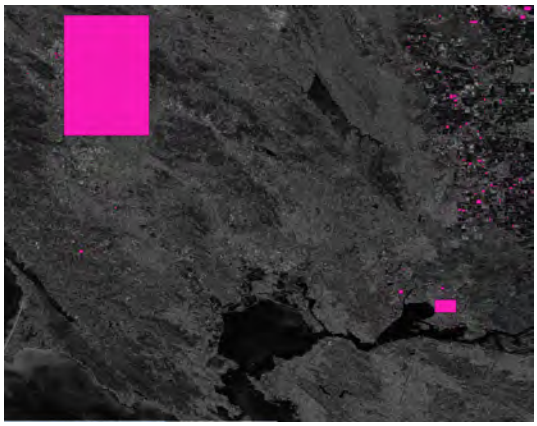


FIGURE 5.9 – ROIs générées par l'algorithme (rose) avec les paramètres `threshold`=0.65 et `minSurface`=10000m^2

La figure 5.9 montre les ROIs générées par l'algorithme avec les paramètres définis précédemment. Le fichier shapefile résultant contient 484 polygones et on

remarque ici que deux polygones se démarquent de par leur taille. Le polygone le plus important situé dans la partie supérieure gauche de l'image représente le feu de forêt Kincade Fire. Le polygone situé dans la partie inférieure droite représente un feu de champs. Les autres polygones de tailles inférieures correspondent à des changements anodins comme le moissonnage de champs ou le remplissage de bassins de rétention d'eau. Pour vérifier que la ROI correspondant au Kincade Fire couvrait bien la surface de l'incendie, nous avons téléchargé le fichier shapefile de ce feu depuis le site du gouvernement californien [2]. La figure 5.10 montre qu'un des polygones générés couvre en majorité la surface impactée par le Kincade Fire. La partie non couverte par la ROI s'explique par la date de la seconde image utilisée pour la détection de changements. Cette seconde image date du 1er novembre 2019. Or l'incendie a duré jusqu'au 6 novembre 2019. La partie qui n'est pas couverte par la ROI correspond à la zone brûlée entre le 1er et le 6 novembre 2019.

FIGURE 5.10 – Surface précise du feu de forêt Kincade Fire (en vert) superposée aux ROIs générés (en rose)

Concernant le feu de champs localisé au sud-est dans la figure 5.9, le fichier shapefile n'était pas fourni pour le gouvernement californien. Cet incendie a eu lieu le 27 octobre 2019 près de la ville de Lafayette en Californie et n'a pas été répertorié par les services de pompiers de Californie. En revanche, il a été couvert par les médias [3]. Nous avons choisi de générer un polygone à partir du raster de changement pour déterminer le périmètre de cet incendie. Ce polygone a été généré avec l'algorithme `Polygonize` de la librairie GDAL. La figure 5.11 montre que le polygone rose généré par l'algorithme de création de ROI couvre l'intégralité de la surface impactée par le feu en gris. L'ensemble de ces données a ensuite été converti en RDF et intégré à la base de connaissances.

5.4.2 Calcul d'indices à partir des images

L'étape suivante consiste à calculer des indices à partir des images satellitaires en fonction du phénomène à étudier. Pour l'étude des incendies, nous avons choisi de calculer le BAI (burned area index) sur la seconde image satellitaire utilisée pour la détection de changement. Cet indice nous permet d'étudier les zones qui ont été

2. `https://frap.fire.ca.gov/frap-projects/fire-perimeters/`
3. `https://www.mercurynews.com/2019/10/28/grizzly-island-fire-scorches-2300-acres-sends-smoke-across-`

FIGURE 5.11 – Surface du feu de champ (gris) couverte par les ROIs générées (rose)

impactées par un incendie. Pour calculer cet indice, nous avons utilisé l'équation 5.1 élaborée par [Filipponi 2018].

$$BAIS2 = \left(1 - \sqrt{\frac{B06 * B07 * B8A}{B4}}\right) * \left(\frac{B12 - B8A}{\sqrt{B12 + B8A}} + 1\right) \tag{5.1}$$

Un des problèmes de cette équation pour le BAI est qu'elle identifie les pixels représentant des surfaces composées d'eau comme étant des zones brûlées. Nous avons donc appliqué la méthode proposée par [Filipponi 2018] qui consiste à créer au préalable un masque de ces pixels pour que l'équation ne les prenne pas en compte. Pour créer ce masque, nous avons utilisé l'indice NDWI qui se calcule selon l'équation ci-dessous, tirée de 5.2, avec laquelle nous avons obtenu de meilleurs résultats qu'avec l'équation `Later Water Pixels` proposée dans [Filipponi 2018].

$$NDWI = \frac{(B03 - B08)}{(B03 + B08)} \tag{5.2}$$

Le calcul du BAI produit un fichier raster de la taille de l'image source utilisée avec des valeurs de pixel comprises entre -1 et 6. Les valeurs comprises entre -1 et 1 représentent les zones non impactées par un incendie. Les pixels avec une valeur comprise entre 1 et 6 représentent les zones brûlées.

Pour générer une liste de polygones représentant les zones brûlées dans l'image, nous avons appliqué au raster du BAI le même algorithme que celui pour les ROI de changement avec en paramètres une valeur de `threshold` de 1.01 et une surface minimale de polygone de 10000 m^2. La valeur du paramètre `threshold` a été définie pour ne récupérer que les éléments brûlés (valeur du pixel > 1). La figure 5.12 montre les régions d'intérêts générées à partir du raster BAI.

La figure 5.13 montre qu'un des polygones générés (en jaune) par l'algorithme de création de ROI couvre bien la majorité de la superficie totale de l'incendie (en vert). La partie inférieure qui n'est pas couverte correspond à une zone non identifiée par l'algorithme de calcul de l'indice BAI.

Grâce à ces polygones, il est possible d'identifier sur l'image les régions qui ont été impactées par un incendie. On remarque que les deux incendies identifiés par le raster de changement sont bien identifiés avec le BAI. Les autres polygones de plus petite taille correspondent à des incendies ou des feux de champs ayant eu lieu avant la période étudiée et ne sont pas pertinents pour notre étude. Ces données

FIGURE 5.12 – Polygones générés par l'algorithme de création de ROI à partir du raster BAI (jaune)

FIGURE 5.13 – Surface du feu de forêt Kincade Fire (vert) sur la ROI générée à partir du raster BAI (jaune)

sont finalement converties en graphe RDF et insérées dans notre triplestore.

5.4.3 Données contextuelles

Une partie du processus consiste à identifier le type de phénomène que l'on souhaite étudier afin de choisir les sources de données adaptées. Dans notre cas, pour les feux de forêts, les sources de données contextuelles utilisées sont la base de données Firecast, les lieux géolocalisés dans YAGO2Geo et les messages transmis sur le réseaux social Twitter. Nous précisons l'utilisation de chacun d'eux.

5.4.3.1 Firecast

Les données collectées par le site Web Firecast [4] proviennent des satellites MODIS et VIIRS qui permettent une étude pratiquement en temps réel des anomalies thermiques à la surface de la Terre [Giglio *et al.* 2003]. Les capteurs thermiques

4. `https://firecast.conservation.org`

présents sur ces satellites permettent d'identifier des incendies avec un indice de confiance compris entre 0 et 1. Le jeu de données fourni par Firecast est un fichier shapefile contenant les points chauds associés à leur indice de confiance et leur source pour une période déterminée. Ces points chauds sont mis à jour toutes les 30 minutes et sont horodatés.

Pour notre étude, nous avons téléchargé les données Firecast disponibles entre les dates des deux images utilisées pour la détection de changement. Afin d'éliminer les points chauds ne correspondant pas à des incendies, nous n'avons conservé que les points avec un indice de confiance supérieur à 0.65. Cette valeur de 0.65 nous permet d'écarter les faux-positifs générés automatiquement par l'algorithme de Firecast. Le fichier shapefile généré contient plus de 2000 points chauds entre les dates du 22 octobre 2019 et du 1er Novembre 2019. La figure 5.14 montre l'ensemble des points chauds téléchargés pour la zone étudiée.

FIGURE 5.14 – Points chauds issus du site Web Firecast pour le feu de forêt Kincade Fire

On remarque une importante concentration de points à l'endroit où a eu lieu le Kincade Fire ainsi que le feu de champs (en bas à droite). Cette visualisation nous permet déjà d'identifier deux incendies distincts pour cette période. Les points isolés correspondent à des feux mineurs et ne sont pas pertinents pour notre étude. L'ensemble de ces points est ensuite converti en RDF grâce au script Python et inséré dans notre base de connaissance.

5.4.3.2 YAGO2Geo

Une autre source de données contextuelles utilisée dans notre chaîne de sémantisation est la base de connaissance YAGO2geo présentée dans la section 3.1.3.3. Cette source de données du LOD nous permet de retrouver n'importe quelle unité administrative et de récupérer sa surface au sol (sa géométrie) sous forme d'un polygone à l'aide d'une requête SPARQL. Ces unités administratives sont organisées par niveaux avec les classifications GADM et OSM. La granularité des unités administratives varie du parc jusqu'à l'état.

Pour notre étude, nous avons extrait au format WKT l'emprise au sol d'une des images utilisées pour la détection de changement, puis nous avons écrit une requête SPARQL (Cf. figure 5.15) pour récupérer toutes les villes enregistrées dans YAGO2Geo situées dans ce polygone. Le résultat de la requête est un fichier au format XML contenant le nom, l'URI et le polygone au format WKT de 27 unités administratives. Nous avons ensuite converti ce fichier en un graphe RDF que nous avons intégré à notre base de connaissance pour une exploitation future.

```
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX y2geoo: <http://kr.di.uoa.gr/yago2geo/ontology/>

SELECT DISTINCT ?admin ?name ?wkt  ?type
WHERE {
  ?admin rdf:type ?type        .
  ?admin y2geoo:hasOSM_Name ?name . FILTER(LANG(?name) = "") .
  ?admin geo:hasGeometry/geo:asWKT ?wkt.
  FILTER(?type IN (y2geoo:OSM_city, y2geoo:OSM_town, y2geoo:OSM_locality))
  FILTER(geof:sfIntersects(?wkt, "MULTIPOLYGON(((-123.000230479793 ↩
      38.8489984833831,-123.000227364091 37.8594419588954,-121.752134591281↩
       37.8528295028533,-121.735039885204 38.842148672443,-123.000230479793↩
       38.8489984833831)))"^^geo:wktLiteral))}
```

FIGURE 5.15 – Requête SPARQL permettent de récupérer les noms de villes présentes dans l'image

5.4.3.3 Twitter

La troisième source de données contextuelles utilisée est le réseau social Twitter. Les mots-clés dans les tweets aident à identifier la nature d'un évènement dans la zone de l'image pour la période du changement identifié par le raster. Les tweets utilisés proviennent de la collection faite par le laboratoire Institut de Recherche en Informatique Toulousain (IRIT) sur la plateforme OSIRIM [5]. Cette plateforme sauvegarde un corpus contenant 1% des tweets mondiaux depuis 2016. Elle collecte entre 20 et 30 tweets par seconde sans aucun critère de restriction. Les tweets sont groupés par heure et stockés au format JSON. Ce corpus ne fournit pas une représentation exacte de la fréquence des mots-clé mais il permet d'établir une tendance.

Nous avons élaboré deux scripts Python pour la recherche de mot-clés dans les tweets. Le premier script, localisé sur la plateforme OSIRIM, est chargé de construire un fichier JSON constitué des tweets contenant le nom de l'unité administrative passé en paramètre entre deux dates. L'ambiguïté sur les noms des unités administratives n'est pas traitée dans ce processus. Nous ne prenons pas en compte volontairement la localisation des tweets car une importante majorité de tweets n'est pas localisée. Or nous souhaitons garder les tweets des personnes qui ne se situent pas à l'endroit de l'évènement. Une fois le fichier construit, nous le

5. https://osirim.irit.fr/

récupérons depuis la plateforme et nous exécutons un second script chargé de générer les mots-clés les plus fréquents. Ce script utilise la librairie Python open source Natural Language ToolKit (NLTK) pour la gestion des mots-vides et le comptage des mots les plus fréquents. Il renvoie les 5 mots-clés les plus fréquents associés à un nom de ville et une période.

Concernant le cas d'étude du feu de forêt, nous avons identifié deux villes en intersection avec la ROI la plus importante de l'image. Nous avons démarré le processus de recherche de mots-clés avec les noms de ville *Healdsburg* et *Windsor* pour la période de la prise de vue des images. Le script exécuté sur les données de la plateforme OSIRIM a retourné un fichier de 3976 tweets contenant le mot *Healdsburg* et un fichier de 1728 tweets contenant le mot *Windsor*.

Windsor
healdsburg
#kincadefire
castle
fire
firefighters

Healdsburg
evacuation
windsor
#kincadefire
fire
healdsburg

FIGURE 5.16 – Résultats retournés par le processus de recherche de mots-clés pour les villes de *Windsor* et *Healdsburg* pendant l'incendie Kincade Fire

La figure 5.16 montre les résultats obtenus par le processus de recherche de mots-clés pour la période de la détection de changement sur l'ensemble des tweets extraits pour les villes de Windsor et Healdsburg. On remarque ici les mots **fire** et **firefighters** qui font partie du champ lexical d'un incendie. Ces mots-clés sont ensuite représentés en RDF et liés à une période et à la ville concernée. Ces graphes de connaissances sont ensuite insérés dans la base de connaissances pour être exploités par la suite.

