Recherche d'information spatio-temporelle : Application aux images satellitaires

Jean-Christophe Desconnets

▶ **To cite this version:**

Jean-Christophe Desconnets. Recherche d'information spatio-temporelle : Application aux images satellitaires. Recherche d'information [cs.IR]. Université de Montpellier, 2017. tel-01649173v1

HAL Id: tel-01649173

https://hal.science/tel-01649173v1

Submitted on 27 Nov 2017 (v1), last revised 7 Feb 2018 (v2)

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

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

HDR

Habilitation à Diriger des Recherches

Spécialité : Informatique

Par

Jean-Christophe Desconnets

Recherche d'information spatio-temporelle : Application aux images satellitaires

Soutenue le 17 novembre 2017. Jury :

Jean-Paul Donnay - Professeur - Unité Géomatique - Université de Liège – Rapporteur

Anne Doucet – Professeur – LIP6 –Université Pierre et Marie Curie – Rapporteure

Jérôme Gensel - Professeur - LIG - Université Grenoble Alpes – Rapporteur

Carmen Gervet - Professeur – ESPACE-DEV - Université de Montpellier – Examinatrice

Thérèse - Libourel Professeur émérite - ESPACE-DEV - Université de Montpellier – Examinatrice

Isabelle Mougenot – Maître de conférence - ESPACE-DEV – Université de Montpellier - Examinatrice

« Le génie est fait d'1% d'inspiration et de 99% de transpiration »
Thomas Alva Edison (1847-1931)

« La valeur d'une idée dépend de son utilisation »
Thomas Alva Edison (1847-1931)

Table des matières

Listes des figures

Listes des tableaux

CV étendu

1. Curriculum Vitae

1.1 Identité

- **Nom** : Desconnets
- **Prénom** : Jean-Christophe
- **Né le** 18 décembre 1964 à Bordeaux
- **Nationalité** : Française
- **Email** : jean-christophe.desconnets@ird.fr
- **Adresse professionnelle** : Maison de la Télédétection. 500, rue Jean François Breton. 34093 Montpellier cedex 05.
- **Adresse ResearchGate** : https://www.researchgate.net/profile/Jean-Christophe_Desconnets

1.2 Formation

- **1999-2000** DESS informatique Appliquée aux Organisations. Université de Montpellier II
- **1991- 1994** Doctorat 3ème Cycle Science de l'Eau. Université de Montpellier II. Ecole doctorale : Génie Mécanique, Mécanique, Génie Civil.
- **1990** : DEA de Science du Sol, Institut National Agronomique Paris Grignon-Paris VI

1.3 Expérience professionnelle

Ingénieur de recherche IRD, spécialité géomatique

- **2010 – aujourd'hui** : Unité mixte de recherche ESPACE-DEV (IRD, UM, UR, UG)
- **2007 – 2010** : Unité de service ESPACE (IRD)
- **2000 – 2007** : Unité de Service Désertification (IRD)
- **1998 - 2000** : Consultant indépendant hydrologue et bases de données spatiales
- **1996 – 1998** : Expert international Nations Unies en poste à Ouagadougou (Burkina Faso)

2. Travaux de recherche

2.1 Note préliminaire

Des sciences de la Terre à la géomatique

Mon doctorat ainsi que mes premiers travaux de recherche (1987-1998) ont été menés dans les domaines des sciences de la Terre et de l'Eau. Ma principale contribution a porté sur l'étude du cycle de l'eau à l'échelle locale en vue de son extrapolation à l'échelle régionale. J'ai été amené à aborder cette problématique par la **modélisation spatialement distribuée des mécanismes de redistribution des eaux de surface**. Un des intérêts d'une telle approche est d'associer aux lois de la mécanique des fluides la représentation du milieu naturel afin de mieux paramétrer les modèles d'écoulement en tenant compte, notamment, de leur variabilité spatiale liée à celle du milieu. Une telle approche

demande de constituer, en amont des travaux de simulation, un ensemble de jeux de données spatio-temporels qui sont les supports à la détermination des paramètres des modèles physiques : pente, exposition, carte d'occupation des sols, carte d'humidité du sol.

Si les outils numériques permettent de reproduire les processus naturels de manière satisfaisante, la constitution des jeux de données spatio-temporelles de qualité (en précision géométrique, thématique et temporelle) relève d'un véritable défi qui va de la connaissance des observations disponibles, du choix de l'observation adéquate (nature, résolution spatiale et temporelle) à son accès puis son formatage en vue de son analyse pour représenter spatialement les paramètres physiques souhaités.

La plupart de ces défis, et notamment ceux portant sur les données ayant une dimension spatiale sont abordés par la **géomatique,** à l'interface entre les sciences et les technologies de mesure de la terre ainsi que les sciences de l'information pour faciliter l'acquisition, le traitement et **la diffusion des données** d'un territoire. Mes travaux de recherche allient donc différents aspects de l'analyse et du traitement de données, en particulier l'interopérabilité et l'intégration de données hétérogènes, interdisciplinaires et spatio-temporelles, pour en extraire de la connaissance.

2.2 Synthèse de mes travaux

Contexte

Les scientifiques, que ce soit dans les domaines de l'hydrologie, du climat, de l'écologie ou des sciences humaines, ont besoin d'observer et de collecter les éléments des systèmes étudiés et examiner précisément les interactions entre les changements climatiques, la dégradation de l'environnement, l'anthropisation afin d'en évaluer l'état, prédire les changements et proposer les adaptations nécessaires à la société.

Pour cela, de nombreux acteurs (scientifique, gouvernement, privé, citoyen) collectent de grandes masses de données en utilisant différents modes d'observation (capteur *in situ*, enquêtes de terrain, capteur aéroporté, satellite) afin de mesurer les conditions environnementales, dénombrer les espèces, identifier les pratiques des sociétés. Quelque soit le système étudié, l'analyse des interactions entre les différents éléments demande de croiser les données issues de différents capteurs pour représenter et analyser les phénomènes environnementaux qui sont, le plus souvent aux interfaces milieu et société. Ce besoin se heurte souvent à une approche mono-disciplinaire de la représentation d'un système ou de ses sous-ensembles. Elle se traduit par la mise à disposition des données et des modèles de données sous-jacents propres au champ disciplinaire. Ces spécificités rendent ainsi difficile les analyses qui doivent être menées conjointement. De fait, toutes ces données ne seront pas ou peu accessibles par les autres communautés scientifiques car les représentations du système étudié ne sont pas partagées ou/et acceptées par les autres champs disciplinaires. Par ailleurs, les descriptions nécessaires à leur identification, leur caractérisation et leur découverte sont souvent manquantes.

Thématiques

Mes travaux de recherche se situent dans ce contexte et se sont concentrés sur la mutualisation et le partage des données ainsi que leur traitements en environnement dans un souci d'ouverture à l'interdisciplinaire. Dans un premier temps, j'ai abordé cette problématique dans le contexte **des systèmes d'information en environnement** (SIE). Mes contributions intègrent différentes

propositions méthodologiques sur le rôle, la structuration des **métadonnées** pour assurer la découverte et la localisation des informations environnementales. Récemment, en mettant à profit les technologies associées à l'émergence des sciences du web, j'ai porté mes travaux dans le domaine du **web sémantique**. Les langages de représentation qui en sont issus constituent le socle pour mieux exprimer et organiser les informations et les connaissances dans le contexte de partage multi-disciplinaire qui pilote mes travaux.

Dans leur ensemble, les travaux de recherche que j'ai développés s'appuient sur les métadonnées en tant que support à la médiation de données et de leur traitements. Je les ai complété par l'utilisation des référentiels terminologiques et spatiaux qui viennent enrichir les métadonnées en vue d'améliorer la découverte, l'accès et le traitement des informations environnementales. Ce volet s'appuie sur la formalisation et l'exploitation des connaissances expertes des domaines étudiés via les terminologies et les ontologies. L'ensemble de mes travaux me permet d'aborder la conception des systèmes de manière formelle et de proposer des solutions innovantes et génériques. Les méthodologies employées relèvent principalement de l'ingénierie des modèles et de la connaissance et mettent en œuvre différents paradigmes tels que les modèles objets, l'algèbre relationnelle et la théorie des graphes. Mes contributions scientifiques sont organisées autour de trois aspects. Le premier porte sur la conception de systèmes d'information en environnement (SIE) en tant que dispositif technique en appui aux observatoires de la recherche [6, 12, 24, 25, 26, 28]. Les deux autres concernent l'interopérabilité des données pour les SIE et pour le web semantique. C'est sur ces deux derniers qui constituent le coeur de mes contributions scientifiques que je focalise ma synthèse. Elles seront détaillées dans mon mémoire.