5.4.4 Exploitation et visualisation du résultat

La dernière étape du processus est l'exploitation des données par interrogation du graphe de connaissances. La totalité des graphes de connaissances générés dans les étapes précédentes ont été intégrés dans une même base de connaissances, ce qui permet leur exploitation grâce à des requêtes SPARQL. Nous avons également choisi de représenter certaines de ces données visuellement grâce au SIG QGIS.

La figure disponible dans l'annexe A.3 présente une partie des instances des graphes RDF générés. On remarque que l'ensemble des types de données est connecté à l'emprise du raster de changement (en jaune) à l'exception des mots-clés issus de Twitter (en bleu) qui concernent une unité administrative. Chacune de ces données possède une représentation spatiale et temporelle qui permet de l'exploiter sur ces deux dimensions. Les données calculées à partir du BAI ne sont pas représentées par souci de lisibilité. Un extrait du graphe de connaissances au format RDF représentant les données issues de Twitter est disponible dans l'annexe A.4.

5.4.4.1 Exploitation avec SPARQL

Pour illustrer l'exploitation possible de cette base de connaissances, nous avons construit une requête SPARQL permettant de répondre aux questions suivantes :

— Est-ce qu'un incendie a eu lieu sur les images acquises ?

— Où est localisé l'évènement dans la zone couverte par ces images ?

— Quelle est l'ampleur de l'événement ?

— Quelles villes ont été impactées ?

— Quels sont les mot-clés pour ces villes à cette période ?

La requête présentée dans l'annexe A.5 se compose de 6 parties. La première partie a pour but de récupérer les données géospatiales du raster image sur lequel a lieu la recherche. La deuxième partie récupère l'ensemble des ROI localisées dans le raster. La troisième partie récupère l'ensemble des points chauds de chaque ROI. Nous avons choisi d'effectuer la fonction SPARQL `COUNT` sur ces points chauds pour ne garder que le nombre de points chauds par ROI et non la liste complète. La quatrième partie de la requête récupère les polygones représentant le BAI. Pour cette partie, nous avons fait le choix de n'afficher que la surface totale des polygones de BAI par ROI en mètres carrés. La cinquième partie de la requête affiche l'ensemble des unités administratives en intersection avec une ROI et la dernière partie affiche les mots-clés associés à ces unités administratives.

start	end	roi	nbFirePoint ROI	totalBurnt Area	adminUnit Name	keyword AdminUnit	keyword Frequency
"2019-10-22T21:44:32"	"2019-11-01T21:30:37"	http://melodi.irit.fr/lod/change /highChange_T10SEH_20191022_20191101_roi2234	2195	550752836	"Healdsburg"	"healdsburg"	15324
"2019-10-22T21:44:32"	"2019-11-01T21:30:37"	http://melodi.irit.fr/lod/change /highChange_T10SEH_20191022_20191101_roi2234	2195	550752836	"Healdsburg"	"evacuation"	8024
"2019-10-22T21:44:32"	"2019-11-01T21:30:37"	http://melodi.irit.fr/lod/change /highChange_T10SEH_20191022_20191101_roi2234	2195	550752836	"Healdsburg"	"windsor"	7207
"2019-10-22T21:44:32"	"2019-11-01T21:30:37"	http://melodi.irit.fr/lod/change /highChange_T10SEH_20191022_20191101_roi2234	2195	550752836	"Healdsburg"	"#kincadefire"	6892
"2019-10-22T21:44:32"	"2019-11-01T21:30:37"	http://melodi.irit.fr/lod/change /highChange_T10SEH_20191022_20191101_roi2234	2195	550752836	"Healdsburg"	"fire"	6228
"2019-10-22T21:44:32"	"2019-11-01T21:30:37"	http://melodi.irit.fr/lod/change /highChange_T10SEH_20191022_20191101_roi2234	2195	550752836	"Healdsburg"	"#healdsburg"	4779

FIGURE 5.17 – Extrait du résultat de la requête SPARQL sur la base de connaissances

La figure 5.17 montre un extrait du résultat de la requête. Seules les lignes traitant de la ROI ayant pour identifiant 2234 (colonne `roi`) sont affichées. On peut voir ici que pour la période choisie pour la détection de changement, cette ROI contient 2195 points chauds, et sa superficie est supérieure à 55000 ha. La ville de *Healdsburg* se trouve dans cette ROI et la liste des mots-clés identifiés pour cette période contient **#kincadefire**, **fire** et **evacuation**. Ce tableau montre qu'il est possible d'avoir plus de contexte sur le raster de changement afin qu'un expert puisse prendre des décisions.

5.4.4.2 Visualisation avec QGIS

QGIS [6] est un SIG libre et open source développé par la fondation OSGeo. Cet outil nous permet de visualiser les données géospatiales sur un fond de carte et sur les images satellitaires pour une validation visuelle. Il peut afficher des images disponibles dans un grand nombre de formats et permet la reprojection de coordonnées lorsque cela est nécessaire. Pour le cas d'utilisation relatif au *Kincade Fire*, nous avons représenté les unités administratives associées à leurs mot-clés dans un fichier shapefile. Nous avons développé un script qui interroge la base de données NoSQL CouchDB contenant les villes et les mot-clés et construit un fichier shapefile en ajoutant ces mot-clés en tant que label sur les polygones.

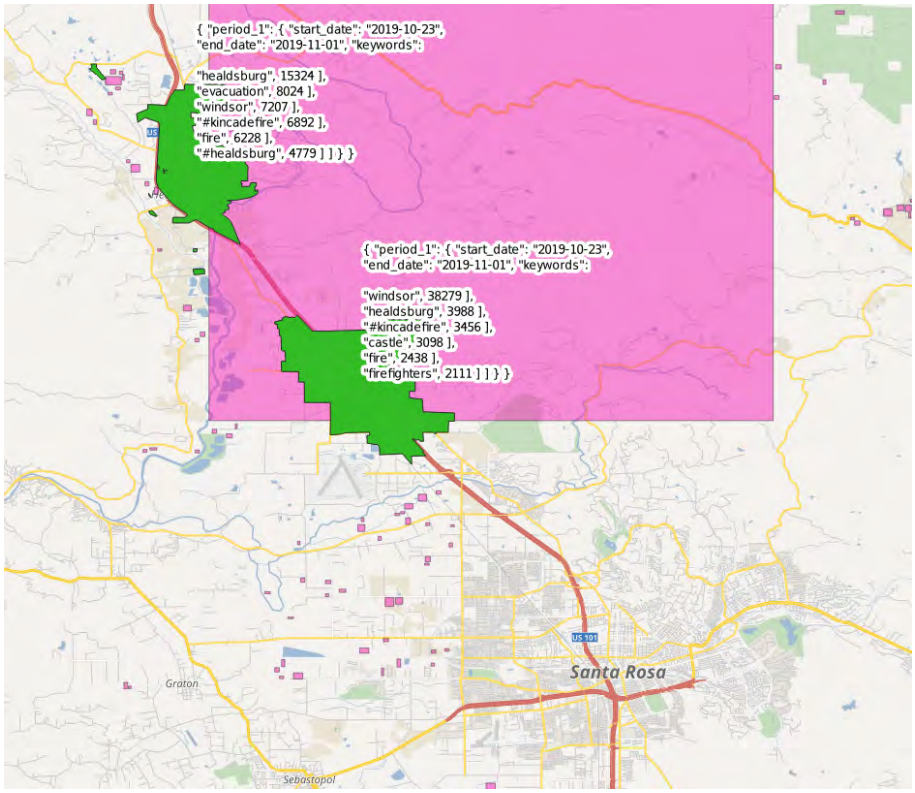


FIGURE 5.18 – Visualisation des mot-clés pour les villes et les ROI dans le SIG QGIS

Sur la figure 5.18, les unités administratives sont représentées en vert et les ROI issues du raster de changement sont en rose. Cette visualisation permet de situer les villes impactées par l'incendie et de prévoir les potentiels dégâts. Le fond de carte OSM permet d'observer l'occupation du sol.

6. `https://www.qgis.org/`

5.4.5 Autres cas d'étude

Afin de valider l'ensemble de la chaîne de sémantisation, nous avons voulu expérimenter celle-ci sur d'autres cas d'utilisation. Les scénarios choisis sont aussi des feux de forêts à l'exception de l'explosion au port de Beyrouth qui a eu lieu le 4 août 2020. Le but est de montrer que les processus et modèles définis peuvent fonctionner sur d'autres données et obtenir un résultat cohérent dans plusieurs scénarios.

5.4.5.1 Camp fire

Le *Camp Fire*[7] est un feu de forêt ayant touché la Californie du 8 au 26 novembre 2018 et brûlé plus de 62000 ha. Nous avons choisi ce cas d'étude pour comparer les résultats obtenus par [Ban *et al.* 2020]. Ces travaux montrent qu'il est possible d'étudier la superficie de ce feu de foret à l'aide d'images Synthetic-Aperture Radar (SAR) Sentinel-1. Ils utilisent un algorithme de détection de changement CNN pour créer un raster de changement. L'avantage majeur des images SAR est qu'elles ne dépendent pas de la couverture nuageuse mais elles ont une moins bonne résolution que les images optiques. Une des difficultés que nous avons rencontrées a été de trouver une image du satellite Sentinel-2 avant l'événement et une image après pour pouvoir effectuer une détection de changement.

Les deux images satellitaires que nous avons utilisées datent respectivement du 22 octobre 2018 et du 31 décembre 2018. La période entre les deux images est de 9 semaines ce qui est assez important. Plus la période est grande entre les images et plus l'algorithme de détection de changement non supervisé est susceptible de détecter des changements de végétations liés aux saisons. Dans notre cas, l'algorithme a bien détecté les changements causés par l'incendie et les ROIs identifiées correspondent au polygone de l'incendie. Nous avons dû ajuster le paramètre `threshold` à une valeur de 0.75 car la valeur de 0.65 ne formait qu'une seule ROI couvrant à la fois l'incendie et une zone non touchée par l'incendie.

Pour les données contextuelles, nous n'avons pas eu de problème particulier pour la récupération des points chauds et le calcul du BAI. Nous avons bien identifié dans YAGO2Gep les villes se trouvant dans l'emprise de l'image. L'étape la plus complexe a été la récupération des mots-clés à partir des tweets. En effet, la ville ayant été impactée par ce feu de foret est la ville de *Paradise*. Notre algorithme ne traitant pas l'ambiguïté sur les noms, les mot-clés obtenus pour la période ne correspondaient pas au champ lexical d'un feu de forêt. Une des solutions était d'utiliser la localisation des tweets pour ne garder que ceux se trouvant dans l'emprise de l'image mais il n'y en avait pas dans le jeu de données Twitter utilisé. Nous avons donc désambiguïsé ce nom de ville en y ajoutant le suffixe *California*, ce qui s'est avéré efficace.

La figure 5.19 montre les résultats obtenus après tous les traitements. La ville de `Paradise` est représentée par le polygone vert et les ROI sont en rose. Les mot-clés retournés par l'algorithme contiennent **fire** et **wildfire** qui font partie du champ lexical d'un feu de forêt. Ce cas d'étude nous a permis de valider l'ensemble de la

7. `https://www.fire.ca.gov/incidents/2018/11/8/camp-fire/`

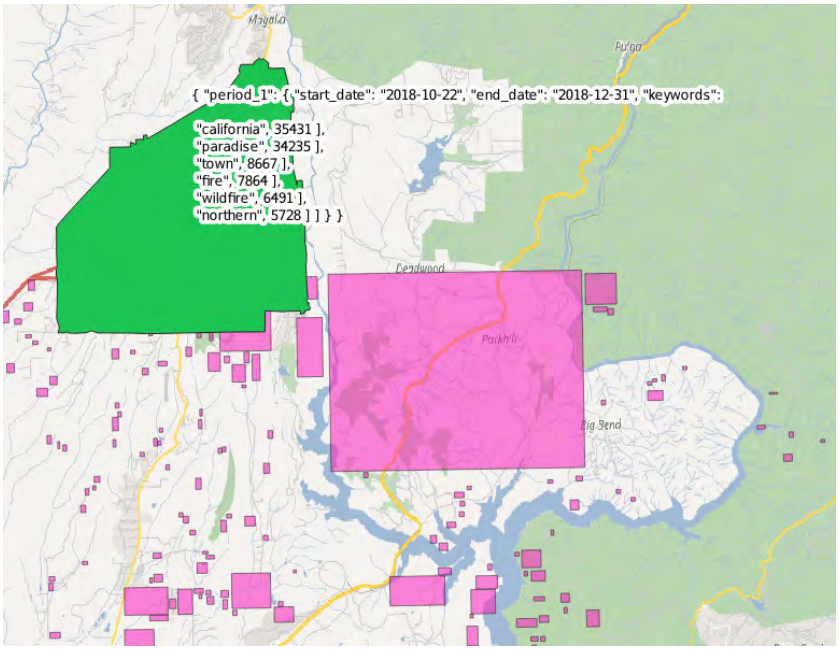


FIGURE 5.19 – Visualisation des mot-clés pour la ville Paradise et les ROI dans le SIG QGIS

chaîne mais aussi de voir les faiblesses au niveau de la détection de mots-clés dans les Tweets. Nous avons pu aussi constater que le paramètre `threshold` utilisé pour la génération de ROI doit être ajusté en fonction du contexte.

5.4.5.2 Turquie

Cette étude porte sur un feu de forêt ayant eu lieu dans le sud-ouest de la Turquie du 6 au 7 septembre 2017 [8]. Cet incendie a eu lieu dans une zone peu urbaine et a brûlé plus de 400 ha de forêt et terrains agricoles. Le choix de ce cas d'étude a été fait pour pouvoir comparer nos résultats avec ceux obtenus par [Kurnaz *et al.* 2020]. Dans leurs travaux, l'incendie est identifié de différentes manières. La première consiste à utiliser plusieurs calculs d'indices comme le NDVI sur une image avant et une image après l'incendie et d'effectuer une différence d'indices pour obtenir une valeur delta appelée dNDVI.

Pour cette étude, nous avons utilisé deux images Sentinel-2 datant du 28 août 2017 et du 7 septembre 2017. L'algorithme de détection de changement non supervisé suivi de l'algorithme de création de ROI de changements ont correctement identifié la zone impactée par l'incendie. Nous avons aussi correctement généré les ROI de BAI. Concernant les données contextuelles, nous avons pu récupérer l'ensemble des points chauds pour la période. Nous avons cependant rencontré des difficultés lors de la recherche des unités administratives ainsi que les mots-clés ve-

8. `https://crisis24.garda.com/insights-intelligence/intelligence/risk-alerts/sge9j2tz4zdvkwcjl/turkey-ongoing-forest-fire-in-mugal-province-july-10`

nant des tweets. En effet, le feu de forêt se trouvant dans une zone peu urbaine, nous n'avons pas trouvé d'unité administrative en intersection avec les ROI de changement générées. Pour résoudre ce problème, nous avons décidé de rechercher les unités administrative au niveau des régions. Pour la recherche de mots-clés à partir de Twitter, l'incendie se situant dans une zone peu urbaine, nous n'avons pas réussi à collecter assez de tweets en utilisant le nom de le région pour pouvoir en extraire des mots-clés.

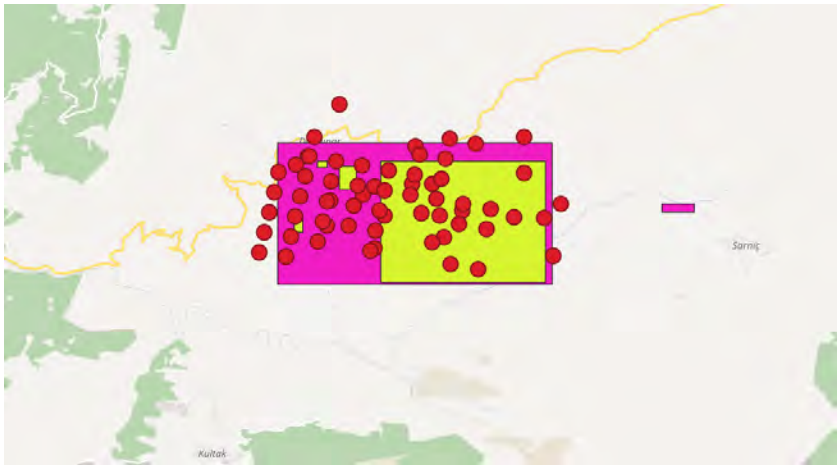


FIGURE 5.20 – Visualisation des données pour l'incendie en Turquie dans le SIG QGIS

La figure 5.20 montre les données que nous avons collectées pour ce cas d'utilisation. On remarque que les points chauds en rouge, les ROI du BAI en jaune ainsi que les ROI de changements en rose, se concentrent à l'endroit même où a eu lieu l'incendie. Ce cas d'utilisation nous a permis de valider notre chaîne de traitement sur un autre feu de forêt de plus petite taille. Nous avons pu également constater que les données contextuelles issues de Twitter ne peuvent pas être identifiées lorsque l'événement à étudier est localisé dans une zone peu habitée.

5.4.5.3 Incendies en Australie

Pour cette étude, nous avons choisi d'appliquer l'ensemble de notre chaîne de traitement aux incendies ayant eu lieu dans l'est de l'Australie entre septembre 2019 et mars 2020[9]. Ces feux sont principalement des feux de brousse localisés dans de multiples endroits isolés. Ces incendies ont détruit près de 12 millions d'hectares de brousses et de forets. La principale difficulté de cette étude réside dans le fait qu'il ne s'agit pas d'un seul événement ponctuel mais d'un ensemble d'évènements se déroulant sur une période longue de plusieurs mois.

Nous avons récupéré des images satellitaires Sentinel-2 de la zone en date du 3 octobre 2019 et du 27 décembre 2019. Cette zone couvre les forets et la brousse

9. `https://www.wwf.org.au/what-we-do/bushfires`

sur la côte est jusqu'à la ville de *Port Macquire*. La détection de changement avec l'algorithme non supervisé n'a pas posé de difficulté particulière tout comme la génération des ROI avec une valeur de `threshold` à 0.65. Les polygones issus du BAI ont aussi été générés sans difficulté. La première difficulté rencontrée lors de cette étude a été le nombre de points chauds obtenus depuis le site . En effet, nous avons obtenu 20 000 points chauds entre les deux dates des images satellitaires. La génération du graphe de connaissances s'est bien déroulée et tous les points sont présents. Cependant, l'interrogation avec SPARQL de la base de connaissances, stockée dans l'entrepôt Virtuoso, a été un échec : la requête permettant de calculer le nombre de points chauds contenus dans une ROI n'a pas fourni de résultat, la machine hébergeant l'entrepôt n'étant pas assez performante.

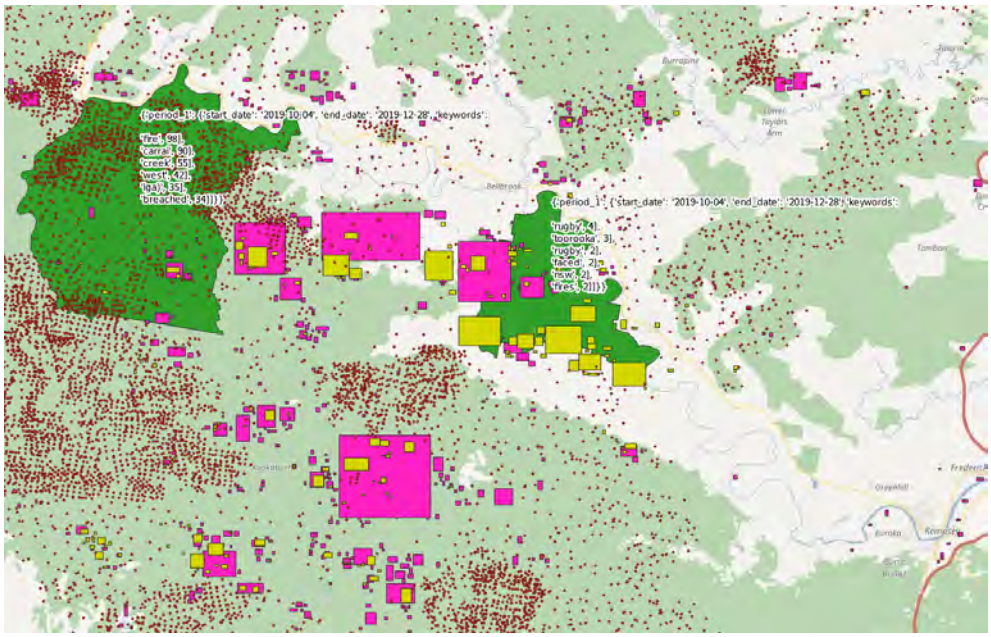


FIGURE 5.21 – Visualisation des données pour les incendies en Australie dans le SIG QGIS

La figure 5.21 présente une partie des données collectées pour la zone correspondant aux images satellitaires. On remarque ici que les ROI de changements en rose et de BAI en jaune sont dispersées tout comme les points chauds en rouge. Nous avons identifié deux unités administratives en vert se trouvant dans la zone étudiée : la ville de *Carrai* ainsi que la ville de *Toorooka*. L'algorithme de recherche de mots-clés a bien identifié les mots **fire** et **fires** pour la période.

5.4.5.4 Beyrouth

Ce dernier cas d'étude a pour but de valider notre approche sur un autre événement qu'un feu de forêt. Nous avons choisi d'étudier l'explosion ayant eu lieu sur le

port de Beyrouth le 4 août 2020 [10]. Nous avons essayé d'effectuer, dans un premier temps, une détection de changement avec des images issues du satellite Sentinel-2 mais la zone impactée par l'explosion ne représentait que quelques pixels sur l'image. Dans un second temps, nous avons utilisé les images fournies par l'entreprise Maxar sous forme de données ouvertes [11]. Ces images proviennent du satellite WorldView-2 et ont une résolution spatiale de 50 cm. L'image avant l'explosion date du 31 juillet 2020 et celle après l'explosion date du 5 août 2020.

L'algorithme de détection de changement non-supervisé a fonctionné correctement et nous avons pu identifier les ROI de changements localisées sur le port de Beyrouth. Pour ce cas d'étude, nous n'avons pas d'autre donnée contextuelle que les noms des unités administratives ainsi que les mots-clés associés à celles-ci.

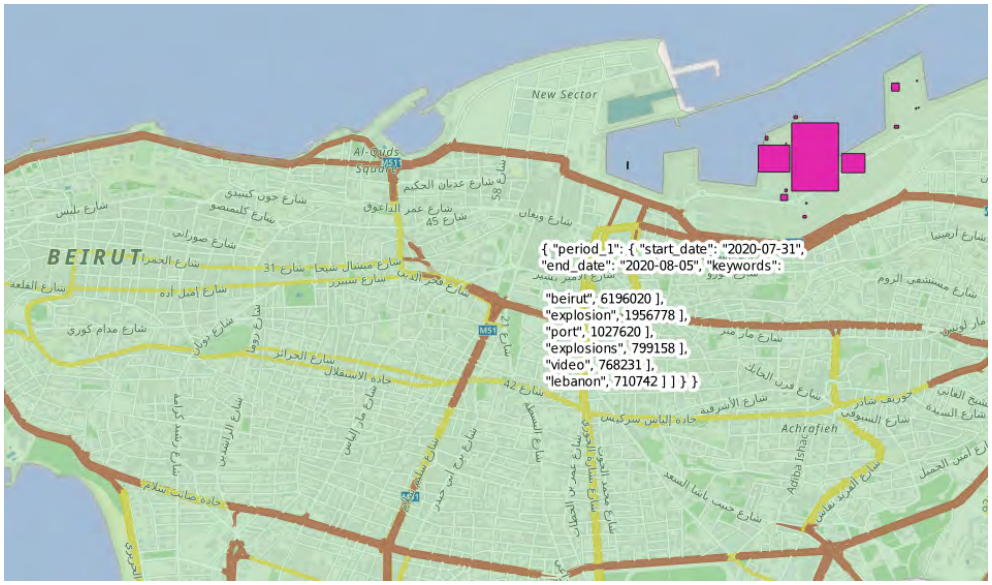


FIGURE 5.22 – Visualisation des données pour l'explosion de Beyrouth dans le SIG QGIS

Sur la figure 5.22, on peut observer les ROI générées par le raster de changement en rose. Ces ROI couvrent l'épicentre de l'explosion dans le port. En effet, beaucoup de bâtiments ont été affectés sans que leur aspect vu du ciel soit suffisamment différent ou alors de manière trop dispersée, ce qui a pour conséquence que l'algorithme de comparaison d'images n'a pas identifié d'autre zone de changement. Nous pouvons également voir le résultat obtenu pour la recherche de mots-clés pour la ville de Beyrouth pour cette période. Les mots **port** et **explosion** apparaissent dans cette liste de mots-clés. Cette étude a permis de montrer que la chaîne de traitement développée fonctionne sur un autre type d'évènement que les incendies.

10. `https://www.bbc.com/news/world-middle-east-53668493`
11. `https://www.maxar.com/open-data/beirut-explosion`

5.5 Conclusions

Dans ce chapitre, nous avons illustré l'utilisation de la chaîne de sémantisation complète au travers de différents cas pratiques, en particulier la chaîne la plus complète définie pour les rasters de changements et en utilisant des zones d'intérêt (ROI). Certains scénarios nous ont permis de tester les limites de notre approche concernant le choix des données ou le paramétrage des variables. Dans l'ensemble des cas, nous avons obtenu un graphe de connaissances qui permet d'ajouter du contexte au raster obtenu grâce à la détection de changements non supervisée ou par analyse d'une image satellitaire.

Les résultats obtenus dans les différentes approches de modélisation spatiale montrent que le concept de ROI est le plus adapté pour l'exploitation des rasters et images satellitaires. L'avantage majeur de cette manière d'agréger les pixels est que l'exploitation du graphe de connaissances est bien plus rapide que l'exploitation de rasters. De plus, elle ne nécessite pas de connaissance poussée dans le domaine de la télédétection. L'apport de cette chaîne de traitement est la facilité d'exploitation de données géospatiales complexes. Les phénomènes sur les images et les rasters sont ici représentés grâce au concept de ROI qui évite de traiter l'image dans sa globalité.