Projets

Par ailleurs, mes travaux de recherche ont été, et sont toujours fortement associés à la réalisation concrètes de projets de systèmes d'information en environnement au niveau national et international. Ils me permettent de transférer les résultats de mes recherches vers la société et le monde industriel. Les sections 2.3 et 4 listent les réalisations opérationnelles auxquelles j'ai contribué, les licences déposées ainsi que les activités de transfert de technologies que j'ai menées.

2.2.1 Activités autour de l'interopérabilité des systèmes d'information en environnement

Rôle et structuration des métadonnées pour la découverte et la localisation des informations environnementales (2001 – 2007).

Collaborations : LIRMM, LIP6, TETIS (ex UMR3S), CIRAD
Publications : [15, 16, 21, 22, 23, 30, 31, 33]
Communications orales : [32, 37, 38, 39, 40, 45]
Réalisations : MDweb version 1.1, 1.2, 1.3, système de diffusion inter-observatoires ROSELT/OSS
Licence : CeCILL et dépôt à l'APP
Projets : ROSELT/OSS, PADOUE
Etudiants : Baranov A. (mastère pro, 100%), Moyroud N. (mastère pro, 100%), Granouillac B. (mastère pro, 100%)

Ces travaux s'inscrivent dans le contexte des systèmes d'information en environnement (SIE). Comme tout système d'information, les SIE englobent données et traitements. Cependant, ils présentent des spécificités relatives à la variété des thèmes abordés et au caractère spatial de l'information traitée. L'information (donnée et expertise) a été accumulée au fil des années et des

projets. La structuration et le stockage de cette information existent sous différents formats. Une des problématiques qui a émergé début des années 2000 a été celle de la diffusion et de l'accès à l'information partagée [Egenhofer 99]. Dans ce contexte, établir des corrélations entre des sources d'information issues de points de vue différents émis par des acteurs d'origines diverses demande à ce que l'accès à l'information se fasse de manière homogène bien que les ressources réelles soient hétérogènes et dispersées. Un service de catalogage est un des moyens de résoudre partiellement l'interopérabilité de ressources distribuées et hétérogènes [Claramunt 98]. Les moteurs de recherche du web pourraient potentiellement assurer un tel service de catalogage mais la description des métadonnées utilisées reste incomplète et peu explicitée [30]. En partant de ces constats, mes travaux se sont resserrés autour des fonctionnalités *de découverte et de localisation d'information environnementale via les métadonnées*. A cet effet, je me suis appuyé sur les métadonnées en leur affectant un double rôle: descriptif ou de représentation des connaissances (rôle sémantique) et un rôle de « facilitateur d'accès » (rôle d'index) [Lamb 01]. C'est en premier lieu sur ce rôle de facilitateur d'accès que des propositions de structuration des métadonnées ont été élaborées [31,33]. Mes efforts ont porté sur le raffinement des standards de la communauté géographique dont le principal est l'ISO 19115 [ISO 03]. L'utilisation de la notion de profil [ISO 04], qui correspond à celle proposé par UML [OMG, 15], permet de raffiner les descriptions proposées par les standards afin les adapter et les enrichir dans le but de tenir compte des besoins applicatifs propre à la communauté environnementale.

Ces réflexions ont été le socle pour la construction de l'outil MDweb [21, 22], outil générique et extensible, qui a été le support de tous les autres développements méthodologiques que je mentionne dans les paragraphes suivants. Ces travaux ont été réalisés grâce à la collaboration de plusieurs partenaires (UMR Tetis (IRSTEA, AgroParisTech, Cirad), Cepralmar, LIRMM et la Région Languedoc-Roussillon).

Association des référentiels terminologiques et des standards de métadonnées (2004-2010).

Collaborations : LIRMM, Cepralmar[1], UMR3S[2], Région Languedoc-Roussillon
Publications : [5, 7, 19, 29]
Communications orales : [44]
Projets : SYSCOLAG, FP7 NatureSDIplus
Réalisations : MDweb 1.4
Etudiants : Barde J. (thèse, 15%)

Dans un deuxième temps, le rôle de représentation de connaissances des métadonnées a été exploré. En effet, conscients que l'intérêt des outils de catalogage, pour les utilisateurs finaux, réside en grande part sur la pertinence des réponses apportées lors des recherches, nous avons mis l'accent sur l'apport complémentaire de sémantique aux rubriques présentes dans les structures de métadonnées [30]. Les apports ont consisté à étendre certains sections sur lesquelles l'indexation des informations portent (mots clés, empreinte géographique, nom de lieu, discipline…) [Barde 06]. Pour ce faire, les notions de référentiel terminologique et spatial ont été introduites et étudiées. Un référentiel terminologique décrit, pour une communauté donnée, la sémantique du domaine considéré par l'intermédiaire de modèles explicites. Un modèle traduit l'expertise d'une

[1] Cepralmar : Centre d'étude pour la promotion des activités lagunaires et maritimes

[2] UMR3S : Structures et Systèmes Spatiaux Cemagref – Engref

communauté. Il est le vecteur de l'interopérabilité sémantique entre acteurs pour partager la compréhension des concepts du domaine [20]. Une démarche de réflexion analogue a été menée sur la spatialité de l'information. Un référentiel spatial est un ensemble d'objets géographiques pertinents et donc de référence pour la communauté concernée. Les objets géographiques référents peuvent être représentés sous différentes formes : identifiants toponymiques, rectangle englobant décrivant l'emprise spatiale de l'objet, géométrie précise de l'objet. Les deux référentiels sont étroitement corrélés. La dimension spatiale constitue un « médiateur » pour guider la recherche d'information sous un angle thématique et sous l'angle des objets d'intérêts qui ont une dimension spatiale. Ces travaux ont été développés dans le cadre de la thèse de Julien Barde puis poursuivis et concrétisés (post-doc) au cours de la phase d 'opérationnalisation du service de catalogage du projet SYSCOLAG (cf. section 2.4) que j'ai coordonné. Concrètement, Plusieurs thesaurii de référence, gérés par un composant logiciel adhoc, ont été intégrées à l'outil MDweb. La standardisation de ces corpus de termes au format SKOS [Miles 05] a permis d'envisager leur gestion et leur utilisation de manière générique.

Exploitation des référentiels terminologique et spatial pour améliorer la recherche d'informations environnementales (2006-2012).

Collaborations : LIRMM, Cepralmar, UMR TETIS[3], Consortium NatureSDIplus, CNES[4]
Publications : [14, 17, 18, 20]
Projets : SYSCOLAG, FP7 NatureSDIplus, REFLECS
Réalisations : MDweb 1.5 et 1.6, portail REFLECS, portail pour le catalogage et la recherche des données de conservation de la nature, API web pour des thésaurus SKOS
Etudiants : Clerc S. (Mastère recherche, 50%), Boisson P. (Mastère recherche, 50%), Sayah H. (Mastère recherche, 50%), Laporte M.A. (post-doctorante, 100%), Boulet R. (post-doctorant, 50%)