Chapitre 6

Conclusion et perspectives

Cette thèse a abordé le problème de l'utilisation de graphes de connaissances pour donner du sens aux changements détectés à partir d'images d'observation de la Terre, grâce à des données contextuelles calculées à partir des images ou disponibles sous forme de données ouvertes. Nos principales contributions sont les suivantes :

— un algorithme pour regrouper de manière optimale les pixels pour définir les régions d'intérêt (ROI) et trouver la division géographique précise qui guidera l'intégration des données ;

— un processus d'intégration sémantique de données qui prend en entrée un raster et qui génère un graphe de connaissances à partir de différentes sources de données contextuelles afin d'aider à identifier ou à qualifier les changements ; ce processus peut être adapté grâce à la définition de templates pour prendre en compte de nouveaux types de données ;

— plusieurs ontologies modulaires permettant la représentation sémantique des données, et une approche pour adapter ces ontologies en fonction des données contextuelles à prendre en compte ;

— une validation de l'approche avec différents cas d'étude (changement du NDVI, de l'occupation du sol, et le suivi d'incendies).

Les travaux de cette thèse s'inscrivent dans la convention Cifre ANR n° 2017/1399 entre Thales Alenia Space et le CNRS.

Dans ce qui suit, nous dressons un bilan de ces contributions, suivi des différentes perspectives envisagées dans la poursuite des travaux réalisés dans le cadre de cette thèse.

6.1 Bilan des contributions

Agrégation de pixels de changement Les algorithmes d'apprentissage automatique dédiés à la détection de changement sont capables d'identifier un degré de changement au niveau du pixel à partir de la comparaison de deux images. Comme les algorithmes d'analyse d'image qui calculent des indices, ils fournissent leurs résultats sous forme de fichier raster géo-référéncés, datés, qui gardent un lien vers la ou les images satellitaires à partir desquels ils sont calculés. Dans ces fichiers raster, l'information sur le changement ou l'indice est une valeur associée à chacun des pixels. Pour donner du sens à ces fichiers, les pixels doivent être regroupés en zones impactées en fonction de leur valeur. Ces regroupements présentent l'avantage de

faciliter l'interprétation des fichiers rasters, et de réduire les temps de traitement pour exploiter ces données au niveau d'une zone et non pixel par pixel.

Dans cette thèse, nous avons expérimenté plusieurs manières de regrouper les pixels : en fonction de régions prédéfinies (tuiles, polygones d'entités géographiques ou administratives, etc) ou calculées en fonction des valeurs des pixels et de leur voisinage. Nous avons proposé la notion de *région d'intérêt* (ROI) pour générer à partir des valeurs de changement, des zones suffisamment étendues, aux contours simples, dans lesquelles le changement ou l'indice étudié à une valeur homogène et "significative".

Une ROI est définie comme un sous-ensemble d'un fichier raster constitué selon deux paramètres : i) la majorité des pixels de ce sous-ensemble a une valeur supérieure à un seuil minimum (ce seuil définit un changement significatif) ; et ii) le polygone constitué à partir des pixels possède une taille minimale. Pour accélérer le calcul et la recherche de ROI, ceux-ci sont représentés par des rectangles correspondant à leur enveloppe. Plus précisément, une ROI est *l'enveloppe rectangulaire d'une taille minimum qui regroupe les pixels d'un fichier raster avec une valeur supérieure à un seuil*. Chaque région d'intérêt est délimitée par quatre points plutôt que leur périmètre exact de forme complexe.

La notion de ROI joue un rôle majeur dans le processus d'intégration de données afin de trouver la division géographique précise qui guidera l'intégration des données. L'identification des ROI à partir d'un raster permet notamment de pouvoir ne générer de la connaissance que sur une zone géographique d'intérêt. Contrairement aux tuiles, toutes les ROI n'ont pas la même taille. Au centre de notre modèle, l'utilisation des ROI permettent une plus grande flexibilité dans la représentation de données géospatiales de part leur forme rectangulaire, facile à traiter et dont la recherche ou l'utilisation réduit les temps de calcul. Le fait de ne représenter sémantiquement que les parties importantes d'un raster, ayant un intérêt scientifique ou pour l'étude de certains phénomènes, par le biais de ces ROI, est une approche innovante pour la représentation sémantique de fichier rasters. Les approches qui visent à représenter l'ensemble des pixels d'un raster sous forme d'entités ou de concepts, atteignent leurs limites lorsqu'il s'agit de représenter des fichiers raster composés de plusieurs millions de pixels. L'approche basée sur des ROI proposée dans cette thèse nécessite un traitement, en amont de la sémantisation, pour regrouper les pixels proches géographiquement et ayant des valeurs proches de raster en polygones. Ce processus est peu coûteux en terme de performance mais il permet d'améliorer grandement l'interrogation des données via SPARQL puisque le nombre de polygones est réduit par rapport au nombre de pixels du raster.

Changement par tuile et polygone D'autres types de découpages spatiaux ont été également explorés dans cette thèse dans le cadre de l'étude de changement. Avec l'étude du NDVI, les indices de végétation sont calculés sur les tuiles Sentinel-2 à partir des images correspondantes. L'ontologie permet d'établir des relations spatiales et temporelles facilitant ainsi l'analyse de l'évolution des indices.

Ce processus peut être appliqué à tous les indices calculables à partir des images satellitaires comme le NDSI ou NDWI. Ceux-ci permettent de découvrir les types de sols présents dans l'image sans avoir à les visualiser. La représentation sémantique de ces indices permet de les lier avec des données disponibles dans le LOD.

Une des perspectives de ce travail consiste à étudier un changement anormal directement à partir des données sémantisées. Une solution envisagée est de fournir des règles de raisonnement ainsi qu'un seuil permettant de détecter automatiquement un changement anormal via les variations de valeurs de ces indices. Ces valeurs peuvent également servir de données contextuelles pour documenter un changement détecté au préalable par un autre processus.

Obtenir le land cover sur une zone choisie a un fort intérêt pour les applications de télédétection. L'API que nous avons développée permet d'obtenir ces valeurs sous forme de graphe de connaissances via une seule requête. Le but de cette API est d'être exploitée en tant que maillon d'une chaîne de traitement. Cela permet d'ajouter de la connaissance sur un lieu particulier quelle que soit sa localisation. Le graphe de connaissances est généré grâce à une ontologie qui reprend l'ensemble des classes du land cover comme concepts. Grâce à ces données il est possible d'établir des tendances pour une zone en choisissant le même land cover sur différentes années. Il est également possible de définir des règles de raisonnement sur les données afin de qualifier un changement de couverture du sol à la manière de [Li *et al.* 2016].

Une des améliorations possible pour ce travail est d'aligner les concepts entre les différents land cover pour pouvoir faire une comparaison de valeurs entre eux. Le but est de pouvoir utiliser une seule ontologie pour l'ensemble des land cover existants dans le système. Une seconde amélioration possible pour ce travail est la possibilité d'ajouter un nouveau land cover directement depuis une interface ou une requête. L'utilisateur fournirait alors le raster du land cover et remplirait un fichier template dans lequel les noms des classes sont associés aux valeurs des pixels pour qu'il soit intégré dans l'API.

Processus d'intégration sémantique Une partie de la démarche proposée consiste à ajouter des données contextuelles après avoir identifié un changement. Grâce à ces données, il est possible de confirmer qu'un changement détecté de manière automatique est dû à un évènement spécifique. Nous avons traité ce problème pour des images issues de divers satellites et des changements détectés grâce à un algorithme non-supervisé utilisant le deep learning. Nous avons défini un processus générique qui prend en entrée un raster de changement géoreferencé et daté, plusieurs sources de données géospatiales ouvertes comme information contextuelle et génère un graphe de connaissances.

L'un des apports majeurs des travaux de cette thèse est le processus de sémantisation appliqué aux fichier raster. L'ensemble des composants de la chaîne a pour but d'ajouter de la connaissance aux images satellitaires ainsi qu'au résultat de leurs exploitations. Le fait d'utiliser des ontologies standards pour la représentation des dimensions spatiale et temporelle permet d'étendre les exploitations possibles de ces

données avec d'autres réutilisant ces mêmes standards. L'ontologie GeoSPARQL est une référence de plus en plus utilisée pour la représentation de propriétés spatiales, qui continue d'évoluer. Concernant la dimension temporelle, l'ontologie OWL-Time permet la représentation de concepts simples ou complexes.

La chaîne de sémantisation proposée permet de collecter un maximum d'informations de manière automatique afin que l'intervention humaine soit le moins nécessaire possible pour identifier un évènement à partir d'images satellites. La structuration de ces données en graphe de connaissances permet également une exploitation facilitée au sein d'un catalogue d'image où les informations contextuelles font partie de l'image au même titre que les méta-données. Il est également possible de considérer l'image en tant que concept faisant partie du LOD et d'exploiter l'ensemble des connaissances disponibles relatives à cette image comme des données statistiques ouvertes ou des informations journalistiques.

Aujourd'hui la majorité des catalogues d'images satellitaires disponibles ne permettent que des recherches d'images via les dimensions spatiale et temporelle. Le Web sémantique permet d'ajouter des connaissances à ces catalogues en proposant d'autres dimensions comme la recherche d'image par évènement et de pouvoir trouver des informations relatives à cet événement directement dans le catalogue. L'ajout de données provenant de réseaux sociaux en tant que données contextuelles du catalogue est aussi intéressante car cela permet d'intégrer les retours de personnes se trouvant parfois sur le terrain lors d'un événement.

6.2 Perspectives

Nous évoquons ici les perspectives envisagées pour la poursuite de ces travaux.

Identifier différents types d'évènements La première perspective est de pouvoir appliquer l'ensemble de ce processus sur d'autres types d'évènements observables via des images satellitaires comme les inondations, les séismes, les éruptions volcaniques ou encore les flux de migrations de populations. Pour ce faire, il faudrait identifier dans un premier temps une source de donnée ouverte sur l'évènement à intégrer. Le but est de trouver des informations datées et géo-localisées qui pourront être utilisées comme données contextuelles. Après l'identification des données, il faut dans un second temps, établir des template pour pouvoir les sémantiser. L'expérience menée avec l'explosion survenue à Beyrouth a montré que la chaîne de traitement est facile à adapter, mais que les images satellitaires ne constituent pas toujours une source adaptée pour observer l'impact de certains événements, comme une explosion dans ce cas.

Améliorer l'algorithme de recherche de mots-clés Une autre piste a été amorcée pour améliorer la recherche de mots-clés au sein de message du réseau social Twitter grâce à un algorithme plus performant de type Tf.Idf comme réalisé par [Marujo *et al.* 2015]. Afin de valider les mots trouvés, une deuxième piste serait

aussi d'exploiter un vocabulaire de catastrophes naturelles comme l'ontologie proposée par [Bouyerbou *et al.* 2019]. Une troisième perspective pour l'amélioration de la recherche de mots-clés serait d'utiliser les tweets collectés à l'aide du scrapper Twint [1]. Cet outil permettrait de ne plus utiliser les données Twitter collectées par la plateforme OSIRIM et d'obtenir une quantité plus importante de tweets. Enfin, une dernière piste envisagée serait d'analyser les sentiments exprimés dans les tweets pour pouvoir les classer en 3 catégories : négatifs, neutres et positifs en utilisant les travaux de [Nielsen 2011]. Ces classes peuvent donner un indice supplémentaire sur la nature de l'évènement qui est en train de se dérouler dans l'image.

Améliorer les algorithmes de détection d'objet par apprentissage automatique L'approche que nous proposons débouche sur la disponibilité d'un graphe de connaissances décrivant le contenu d'images satellitaires. Ce graphe étant disponible, une perspectives de ces travaux serait de l'exploiter en retour pour mieux analyser les images à l'aide d'algorithmes d'apprentissage automatique. En effet, les connaissances géolocalisées représentées dans les graphes RDF peuvent être considérées comme des annotations de zones géographiques dans les images. Ces annotations peuvent servir d'exemples pour entraîner des algorithmes d'apprentissage supervisé, comme celui proposé par [Redmon *et al.* 2016], à détecter des objets ou des phénomènes, et ainsi améliorer les résultats de ces algorithmes. D'une certaine manière, les graphes de connaissances fournissent à ces algorithmes des labels pour les pixels des zones ainsi caractérisées, donc des jeux de données étiquetées, soit par des entités sémantiques, soit par des mots-clés extraits des réseaux sociaux. Dans cette perspective, l'objectif serait de mettre en oeuvre la démarche proposée dans la thèse pour analyser des images satellitaires en vue de repérer un phénomène particulier, d'identifier les ROI dans lesquelles ce phénomène est présent. Les ROI identifiées servent à découper des *patches* dans les images satellitaires pour les fournir aux algorithmes d'apprentissages comme jeu de données d'entraînement. Ces patches sont des vignettes étiquetées permettent d'associer du texte, ici le nom d'un phénomène, à un groupe de pixels. L'objectif est de pouvoir identifier automatiquement ce même phénomène sur de nouvelles images sans passer par une détection de changement par comparaison entre deux images.