Ces travaux viennent compléter les précédents et investissent les questions autour de l'aide apportée à un utilisateur dans ses activités de recherche d'information au sein des services de catalogage. Dans ce cadre, mes préoccupations ont été d'exploiter les référentiels terminologiques associés aux métadonnées (cf. paragraphe rôle et structuration des métadonnées) pour apporter des alternatives aux moteurs de recherche classiques et faciliter ainsi l'expérience utilisateur dans sa recherche de données spatio-temporelles. Les questions soulevées relèvent de la recherche d'information [Frakes 92] et du web sémantique [Berners-Lee 01]. L'originalité de la démarche consiste à intégrer la connaissance du domaine dans le processus de recherche en s'appuyant sur les propositions de représentation des connaissances expertes issues du web sémantique tel que le vocabulaire RDF SKOS [20]. L'annotation sémantique des métadonnées à l'aide des référentiels terminologiques nous apporte le contrôle des valeurs nécessaires à la mise en œuvre des différents mécanismes explorés. L'exploitation des concepts spatiaux et thématiques formalisés au sein des référentiels terminologiques ont donc guidé la réflexion et déterminé les axes méthodologiques explorés :

- ***Expansion de requêtes spatio-temporelles pour un service de catalogage***: L'expansion de requêtes peut être définie comme un processus de transformation d'une requête d'un utilisateur dans le but de lui apporter des réponses les plus pertinentes possibles. Des processus de

[3] UMR TETIS : Territoires, Environnement, Télédétection et Information Spatiale.
[4] CNES : Centre National d'Etudes Spatiales

reformulation et d'expansion de requêtes existent depuis longtemps et reposent sur différentes techniques [Baziz 03, Claveau 04]. Ces travaux s'apparentent à ceux menées sur l'expansion sémantique associée à des ressources lexicales [Voorhees 94]. Les développements méthodologiques explorés reposent sur l'exploitation des différentes relations (synonymie, hyponymie, hypernomie) entre les termes d'un thesaurus ou d'une ontologie [20, 30]. En effet, le fait de s'appuyer sur un thesaurus qu'il soit thématique ou toponymique offre la possibilité d'utiliser les liens entre termes pour reformuler une requête. Le principe général est le suivant : le moteur de recherche complète ou affine la requête de l'utilisateur en ajoutant ou en remplaçant des mots clés du thesaurus en naviguant sur les relations. Différentes stratégies et différents types d'expansion de requêtes ont été envisagées : thématique, spatiale ou croisée ; automatique ou interactive. Le choix de la stratégie est alors guidée par la réponse du moteur de recherche pour reformuler la requête soit en enrichissant la requête par des termes plus généraux (cas du moteur « muet ») soit en filtrant sur des termes plus spécifiques (cas du moteur « bavard »). L'expansion spatiale quant à elle s'appuie sur l'algèbre d'Egenhofer [Egenhofer 95] qui formalise les relations topologiques entre entités spatiales. Les résultats de cette recherche ont été prototypés puis opérationnalisés dans l'outil MDweb [21] pour être utilisés dans le service de catalogage régional SYSCOLAG (cf. section 2.4).

- ***Aide à la formulation des requêtes utilisateurs*** : Cette approche part du postulat suivant : la formulation d'une requête adressée par un utilisateur à un moteur de recherche est constituée de divers critères dont le principal est le «Quoi ? ». Une des pistes explorées est l'appui de la représentation visuelle d'un référentiel terminologique (thésaurus) comme outil de découverte et de sélection du vocabulaire spécifique à un domaine. Dans ces travaux, les référentiels terminologiques sont vus comme des graphes orientés et étiquetés qui peuvent être regroupés en sous-ensembles de termes (sous domaine thématique) [Boulet 11]. Les travaux ont porté sur la mise au point d'une méthodologie qui permette d'appréhender puis d'analyser la structure d'un thésaurus afin de choisir et exécuter des algorithmes de partitionnement de graphes. Les outils d'analyse relevant de la théorie des graphes [Sowa 84] et notamment ceux exploitant la notion de « petit monde » [Milgram 67] ont été mis en œuvre. Divers algorithmes de partitionnements existants [Clauset 04, Pons 07] (fastgreedy, walktrap et leading eigenvector) ont été évalués sur divers thésaurus. Un prototype, implémentant la méthodologie, assure au sein de la plateforme MDweb l'exécution des différentes étapes qui mènent au partitionnement du thésaurus en sous domaines. Un composant graphique permet de représenter le thésaurus au travers du partitionnement effectué.
- ***Adaptation des valeurs de métadonnées et système de facettes guidés par les thésaurus*** : La dernière piste explorée consiste à rendre intuitive la recherche d'images satellites, là où il est habituellement demandé de manipuler un grand nombre de critères issus de la sphère des producteurs d'images. Ces derniers ont souvent une sémantique incompréhensible pour les utilisateurs finaux. Pour dépasser ces limites, le processus de recherche à travers un mécanisme interactif de filtrage des résultats retournés, dit de « recherche à facettes » [Uddin 07] est couplé à un référentiel terminologique a été expérimenté [4,13,23]. La mise en place de relation d'équivalence au sein du référentiel entre les catégories issues de la sémantique « producteur » et celles issues des « utilisateurs » permettent durant un processus d'harmonisation des métadonnées une adaptation, voire un enrichissement des métadonnées et présenter ainsi des facettes sémantiquement «accessibles » pour l'utilisateur final. La représentation RDF associée à la terminologie nous permet d'envisager de tels mécanismes

sur des référentiels aussi internes qu'externes à notre plateforme [13, 25]. Cette approche a été intégralement implémenté sur l'infrastructure de données spatiales nationale GEOSUD et assure ainsi l'accès à des ressources numériques multi sources pour des utilisateurs non expert de la télédétection.

Méta modélisation pour les services de catalogage et de localisation de ressources environnementales (2007 – 2011).

Collaborations : LIRMM, Consortium NatureSDIplus, Geomatys SARL, Consortium GEONetCab,
Communications orales : [41, 42, 43]
Outils : MDweb 2.0
Licence : LGPL v3
Projets : FP7 NatureSDIplus, FP7 GeoNetCab
Etudiants : Sidhoum M. (30%), Legal G. (30%)

Ces travaux sont dans le prolongement logique de ceux portant sur la structuration des métadonnées et le prototypage d'un outil de catalogage initié dans les années 2000. Dans ce contexte, l'objectif a été d'améliorer la généricité de l'outil afin qu'il puisse constituer, gérer, administrer et consulter des catalogues multi-standards, multi-langues via le web. Conscients de la variété et de l'extensibilité potentielles des standards de métadonnées et afin de proposer une approche générique et modulaire dans le déploiement de services de métadonnées, la conception d'une telle plateforme a été abordée suivant les approches de méta-modélisation et de modularité introduites en génie logiciel. L'approche adoptée s'inspire de la vision préconisée par l'OMG au travers de Meta Object Facility [OMG 04] qui propose un langage standardisé qui permet de définir et contrôle la définition méta-modèles et de leur instanciation. La généricité de la plateforme repose sur celle de la base de données (métabase), deux référentiels de valeurs (thématique et spatial) qui vont intervenir dans le contrôle sémantique des métadonnées lors de la saisie puis lors de la phase de recherche. Sur ce socle, l'adaptation de l'outil à un nouveau contexte d'application consiste à instancier le métamodèle proposé en modèle de métadonnées et référentiel thématiques propres à la communauté [41].

La maturité et la pertinence de l'approche a permis d'aboutir à une plateforme générique et open source : MDweb. Elle a été distribuée sous licence GPL puis LGPL et téléchargée plusieurs milliers de fois par la communauté environnementale francophone et européenne. Elle a ainsi servi les besoins de partage de données en biodiversité, océanographie, foresterie, gestion des territoires, hydrologie, glaciologie, pastoralisme, …. Par ailleurs, ces innovations logicielles ont été par la suite transférées à la société Geomatys pour porter la plateforme MDweb au sein de la solution industrielle et open source Constellation-SDI [42,43].