Proposer un indice de confiance sur les données contextuelles Une dernière piste envisagée suite aux travaux réalisés pendant cette thèse est d'établir un indice de confiance sur l'évènement identifié à partir du raster de changement. Cette indice serait calculé en fonction du nombre de données contextuelles qui viennent confirmer la nature de cet événement. Plus il y a de données qui indiquent la même information, plus l'indice serait élevé. Pour le cas d'un incendie par exemple, nous avons disposé de trois sources de données : la base de données Firecast, Twitter et le calcul du BAI. Il serait également possible de combiner d'autres indices calculés comme le NDWI ou le NDSI pour définir la confiance d'un phénomène lié à ces in-

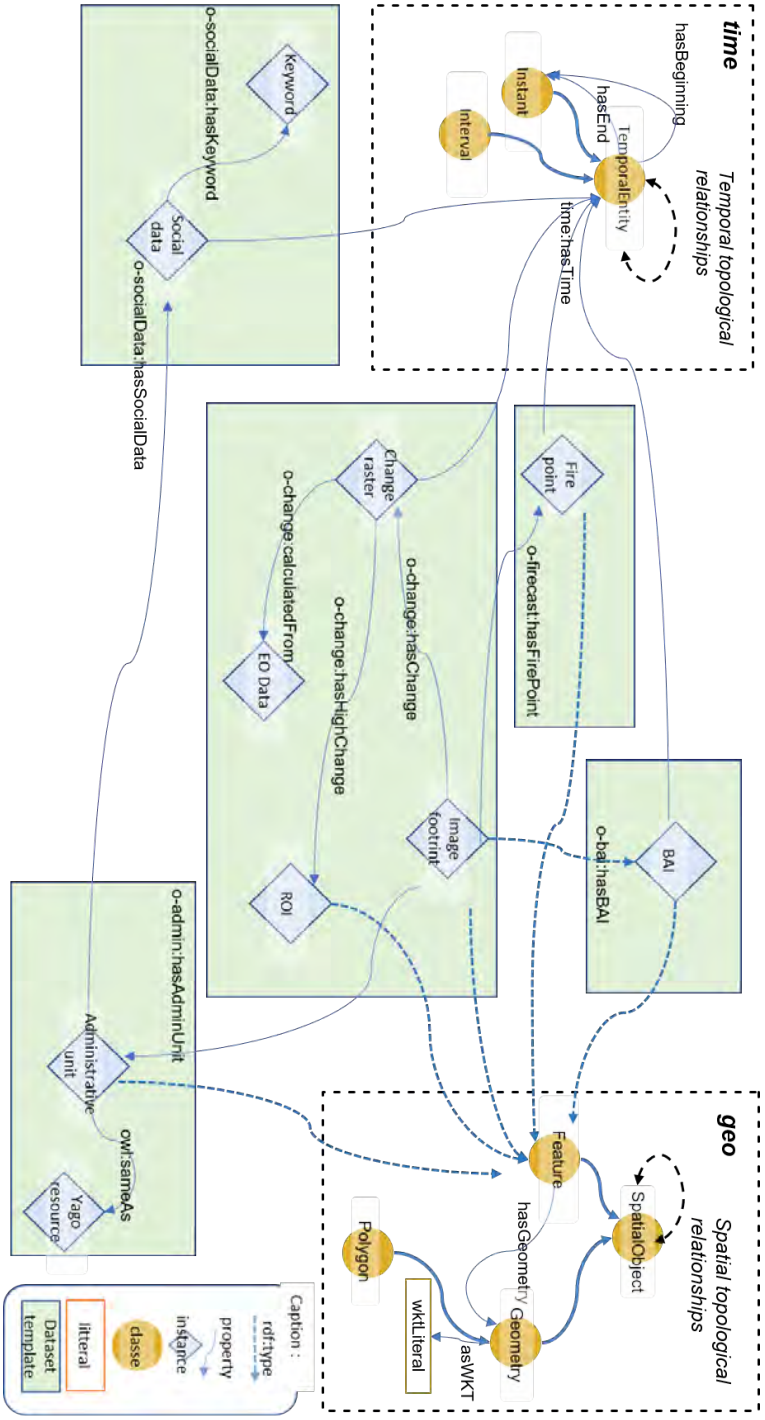
1. `https://github.com/twintproject/twint`

dicateurs. Certaines sources peuvent avoir une fiabilité plus ou moins élevée ce qui jouerait un rôle dans la valeur finale de cet indicateur. Cet indicateur permettrait d'éliminer les résultats faux-positifs dans le cas d'une automatisation complète du processus de sémantisation.

Annexe A

Annexes

A.1 Représentation graphique de la modélisation des rasters de changements associées aux données contextuelles



A.2 Requête SPARQL permettant l'étude de de l'évolution du land cover sur la ville de Blagnac

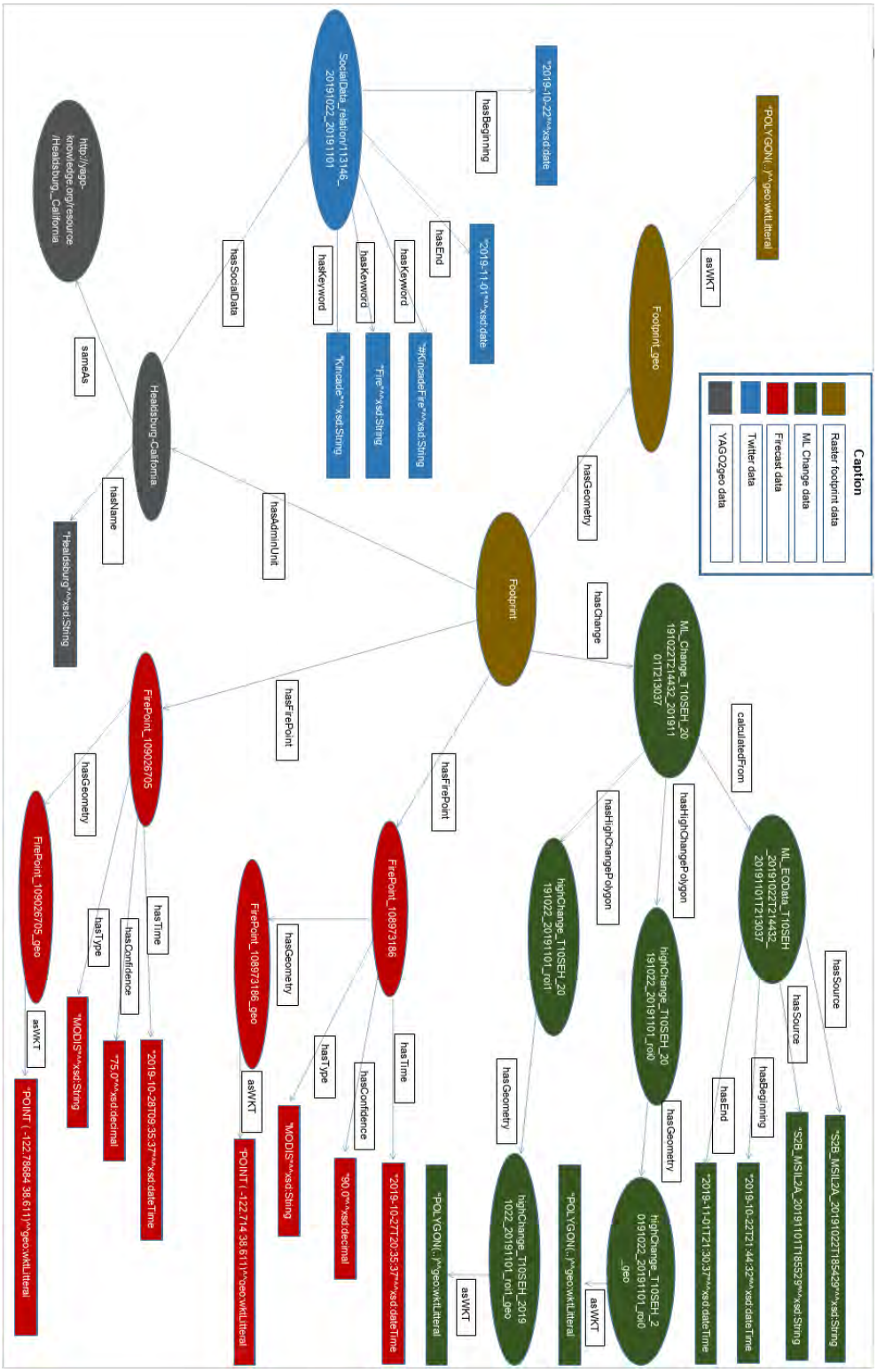
```
prefix time: <http://www.w3.org/2006/time#>
prefix lci: <http://melodi.irit.fr/ontologies/lci.owl#>

SELECT ?landcoverStart ?landcoverEnd SUM(?vegetation) AS ?totalVegetationCESBIO
    SUM(?urbain) AS ?totalUrbainCESBIO
WHERE{
    ?lc a lci:LandCoverDataset .
    ?lc time:hasTime/time:hasBeginning/time:inXSDDateTime ?landcoverStart .
    ?lc time:hasTime/time:hasEnd/time:inXSDDateTime ?landcoverEnd .
    ?lc lci:hasData ?lcClass.

    #VEGETATION
    {?lcClass a lci:CESBIO-CultureEte;lci:hasLandCoverPercentage ?vegetation.}UNION
    {?lcClass a lci:CESBIO-CultureHiver;lci:hasLandCoverPercentage ?vegetation.} UNION
    {?lcClass a lci:CESBIO-ForetConiferes;lci:hasLandCoverPercentage ?vegetation.} UNION
    {?lcClass a lci:CESBIO-ForetFeuilleus;lci:hasLandCoverPercentage ?vegetation.} UNION
    {?lcClass a lci:CESBIO-LandesLigneuse;lci:hasLandCoverPercentage ?vegetation.} UNION
    {?lcClass a lci:CESBIO-Pelouses;lci:hasLandCoverPercentage ?vegetation.} UNION
    {?lcClass a lci:CESBIO-Prairies;lci:hasLandCoverPercentage ?vegetation.} UNION
    {?lcClass a lci:CESBIO-Vergers;lci:hasLandCoverPercentage ?vegetation.} UNION
    {?lcClass a lci:CESBIO-Vignes;lci:hasLandCoverPercentage ?vegetation.} UNION

    #URBAIN
    {?lcClass a lci:CESBIO-UrbainDense;lci:hasLandCoverPercentage ?urbain.} UNION
    {?lcClass a lci:CESBIO-UrbainDiffus;lci:hasLandCoverPercentage ?urbain.} UNION
    {?lcClass a lci:CESBIO-ZonesIndEtCom;lci:hasLandCoverPercentage ?urbain.} UNION
    {?lcClass a lci:CESBIO-SurfacesRoute;lci:hasLandCoverPercentage ?urbain.}
}
ORDER BY ?landcoverStart
```

A.3 Extrait des graphes RDF générés



A.4 Extrait du graphe de connaissance représentant des données issues de Twitter

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix time: <http://www.w3.org/2006/time#> .
@prefix o-sd: <http://melodi.irit.fr/ontologies/socialData.owl#> .
@prefix g-sd: <http://melodi.irit.fr/lod/socialData/> .
g-admin:Healdsburg-California o-sd:hasSocialData g-sd:sd_Healdsburg-↩
    California_20191023_20191101 .
g-sd:Instant_20191023 a time:Instant .
g-sd:Instant_20191023 time:inXSDDateTime "2019-10-23"^^xsd:date .
g-sd:sd_Healdsburg-California_20191023_20191101 time:hasBeginning g-↩
    sd:Instant_20191023 .
g-sd:Instant_20191101 a time:Instant .
g-sd:Instant_20191101 time:inXSDDateTime "2019-11-01"^^xsd:date .
g-sd:sd_Healdsburg-California_20191023_20191101 time:hasEnd g-↩
    sd:Instant_20191101 .
g-sd:Keyword_healdsburg o-sd:hasValue "healdsburg"^^xsd:String .
g-sd:Keyword_healdsburg o-sd:hasFrequency "15324"^^xsd:decimal .
g-sd:sd_Healdsburg-California_20191023_20191101 o-sd:hasKeyword g-↩
    sd:Keyword_healdsburg .
g-sd:Keyword_evacuation o-sd:hasValue "evacuation"^^xsd:String .
g-sd:Keyword_evacuation o-sd:hasFrequency "8024"^^xsd:decimal .
g-sd:sd_Healdsburg-California_20191023_20191101 o-sd:hasKeyword g-↩
    sd:Keyword_evacuation .
g-sd:Keyword_windsor o-sd:hasValue "windsor"^^xsd:String .
g-sd:Keyword_windsor o-sd:hasFrequency "7207"^^xsd:decimal .
g-sd:sd_Healdsburg-California_20191023_20191101 o-sd:hasKeyword g-↩
    sd:Keyword_windsor .
g-sd:Keyword_kincadefire o-sd:hasValue "#kincadefire"^^xsd:String .
g-sd:Keyword_kincadefire o-sd:hasFrequency "6892"^^xsd:decimal .
g-sd:sd_Healdsburg-California_20191023_20191101 o-sd:hasKeyword g-↩
    sd:Keyword_kincadefire .
g-sd:Keyword_fire o-sd:hasValue "fire"^^xsd:String .
g-sd:Keyword_fire o-sd:hasFrequency "6228"^^xsd:decimal .
g-sd:sd_Healdsburg-California_20191023_20191101 o-sd:hasKeyword g-↩
    sd:Keyword_fire .
g-sd:Keyword_healdsburg o-sd:hasValue "#healdsburg"^^xsd:String .
g-sd:Keyword_healdsburg o-sd:hasFrequency "4779"^^xsd:decimal .
g-sd:sd_Healdsburg-California_20191023_20191101 o-sd:hasKeyword g-↩
    sd:Keyword_healdsburg .
```

A.5 Requête SPARQL permettant l'interrogation de l'ensemble des graphes de connaissances générés

```
SELECT ?start ?end ?roi  (count(distinct ?firepoint) as ?nbFirePointROI)
(sum(distinct ?BaiPolyArea) as ?totalBurntArea) ?adminUnitName
?keywordAdminUnit ?keywordFrequency

WHERE{

   #Step 1 - Change raster data
   ?MLchange o-change:calculatedFrom ?ML_EO_Data .
   ?ML_EO_Data o-change:hasSource ?image .
   ?ML_EO_Data time:hasBeginning/time:inXSDDateTimeStamp ?start .
   ?ML_EO_Data time:hasEnd/time:inXSDDateTimeStamp ?end .
   ?MLchange ^o-change:hasChange ?footprint .

   #Step 2 - ROI list
   ?MLchange o-change:hasHighChangePolygon ?roi .
   ?roi geo:hasGeometry/geo:asWKT ?roiWkt .

   #Step 3a - Firepoints
   ?footprint o-firecast:hasFirePoint ?firepoint .
   ?firepoint geo:hasGeometry/geo:asWKT ?firepointWKT .
   ?firepoint time:hasTime/time:inXSDDateTimeStamp ?firepointTime .
   FILTER(bif:st_intersects (?firepointWKT, ?roiWkt)) .
   FILTER(?firepointTime >= ?start && ?firepointTime <= ?end) .

   #Step 3b - BAI
   ?footprint o-bai:hasBAI ?BAI .
   ?BAI o-bai:hasBaiPolygon ?BaiPoly .
   ?BAI o-bai:calculatedFrom ?BAI_EO_Data .
   ?BAI_EO_Data time:hasTime/time:inXSDDateTimeStamp ?BAI_time .
   ?BaiPoly o-bai:hasArea ?BaiPolyArea .
   ?BaiPoly geo:hasGeometry/geo:asWKT ?BaiPolyWkt .
   FILTER(bif:st_intersects (?BaiPolyWkt, ?roiWkt)) .
   FILTER(?BAI_time >= ?start && ?BAI_time <= ?end) .

   #Step 4 - Administrative units
   ?footprint o-admin:hasAdminUnit ?adminUnit .
   ?adminUnit rdfs:label ?adminUnitName .
   ?adminUnit geo:hasGeometry/geo:asWKT ?adminWKT .
   FILTER(bif:st_intersects (?adminWKT, ?roiWkt)) .

   #Step  5 - Keywords from Twitter
   ?adminUnit o-socialData:hasSocialData ?adminSocialData .
   ?adminSocialData time:hasBeginning/time:inXSDDateTimeStamp ?↩
       socialDataStart .
   ?adminSocialData time:hasEnd/time:inXSDDateTimeStamp ?socialDataEnd .
   ?adminSocialData o-socialData:hasKeyword ?socialDataKeyword .
   ?socialDataKeyword o-socialData:hasValue ?keywordAdminUnit .
   ?socialDataKeyword o-socialData:hasFrequency ?keywordFrequency .
   FILTER(?socialDataStart >= ?start && ?socialDataEnd <= ?end)