2.2.2 Activités autour de l'interopérabilité des métadonnées dans le contexte du web sémantique

Profil d'application pour les données d'observation de la terre (2013 – 2016)

Collaborations : LIRMM, Consortium FP7 GeoNetCab, Consortium FP7 EOPOWER
Publications : [3, 4, 10, 13]
Outils : [L7, L8]
Projets : FP7 EOPOWER, Equipex-GEOSUD
Livrables : [L2, L3, L4, L5, L10, L12, L13]
Etudiants : Chahdi H. (50%), Al Hassouni N. (50%), Loukili A. (50%), Toulet A. (80%)

Depuis déjà quelques années, d'importants efforts ont été réalisés, pour ouvrir les données à l'ensemble des communautés d'intérêt (cf. section précédente), au niveau européen, avec notamment les directives INSPIRE[5] [Craglia 07], au niveau Mondial avec le GEOSS[6] [Christian 08], GBIF[7] et GEO BON[8] dans le domaine de la Biodiversité, par exemple. Elles ont ainsi mis à disposition des scientifiques de grands catalogues regroupant les données portant sur l'observation de la Terre, l'environnement et offrant des services en ligne pour accéder à tous ces jeux de données [Maguire 05, Friis-Christensen 07]. Malgré les efforts réalisés pour la définition d'un cadre d'intéropérabilité, souvent par le biais de standardisation des métadonnées ou l'abstraction de modèle (méta-modélisation) [Haslhofer 10], les démarches engagées sont la plupart du temps centrées autour d'une communauté. Elles s'attachent uniquement à couvrir les exigences fonctionnelles propres à cette dernière [11]. Elle rend ainsi délicate la mise en œuvre d'outils communs de découverte tel que l'on devrait l'envisager pour offrir une vision globale des jeux de données aux utilisateurs cibles. Ce constat nous amène à poser une réflexion autour du rôle facilitateur joué par les métadonnées dans ce contexte d'exploitation de grandes quantités de données et autour des démarches menant à l'interconnexion de jeux de données géospatiaux issues de diverses communautés.

En s'appuyant sur la notion de profil d'application telle que proposée par l'initiative Dublin Core [Coyle 08] dans le cadre du web sémantique, une méthodologie de construction de profil d'application a été proposé à destination des communautés d'observation de la Terre et environnementale [3]. Cette méthodologie a pour objectif de fournir des modèles ouverts, extensibles et exploitables dans le contexte du web sémantique pour assurer l'interconnexion de jeux de données géospatiaux issus de différentes communautés. Ils sont destinés, à terme, à couvrir les besoins de partage de données géospatiales, et plus précisément les besoins de découverte, de localisation, de consultation et de traitements des données à des fins d'analyse. Suite à des travaux menés dans le cadre de mastères 2 en 2013 et en 2014 (voir section 2.4.3), nous avons construit un profil d'application Dublin Core nommé Earth Observation Application Profile (EOAP) [10] et la plateforme Java JEE qui l'exploite. Il permet d'utiliser, à la fois, l'interopérabilité des standards de métadonnées et les principes de partage de données sur le web. Le profil d'application offre de plus un cadre descriptif suffisamment flexible,

[5] INSPIRE Directive : http://inspire.jrc.ec.europa.eu/

[6] Global Earth Observation System of System : http://www.earthobservations.org/

[7] Système mondial d'information sur la Biodiversité : http://www.gbif.org

[8] GEO Biodiversity Network : http://www.earthobservations.org/geobon.shtml

pour prendre en charge, de nombreuses problématiques environnementales et différents points de vue des utilisateurs [11].