}
ORDER BY ?adminUnitName
```

Liste des acronymes

API Application Programming Interface. 45, 79, 81, 82, 90, 93, 94, 113, 128

BAI Burned Area Index. 59, 69, 70, 77, 78, 97, 98, 102, 103, 105–108, 115

BFO Basic Formal Ontology. 74

CCT Centre Canadien de Télédétection. 16–19, 127

CESBIO Centre d'Etudes Spatiales de la Biosphère. 20, 21, 70, 72, 73, 80, 81, 93, 127, 128

CLC CORINE Land Cover. 65

CNN Convolutional Neuron Network. 22, 52, 105

CNRS Centre National de la Recherche Scientifique. 7, 111

DBSCAN Density-Based Spatial Clustering of Applications with Noise. 56, 67

DCMI Dublin Core Metadata Initiative. 50

DUL DOLCE+DnS Ultra Lite. 52

EPSG European Petroleum Survey Group. 11, 75, 81

ESA European Space Agency. 1, 5, 12, 16, 56, 68, 75, 82

FAIR Findable, Accessible, Interoperable, and Reusable. 10

FAO Food and Agriculture Organization. 21, 72

FOAF Friend Of A Friend. 31, 32

FOSS4G Free and Open Source Software for Geospatial. 10

GADM Global Administrative Areas. 46, 57, 100

GDAL Geospatial Data Abstraction Library. 10, 81, 82, 84, 97

GeoJSON Geographic JSON. 14, 15

GeoTIFF Geographic Tagged Image File Format. 12, 81

GLC-SHARE Global Land Cover-SHARE. 21, 73, 81

GML Geography Markup Language. 14, 38, 41

GPS Global Positioning System. 1, 11, 15

GPU Graphics Processing Unit. 22

HDFS Hadoop Distributed File System. 56

HTTP HyperText Transfer Protocol. 25

IETF Internet Engineering Task Force. 15

INSEE Institut National de la Statistique et des Études Économiques. 93

IoT Internet of Things. 15

IRIT Institut de Recherche en Informatique Toulousain. 101

JPEG Joint Photographic Experts Group. 12

JSON JavaScript Object Notation. 12, 79, 82, 86, 101

LBARD Linked Brazilian Amazon Rainforest Data. 48

LC3 Land Cover Change Continuum. 49, 50, 127

LOD Linked Open Data. 24–26, 29, 31, 36, 44, 45, 48, 66, 69, 90, 91, 100, 113, 114, 127

N3 Notation3. 28

NASA National Aeronautics and Space Administration. 16, 44

NDCI Normalized Difference Cloud Index. 21

NDSI Normalized Difference Snow index. 21, 54, 70, 113, 115

NDVI Normalized Difference Vegetation Index. 6, 7, 12, 21, 22, 54, 59, 65, 66, 69, 73–78, 82, 83, 90–92, 106, 111, 112, 128

NDWI Normalized Difference Water index. 21, 98, 113, 115

NLTK Natural Language ToolKit. 102

NOA National Observatory of Athens. 48

NOAA National Oceanic and Atmospheric Administration. 45

NUTS Nomenclature des Unités Territoriales Statistiques. 51

OGC Open Geospatial Consortium. 10, 12, 14, 15, 37, 39, 41, 43, 66

OLA Open Linked Amazon. 47

ONG Organisation Non Gouvernementale. 10

OSGeo Open Source Geospatial Foundation. 10, 104

OSM OpenStreetMap. 45, 46, 48, 53, 77, 100, 104

OWL Web Ontology Language. 27, 29, 30, 32, 38, 39, 42, 54, 127

PAV Provenance, Authoring and Versioning. 50

PROV-O PROV Ontology. 74

R2RML RDB to RDF Mapping Language. 51

RCC8 Region Connection Calculus 8. 37, 38, 40, 42, 44, 74, 127

RDF Resource Description Framework. 25, 27–31, 35, 36, 38, 40, 41, 45–48, 51, 54, 55, 57, 59, 67, 69, 72, 79–83, 86, 87, 91, 93, 97, 100, 102, 115, 127, 128

RDFS RDF Schema. 25, 27, 29, 30

RFC Request for comments. 15

ROI Region Of Interest. 6, 7, 65, 67, 68, 70, 77, 78, 83–86, 90, 95–98, 102–110, 112, 115, 129

RRT Rapidly-exploring Random Tree. 53

RSS Really Simple Syndication. 38, 56

RuleML Rule Markup Language. 32

SAR Synthetic-Aperture Radar. 105

SIG Système d'Information Géographique. 10, 11, 13–15, 102, 104, 106–109, 127, 129

SKOS Simple Knowledge Organization System. 72

SNCF Société Nationale des Chemins de Fer français. 45

SOSA Sensor, Observation, Sample, and Actuator. 69, 74, 75, 128

SPARQL SPARQL Protocol And RDF Query Language. 25, 27, 31–33, 39–41, 45, 46, 48, 49, 51, 52, 55, 57, 82, 83, 90, 100, 101, 108, 112, 127

SQL Structured Query Language. 31

SSN Semantic Sensor Network. 74

SVM Support Vector Machines. 20, 58

SWEET Semantic Web for Earth and Environmental Terminology. 42–44, 48, 66, 127

SWRL Semantic Web Rule Language. 32

TIFF Tagged Image File Format. 12

TISC Open Time and Space Core Vocabulary. 47

TSN Territorial Statistical Nomenclature. 50, 51, 127

UCS Union of Concerned Scientists. 15

URI Uniform Resource Identifier. 25–29, 81, 82, 87, 91, 101

URL Uniform Resource Locator. 28

USGS United States Geological Survey. 16

UTM Universal Transverse Mercator. 11

VRT Virtual Raster. 81

W3C World Wide Web Consortium. 23, 25, 27–31, 37, 38, 40–44, 66, 127

WGS84 World Geodetic System 1984. 11, 38, 39, 45, 81

WKT Well-Known Text. 14, 40, 41, 78, 81, 93, 101

WPS Web Processing Service. 10

XML Extensible Markup Language. 12, 14, 15, 27, 28, 81, 101

Table des figures

Bibliographie

[Abbas *et al.* 2016] Arbab Abbas, N. Minallh, Nasir Ahmad, Sahibzada Abdur Rehman Abid et Muhammad Khan. *K-Means and ISODATA Clustering Algorithms for Landcover Classification Using Remote Sensing.* Sindh University Research Journal (Science Series), vol. 48, pages 315–318, 04 2016. (Cité en page 20.)

[Alirezaie *et al.* 2017] Marjan Alirezaie, Andrey Kiselev, Martin Längkvist, Franziska Klügl et Amy Loutfi. *An Ontology-Based Reasoning Framework for Querying Satellite Images for Disaster Monitoring.* Sensors (Switzerland), vol. 17, 11 2017. (Cité en pages 40, 52, 53 et 127.)

[Allen 1983] James F. Allen. *Maintaining Knowledge about Temporal Intervals.* Commun. ACM, vol. 26, no. 11, page 832–843, November 1983. (Cité en pages 42 et 127.)

[Amin *et al.* 2016] Arabi Mohammed El Amin, Qingjie Liu et Yunhong Wang. *Convolutional neural network features based change detection in satellite images.* Dans Xudong Jiang, Guojian Chen, Genci Capi et Chiharu Ishll, editeurs, First International Workshop on Pattern Recognition, volume 10011, pages 181 – 186. International Society for Optics and Photonics, SPIE, 2016. (Cité en pages 3 et 22.)

[Arano 2005] Silvia Arano. *Thesauruses and ontologies.* Hipertext. net, vol. 3, 2005. (Cité en page 24.)

[Arenas *et al.* 2016] Helbert Arenas, Nathalie Aussenac-Gilles, Catherine Comparot et Cassia Trojahn. *Semantic Integration of Geospatial Data from Earth Observations.* Dans Knowledge Engineering and Knowledge Management - EKAW 2016 Satellite Events, pages 97–100, Bologna (It), 2016. Springer. (Cité en page 79.)

[Arenas *et al.* 2018] Helbert Arenas, Nathalie Aussenac-Gilles, Catherine Comparot et Cássia Trojahn. *Ontologie pour l'intégration de données d'observation de la Terre et contextuelles basée sur les relations topologiques.* Dans 29es Journées Francophones d'Ingénierie des Connaissances, pages 5–20, Nancy, France, 2018. AFIA. (Cité en pages 69 et 73.)

[Asokan & Anitha 2019] Anju Asokan et J. Anitha. *Change detection techniques for remote sensing applications : a survey.* Earth Science Informatics, vol. 12, no. 2, pages 143–160, Jun 2019. (Cité en page 2.)

[Aubrun *et al.* 2020] M. Aubrun, A. Troya-Galvis, M. Albughdadi, R. Hugues et M. Spigai. *Unsupervised learning of robust representations for change de-*

tection on Sentinel-2 Earth Observation images. Dans 13th Int. Symp. on Environmental Software Systems, 2020. (Cité en pages 23, 83 et 95.)

[Augustin *et al.* 2018] Hannah Augustin, Martin Sudmanns, Dirk Tiede et Andrea Baraldi. *A Semantic Earth Observation Data Cube for Monitoring Environmental Changes during the Syrian Conflict.* GI_Forum, vol. 1, pages 214–227, 2018. (Cité en page 52.)

[Ban *et al.* 2020] Yifang Ban, Puzhao Zhang et et. al. Nascetti. *Near Real-Time Wildfire Progression Monitoring with Sentinel-1 Time Series and Deep Learning.* SR, vol. 10, 2020. (Cité en page 105.)

[Battle & Kolas 2012] Robert Battle et Dave Kolas. *Enabling the Geospatial Semantic Web with Parliament and GeoSPARQL.* Semantic Web, vol. 3, no. October 2012, pages 355–370, 2012. (Cité en page 39.)

[Baucic & Medak 2014] Martina Baucic et Damir Medak. *Building the Semantic Web for Earth Observations.* Dans DailyMeteo.org/2014 Conference - Belgrade, Serbia, pages 76–81, 06 2014. (Cité en page 47.)

[Behera *et al.* 2021] Mukunda Behera, Surbhi Barnwal, Somnath Paramanik, Pulakesh Das, Bimal Bhattyacharya, Jagadish Buddolla, Parth Roy, Sujit Ghosh, Soumit Behera et Javier Marcello-Ruiz. *Species-Level Classification and Mapping of a Mangrove Forest Using Random Forest-Utilisation of AVIRIS-NG and Sentinel Data.* Remote Sensing, vol. 13, 05 2021. (Cité en page 2.)

[Bernard *et al.* 2018] Camille Bernard, Marlène Villanova-Oliver, Jérôme Gensel et Hy Dao. *Modeling changes in territorial partitions over time : ontologies TSN and TSN-change.* Dans the 33rd Annual ACM Symposium, Pau, France, April 2018. ACM Press. (Cité en pages 50, 51 et 127.)

[Berners-Lee *et al.* 2001] Tim Berners-Lee, James Hendler et Ora Lassila. *The Semantic Web.* Scientific American, vol. 284, no. 5, pages 34–43, May 2001. (Cité en page 1.)

[Berners-Lee 1998] Tim Berners-Lee. *The Semantic Web road map.* `https://www.w3.org/DesignIssues/Semantic.html`, 1998. (Cité en page 23.)

[Blázquez *et al.* 2012] Luis Manuel Vilches Blázquez, Victor Saquicela et Óscar Corcho. *Interlinking Geospatial Information in the Web of Data.* Dans AGILE Conference, pages 119–139, 2012. (Cité en page 47.)

[Blower *et al.* 2017] Jon Blower, Maik Riechert et Bill Roberts. *Overview of the CoverageJSON format.* `https://www.w3.org/TR/covjson-overview/`, July 2017. (Cité en page 41.)

[Borgo & Masolo 2010] Stefano Borgo et Claudio Masolo. Ontological foundations of dolce, pages 279–295. Springer Verlag, Berlin, 09 2010. (Cité en page 25.)

[Borst 1997] Willem Nico Borst. *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*. PhD thesis, Institute for Telematica and Information Technology, University of Twente, Enschede, The Netherlands, 1997. (Cité en page 24.)

[Bouyerbou *et al.* 2019] Hafidha Bouyerbou, Kamal Bechkoum et Richard Lepage. *Geographic ontology for major disasters : Methodology and implementation*. International Journal of Disaster Risk Reduction, vol. 34, pages 232–242, 2019. (Cité en pages 52 et 115.)

[Brickley & Guha 2014] Dan Brickley et R.V. Guha. *RDF Schema 1.1*. `https://www.w3.org/TR/rdf-schema/`, February 2014. (Cité en page 29.)

[Brickley 2003] Dan Brickley. *Basic Geo (WGS84 lat/long) Vocabulary*. `https://www.w3.org/2003/01/geo/`, January 2003. (Cité en page 38.)

[Butler *et al.* 2016] H. Butler, M. Daly, A. Doyle, Sean Gillies, T. Schaub et T. Schaub. *The GeoJSON Format*. RFC 7946, August 2016. (Cité en page 15.)

[Charalambidis *et al.* 2015] Angelos Charalambidis, Antonis Troumpoukis et Stasinos Konstantopoulos. *SemaGrow : Optimizing Federated SPARQL Queries*. Dans Proceedings of the 11th International Conference on Semantic Systems, SEMANTICS '15, page 121–128, New York, NY, USA, 2015. Association for Computing Machinery. (Cité en page 57.)

[de Jong & Bosman 2019] Kevin Louis de Jong et Anna Sergeevna Bosman. *Unsupervised Change Detection in Satellite Images Using Convolutional Neural Networks*, 2019. (Cité en page 4.)

[Ding *et al.* 2020] Linfang Ding, Guohui Xiao, Diego Calvanese et Liqiu Meng. *A Framework Uniting Ontology-Based Geodata Integration and Geovisual Analytics*. ISPRS International Journal of Geo-Information, vol. 9, no. 8, 2020. (Cité en pages 36 et 47.)

[Dorne *et al.* 2018] Jordane Dorne, Nathalie Aussenac-Gilles, Catherine Comparot, Romain Hugues, Jean-Guy Planès et Cassia Trojahn dos Santos. *Une approche sémantique pour représenter l'indice de végétation d'images Sentinel-2 et son évolution*. Dans Spatial Analytics and GEOmatics (SAGEO 2018), pages 49–54, Montpellier, France, November 2018. (Cité en page 76.)

[Dorne *et al.* 2020] Jordane Dorne, Nathalie Aussenac-Gilles, Catherine Comparot, Romain Hugues et Cassia Trojahn. *LandCover2RDF : An API for Computing the Land Cover of a Geographical Area and Generating the RDF Graph*. Dans The Semantic Web : ESWC 2020 Satellite Events, pages 73–78, Cham, 2020. Springer International Publishing. (Cité en page 73.)

[Dorne *et al.* 2021] Jordane Dorne, Nathalie Aussenac-Gilles, Catherine Comparot, Romain Hugues et Cassia Trojahn dos Santos. *From EO Change Rasters to Knowledge Graphs : An approach Based on Regions of Interest.* Dans ESWC 2021 workshops : GeoLD 2021 , pages 1–12, Hersonissos, Greece, June 2021. CEUR-WS. (Cité en page 78.)

[Drummond & Shearer 2006] Nick Drummond et Rob Shearer. *The Open World Assumption.* `https://www.cs.man.ac.uk/~drummond/presentations/OWA.pdf`, 2006. (Cité en page 27.)

[Dumitru *et al.* 2018a] Octavian Dumitru, Mihai Datcu et Gottfried Schwarz. *CANDELA deliverable D2.1 : Data Mining v1.* Rapport technique, CANDELA Consortium, H2020 Grant Agreement No. 776193, 2018. (Cité en page 58.)

[Dumitru *et al.* 2018b] Octavian Dumitru, Mihai Datcu et Gottfried Schwarz. *CANDELA deliverable D2.7 : Data Fusion v1.* Rapport technique, CANDELA Consortium, H2020 Grant Agreement No. 776193, 2018. (Cité en page 58.)

[Dumitru *et al.* 2019] Corneliu Octavian Dumitru, Gottfried Schwarz, Anna Pulak-Siwiec, Bartosz Kulawik, Jose Lorenzo et Mihai Datcu. *Earth Observation Data Mining : A Use Case for Forest Monitoring.* Dans IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, pages 5359–5362, 2019. (Cité en page 58.)

[Dumitru *et al.* 2020] Octavian Dumitru, Mihai Datcu et Gottfried Schwarz. *CANDELA deliverable D2.8 : Data Fusion v2.* Rapport technique, CANDELA Consortium, H2020 Grant Agreement No. 776193, 2020. (Cité en page 58.)

[Egenhofer 2002] Max J. Egenhofer. *Toward the Semantic Geospatial Web.* Dans Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems, GIS '02, pages 1–4, New York, NY, USA, 2002. ACM. (Cité en page 1.)

[Espinoza-Molina *et al.* 2015] D. Espinoza-Molina, C. Nikolaou, O. Dumitru, K. Bereta, M. Koubarakis, G. Schwarz et M. 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, vol. 8, no. 4, pages 1696 – 1708, 2015. (Cité en pages 52 et 72.)

[Filipponi 2018] Federico Filipponi. *BAIS2 : Burned Area Index for Sentinel-2.* Dans Proc. of the 2nd International Electronic Conference on Remote Sensing, volume 2 (7), page 7, 2018. (Cité en page 98.)

[F.Y *et al.* 2017] Osisanwo F.Y, T. AkinsolaJ.E., O. Awodele, O. HinmikaiyeJ., O. Olakanmi et J. Akinjobi. *Supervised Machine Learning Algorithms :*

Classification and Comparison. International Journal of Computer Trends and Technology, vol. 48, pages 128–138, 2017. (Cité en page 20.)

[Gandhi *et al.* 2015] G. Meera Gandhi, S. Parthiban, Nagaraj Thummalu et A. Christy. *Ndvi : Vegetation Change Detection Using Remote Sensing and Gis – A Case Study of Vellore District*. Procedia Computer Science, vol. 57, pages 1199–1210, 2015. 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015). (Cité en page 22.)

[Giglio *et al.* 2003] Louis Giglio, Jacques Descloitres, Christopher O. Justice et Yoram J. Kaufman. *An Enhanced Contextual Fire Detection Algorithm for MODIS*. Remote Sensing of Environment, vol. 87, no. 2, pages 273–282, 2003. (Cité en page 99.)

[Gong *et al.* 2017] Maoguo Gong, Hailun Yang et Puzhao Zhang. *Feature learning and change feature classification based on deep learning for ternary change detection in SAR images*. ISPRS Journal of Photogrammetry and Remote Sensing, vol. 129, pages 212 – 225, 2017. (Cité en pages 3 et 22.)

[Gruber 1993] Thomas R Gruber. *A translation approach to portable ontology specifications*. Knowledge acquisition, vol. 5, no. 2, pages 199–220, 1993. (Cité en page 24.)

[Gu *et al.* 2017] H. Gu, Haitao Li, Li Yan, Zhengjun Liu, T. Blaschke et U. Soergel. *An Object-Based Semantic Classification Method for High Resolution Remote Sensing Imagery Using Ontology*. Remote. Sens., vol. 9, page 329, 2017. (Cité en page 52.)

[Guarino *et al.* 2009] Nicola Guarino, Daniel Oberle et Steffen Staab. What is an ontology ?, pages 1–17. Springer Verlag, Berlin, 05 2009. (Cité en page 24.)

[Haas & Ban 2017] Jan Haas et Yifang Ban. *Sentinel-1A SAR and sentinel-2A MSI data fusion for urban ecosystem service mapping*. Remote Sensing Applications : Society and Environment, vol. 8, pages 41–53, 2017. (Cité en page 2.)

[Harbelot *et al.* 2015] Benjamin Harbelot, Helbert Arenas et Christophe Cruz. *LC3 : un modèle spatial et sémantique pour découvrir la connaissance dans les jeux de données géospatiaux*. Dans Extraction et Gestion de la Connaissance, Luxembourg, Luxembourg, January 2015. (Cité en pages 49, 50, 59, 66 et 127.)

[Hobbs & Pan 2004] Jerry R. Hobbs et Feng Pan. *An Ontology of Time for the Semantic Web*. ACM Transactions on Asian Language Information Processing, vol. 3, no. 1, page 66–85, March 2004. (Cité en page 43.)

[Hoffart *et al.* 2013] Johannes Hoffart, Fabian Suchanek, Klaus Berberich et Gerhard Weikum. *YAGO2 : A Spatially and Temporally Enhanced Knowledge*

Base from Wikipedia. Artificial Intelligence, vol. 194, pages 28–61, 01 2013. (Cité en page 45.)

[Homburg *et al.* 2020] Timo Homburg, Steffen Staab et Daniel Janke. *GeoSPARQL+ : Syntax, Semantics and System for Integrated Querying of Graph, Raster and Vector Data.* Dans Jeff Z. Pan, Valentina Tamma, Claudia d'Amato, Krzysztof Janowicz, Bo Fu, Axel Polleres, Oshani Seneviratne et Lalana Kagal, editeurs, The Semantic Web – ISWC 2020, pages 258–275, Cham, 2020. Springer International Publishing. (Cité en page 41.)

[Janowicz *et al.* 2012] Krzysztof Janowicz, Simon Scheider, Todd Pehle et Glen Hart. *Geospatial Semantics and Linked Spatiotemporal Data – Past , Present , and Future.* Semantic Web, no. 3, pages 321–332, 2012. (Cité en page 1.)

[Janowicz *et al.* 2019] Krzysztof Janowicz, Armin Haller, Simon J.D. Cox, Danh Le Phuoc et Maxime Lefrançois. *SOSA : A lightweight ontology for sensors, observations, samples, and actuators.* Journal of Web Semantics, vol. 56, pages 1–10, 2019. (Cité en pages 74, 75 et 128.)

[Jia *et al.* 2014] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama et Trevor Darrell. *Caffe : Convolutional Architecture for Fast Feature Embedding.* arXiv preprint arXiv :1408.5093, 2014. (Cité en pages 3 et 22.)

[Jovanovik *et al.* 2021] Milos Jovanovik, Timo Homburg et Mirko Spasić. *A GeoSPARQL Compliance Benchmark*, 2021. (Cité en page 40.)

[Karalis *et al.* 2019] Nikolaos Karalis, Georgios Mandilaras et Manolis Koubarakis. *Extending the YAGO2 Knowledge Graph with Precise Geospatial Knowledge.* Dans Chiara Ghidini, Olaf Hartig, Maria Maleshkova, Vojtěch Svátek, Isabel Cruz, Aidan Hogan, Jie Song, Maxime Lefrançois et Fabien Gandon, editeurs, The Semantic Web – ISWC 2019, pages 181–197, Cham, 2019. Springer International Publishing. (Cité en page 46.)

[Kauppinen *et al.* 2008] Tomi Kauppinen, Jari Väätäinen et Eero Hyvönen. *Creating and Using Geospatial Ontology Time Series in a Semantic Cultural Heritage Portal.* Dans ESWC2008, volume LNCS 5021, pages 110–123. Springer-Verlag, Berlin, 06 2008. (Cité en page 47.)

[Kauppinen *et al.* 2013] Tomi Kauppinen, Giovana Espindola, Jim Jones, Alber Ipia, Benedikt Graeler et Thomas Bartoschek. *Linked Brazilian Amazon Rainforest Data.* Semantic Web Journal, vol. 5, 01 2013. (Cité en page 47.)

[Kolas *et al.* 2013] Dave Kolas, Matthew Perry et John Herring. *Getting started with GeoSPARQL.* Rapport technique, OGC, 2013. (Cité en page 39.)

[Koubarakis & Kyzirakos 2010] Manolis Koubarakis et Kostis Kyzirakos. *Modeling and Querying Metadata in the Semantic Sensor Web : The Model stRDF*

and the Query Language stSPARQL. Dans Lora Aroyo, Grigoris Antoniou, Eero Hyvönen, Annette ten Teije, Heiner Stuckenschmidt, Liliana Cabral et Tania Tudorache, editeurs, The Semantic Web : Research and Applications, pages 425–439, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. (Cité en pages 40, 41 et 127.)

[Kurnaz *et al.* 2020] Bahadır Kurnaz, Caglar Bayik et Saygin Abdikan. *Forest Fire Area Detection by Using Landsat-8 and Sentinel-2 Satellite Images : A Case Study in Mugla, Turkey*, 05 2020. (Cité en page 106.)

[Kyzirakos *et al.* 2012] Kostis Kyzirakos, Manos Karpathiotakis et Manolis Koubarakis. *Strabon : A Semantic Geospatial DBMS.* Dans Philippe Cudré-Mauroux, Jeff Heflin, Evren Sirin, Tania Tudorache, Jérôme Euzenat, Manfred Hauswirth, Josiane Xavier Parreira, Jim Hendler, Guus Schreiber, Abraham Bernstein et Eva Blomqvist, editeurs, The Semantic Web – ISWC 2012, pages 295–311, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. (Cité en page 41.)

[Kyzirakos *et al.* 2014a] K. Kyzirakos, M. Karpathiotakis, G. Garbis, C. Nikolaou, K. Bereta, I. Papoutsis, T. Herekakis, D. Michail, M. Koubarakis et C. Kontoes. *Wildfire monitoring using satellite images, ontologies and linked geospatial data.* Journal of Web Semantics, vol. 24, pages 18–26, 2014. The Semantic Web Challenge 2012. (Cité en pages 48 et 66.)

[Kyzirakos *et al.* 2014b] Kostis Kyzirakos, Ioannis Vlachopoulos, Dimitrianos Savva, Stefan Manegold et Manolis Koubarakis. *GeoTriples : A tool for publishing geospatial data as RDF graphs using R2RML mappings.* CEUR Workshop Proceedings, vol. 1401, pages 33–44, 01 2014. (Cité en page 51.)

[Kyzirakos *et al.* 2018] Kostis Kyzirakos, Dimitrianos Savva, Ioannis Vlachopoulos, Alexandros Vasileiou, Nikolaos Karalis, Manolis Koubarakis et Stefan Manegold. *GeoTriples : Transforming geospatial data into RDF graphs using R2RML and RML mappings.* Journal of Web Semantics, vol. 52-53, pages 16–32, 2018. (Cité en page 57.)

[Langlois 2004] Patrice Langlois. *Géomatique.* `http://www.hypergeo.eu/IMG/_article_PDF/article_68.pdf`, June 2004. (Cité en page 1.)

[Laublet *et al.* 2004] Philippe Laublet, Jean Charlet et Chantal Reynaud. *Introduction au Web sémantique.* Revue I3, vol. Hors série sur le Web sémantique, pages 7–20, 2004. (Cité en page 23.)

[LaValle 1998] Steven M. LaValle. *Rapidly-Exploring Random Trees : A New Tool for Path Planning.* The annual research report, 1998. (Cité en page 53.)

[Li *et al.* 2016] Wenwen Li, Xiran Zhou et Sheng Wu. *An Integrated Software Framework to Support Semantic Modeling and Reasoning of Spatiotemporal*

Change of Geographical Objects : A Use Case of Land Use and Land Cover Change Study. ISPRS International Journal of Geo-Information, vol. 5, no. 10, 2016. (Cité en page 113.)

[Lieberman *et al.* 2007] Joshua Lieberman, Raj Singh et Chris Goad. *W3C Geospatial Vocabulary.* `https://www.w3.org/2005/Incubator/geo/XGR-geo/`, October 2007. (Cité en pages 38 et 127.)

[Lin *et al.* 2016] Ya Lin, Hao Xu et Yuqi Bai. *Semantically Enhanced Catalogue Search Model for Remotely Sensed Imagery.* IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. PP, pages 1–9, 11 2016. (Cité en page 52.)

[Marujo *et al.* 2015] Luís Marujo, Wang Ling, Isabel Trancoso, Chris Dyer, Alan W. Black, Anatole Gershman, David Martins de Matos, João Neto et Jaime Carbonell. *Automatic Keyword Extraction on Twitter.* Dans Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2 : Short Papers), pages 637–643, Beijing, China, July 2015. Association for Computational Linguistics. (Cité en page 114.)

[Miron 2009] Alina Dia Miron. *Découverte d'associations sémantiques pour le Web Sémantique Géospatial - le framework ONTOAST.* Theses, Université Joseph-Fourier - Grenoble I, December 2009. (Cité en pages 39 et 127.)

[Neptune 2020] Nathalie Neptune. *Automatic Annotation of Change in Earth Observation Imagery.* Dans CIRCLE, 2020. (Cité en page 52.)

[Nielsen 2011] Finn Årup Nielsen. *A new ANEW : Evaluation of a word list for sentiment analysis in microblogs.* CoRR, vol. abs/1103.2903, 2011. (Cité en page 115.)

[Noy & Mcguinness 2001] N. Noy et Deborah Mcguinness. *Ontology Development 101 : A Guide to Creating Your First Ontology.* Knowledge Systems Laboratory, vol. 32, 01 2001. (Cité en page 24.)

[Pathak *et al.* 2018] Ajeet Ram Pathak, Manjusha Pandey et Siddharth Rautaray. *Application of Deep Learning for Object Detection.* Procedia Computer Science, vol. 132, pages 1706 – 1717, 2018. International Conference on Computational Intelligence and Data Science. (Cité en page 22.)

[Pavlis *et al.* 2012] Nikolaos K. Pavlis, Simon A. Holmes, Steve C. Kenyon et John K. Factor. *The development and evaluation of the Earth Gravitational Model 2008 (EGM2008).* Journal of Geophysical Research : Solid Earth, vol. 117, no. B4, 2012. (Cité en pages 11 et 127.)

[Perry *et al.* 2012] Matthew Perry, John Herring, Nicholas J. Car, Timo Homburg et Simon J.D. Cox. *OGC GeoSPARQL - A geographic query language for*

RDF data : GeoSPARQL 1.1 draft. OGC implementation standard draft. Rapport technique, OGC, 2012. (Cité en pages 39 et 41.)

[Pittaras *et al.* 2019] Nikiforos Pittaras, George Papadakis, George Stamoulis, Giorgos Argyriou, Efi Karra Taniskidou, Emmanouil Thanos, George Giannakopoulos, Leonidas Tsekouras et Manolis Koubarakis. *GeoSensor : Semantifying Change and Event Detection over Big Data.* Dans Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, SAC '19, page 2259–2266, New York, NY, USA, 2019. Association for Computing Machinery. (Cité en pages 55, 56, 60, 67 et 127.)

[Pritt & Chern 2020] Mark Pritt et Gary Chern. *Satellite Image Classification with Deep Learning*, 2020. (Cité en page 22.)

[Pérez Luque *et al.* 2015] Antonio Jesús Pérez Luque, Ramón Pérez Pérez, Francisco Javier Bonet et P.J. Magaña. *An ontological system based on MODIS images to assess ecosystem functioning of Natura 2000 habitats : A case study for Quercus pyrenaica forests.* International Journal of Applied Earth Observation and Geoinformation, vol. 37, 05 2015. (Cité en pages 54, 55, 66, 67 et 127.)

[Radke *et al.* 2005] R. J. Radke, S. Andra, O. Al-Kofahi et B. Roysam. *Image change detection algorithms : a systematic survey.* IEEE Transactions on Image Processing, vol. 14, no. 3, pages 294–307, March 2005. (Cité en page 2.)

[Randell *et al.* 1992] David Randell, Zhan Cui et Anthony Cohn. *A Spatial Logic based on Regions and Connection.* Principles of Knowledge Representation and Reasoning : Proceedings of the 1st International Conference, pages 165–176, 01 1992. (Cité en pages 37 et 127.)

[Raskin 2005] R. Raskin. *SWEET- An Upper Level Ontology for Earth System Science.* Dans AGU Fall Meeting Abstracts, volume 2005, pages IN41B–06, December 2005. (Cité en page 44.)

[Redmon *et al.* 2016] Joseph Redmon, Santosh Divvala, Ross Girshick et Ali Farhadi. *You Only Look Once : Unified, Real-Time Object Detection*, 2016. (Cité en page 115.)

[Reitsma & Albrecht 2005] F. Reitsma et J. Albrecht. *Modeling with the semantic Web in the geosciences.* IEEE Intelligent Systems, vol. 20, no. 2, pages 86–88, 2005. (Cité en page 1.)

[Ritter & Ruth 1997] N. Ritter et M. Ruth. *The GeoTiff data interchange standard for raster geographic images.* International Journal of Remote Sensing, vol. 18, no. 7, pages 1637–1647, 1997. (Cité en page 12.)

[Rolland *et al.* 2020] Jean-François Rolland, Fabien Castel, Anne Haugommard, Michelle Aubrun, Wei Yao, Corneliu Octavian Dumitru, Mihai Datcu, Michal Bylicki, Ba-Huy Tran, Nathalie Aussenac-Gilles, Catherine Comparot et Cássia Trojahn. *Candela : A Cloud Platform for Copernicus Earth Observation Data Analytics.* Dans IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2020, Waikoloa, HI, USA, September 26 - October 2, 2020, pages 3104–3107. IEEE, 2020. (Cité en page 58.)

[Smeros & Koubarakis 2016] Panayiotis Smeros et Manolis Koubarakis. *Discovering Spatial and Temporal Links among RDF Data.* Dans Workshop on Linked Data on the Web, 2016. (Cité en page 47.)

[Stadler *et al.* 2012] Claus Stadler, Jens Lehmann, Konrad Höffner et Sören Auer. *LinkedGeoData : A Core for a Web of Spatial Open Data.* Semantic Web Journal, vol. 3, no. 4, pages 333–354, 2012. (Cité en page 45.)

[Stoian *et al.* 2019] Andrei Stoian, Vincent Poulain, Jordi Inglada, Victor Poughon et Dawa Derksen. *Land Cover Maps Production with High Resolution Satellite Image Time Series and Convolutional Neural Networks : Adaptations and Limits for Operational Systems.* Remote Sensing, vol. 11, page 1986, 08 2019. (Cité en page 21.)

[Studer *et al.* 1998] Rudi Studer, V Richard Benjamins et Dieter Fensel. *Knowledge engineering : principles and methods.* Data & knowledge engineering, vol. 25, no. 1-2, pages 161–197, 1998. (Cité en page 23.)

[Théau 2008] Jérôme Théau. Change detection, pages 77–84. Springer US, Boston, MA, 2008. (Cité en page 2.)

[Thiéblin 2019] Elodie Thiéblin. *Automatic Generation of Complex Ontology Alignments.* Theses, Université Paul Sabatier - Toulouse III, October 2019. (Cité en page 25.)

[Wang *et al.* 2015] Xiaolei Wang, Nengcheng Chen, Zeqiang Chen, Xunliang Yang et Jizhen Li. *Earth observation metadata ontology model for spatiotemporal-spectral semantic-enhanced satellite observation discovery : A case study of soil moisture monitoring.* GIScience & Remote Sensing, vol. 53, pages 1–23, 09 2015. (Cité en page 52.)

[Wang *et al.* 2018] Chao Wang, Nengcheng Chen, Wei Wang et Zeqiang Chen. *A Hydrological Sensor Web Ontology Based on the SSN Ontology : A Case Study for a Flood.* ISPRS International Journal of Geo-Information, vol. 7, no. 1, 2018. (Cité en page 47.)

[Wilkinson & et al 2016] Mark Wilkinson et et al. *The FAIR Guiding Principles for scientific data management and stewardship.* Nature Scientific Data, no. 160018, 2016. (Cité en page 10.)

[Yao *et al.* 2020] Wei Yao, Corneliu Octavian Dumitru, Jose Lorenzo et Mihai Datcu. *Data Mining on the Candela Cloud Platform.* Dans IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, pages 6945–6948, 2020. (Cité en pages 57 et 59.)

[Yao *et al.* 2021] Wei Yao, Anastasia Moumtzidou, Corneliu Octavian Dumitru, Stelios Andreadis, Ilias Gialampoukidis, Stefanos Vrochidis, Mihai Datcu et Ioannis Kompatsiaris. *Early and Late Fusion of Multiple Modalities in Sentinel Imagery and Social Media Retrieval.* Dans Pattern Recognition. ICPR International Workshops and Challenges, pages 591–606, Cham, 2021. Springer International Publishing. (Cité en page 58.)