La méthodologie employée s'adosse à la démarche d'architecture dirigée par les modèles qui est couramment utilisée en génie logiciel. Deux de ses principaux standards : Meta Object Facility et le langage de modélisation UML, sont utilisés. De plus, les formats de données RDF et RDF Schema sont mis à contribution. Ces langages semblent particulièrement adaptés pour l'instanciation de métadonnées dans un contexte de partage des données sur le web. Ils jouent le rôle de méta-modèle dans l'explicitation des modèles structurels associés au profil d'application. La construction du profil d'application EOAP puise également sa méthodologie dans les recommandations émises par la communauté Dublin Core autour de la notion de DCAP (Dublin Core Application Profile) et du Singapore framework [Nilsson 08]. Elle s'adosse aux modèles conceptuels : DCAM (Dublin Core Abstract Model) [Powell 07] qui explicite la notion de ressource et le modèle DSP (Description Set Profile) [Nilsson 09] qui fournit un cadre prescriptif à la construction du profil d'application. Un profil d'application est envisagé alors comme un profil d'ensemble de descriptions. Ces travaux ont été développés dans le cadre du projet Equipex-GEOSUD. Ils viennent répondre aux besoins exprimés par les acteurs publics français en matière de diffusion d'images satellitaires multi-sources et multi-résolutions. Le profil d'application et les réflexions autour de l'interopérabilité des métadonnées sont implémentés dans l'infrastructure de données spatiales GEOSUD [L1, L2, L3, L4, L8]. Le profil d'application a un rôle central. Il constitue le modèle « pivot » pour l'harmonisation des métadonnées issues des différentes sources de métadonnées. Il apporte ainsi une grande flexibilité pour assurer la gestion de nouvelles sources d'images, par exemple. Il prend en charge les principales fonctionnalités (découverte, visualisation, téléchargement, archivage, traitement) délivrées par l'infrastructure de données. De par sa généricité, le socle logiciel va être réutilisé dans divers contextes d'acquisition et de diffusion d'images multi-sources (Guyane, Réunion, Haïti).

2.2.3 Perspectives

Harmonisation des métadonnées et « crosswalks » (2016 ->)

Collaborations : FIOCRUZ[9], Institut Pasteur, UnB[10]
Publications : à venir en 2017
Projets : GAPAM Sentinela
Réalisations : Plateforme d'accès et de représentation de données hétérogènes en santé
Etudiants : Briand D. (Doctorant, 35%), El Ghallab A. (Mastère géomatique, 50%)

Les travaux menés actuellement dans le domaine de la santé et autour de la problématique de surveillance des maladies vectorielles dans un contexte transfrontalier, viennent se nourrir de la méthodologie de construction de profil d'application. A travers ces nouvelles recherches méthodologiques, menées dans le cadre de travaux de doctorat de Dominique Briand, nous souhaitons adresser les problématiques d'hétérogénéité des données épidémiologiques et environnementales par la mise en œuvre de techniques de transformation automatique de standards de métadonnées « crosswalks » [Zarazaga-Soria 03]. Ces travaux de mise en correspondance de standards visent *in fine* à proposer de nouvelles plateformes pour faciliter l'analyse de données transnationales en santé.

[9] FIOCRUZ : Fundaçao Oswaldo Cruz (Brésil) : Ciência e tecnologia em saúde (Science et technologie en santé)
[10] UnB : Universidade de Brasilia (Brésil)

Interopérabilité pour l'exécution de traitements hétérogènes et distribués (2016 ->)

Collaborations : CINES, HPC@LR, Unversity of Nottimgham, CNR-IIA
Projets : Equipex-GEOSUD, dépôt H2020 – EINFRA 22.1
Livrables : [L1, L6, L9]
Etudiants : Lin Y. (post-doctorant, 50%)

Cette dernière problématique initiée dans le cadre de la thèse Yuan Lin [Lin 11] et dans le cadre du projet Equipex-GEOSUD [L1, L6, L9] vise à remobiliser des algorithmes, modèles numériques, au sein de systèmes ouverts et distribués, pour leur réutilisation sur de nouvelles données en vue de produire de nouvelles connaissances. Comme pour les données, les métadonnées sont le support à la médiation et à la configuration des traitements [15, 16]. Nous souhaitons par ailleurs y associer des règles d'exécution qui permettront d'envisager l'adaptation du traitement à la donnée en fonction du contexte d'exécution. Les langages de formalisation de connaissances [W3C 04, Martin 04] et la formalisation de règles [Horrocks 04] seront mis à contribution. Cette approche vient compléter les méthodologies précédemment développées pour poursuivre l'implémentation d'infrastructures de recherche ouvertes et accessibles aux communautés de scientifiques en environnement. Un premier projet H2020 a été soumis cette année pour développer une telle approche.

2.3 Projets scientifiques

2.3.1 Projets soumis

- **2016** : Appel d'offres H2020 - EINFRA-22-2016 Topic: User-driven e-infrastructure innovation. Societal challenges : Health. Nom du projet: **Workflow Research e-infrastructure for Geospatial Environmental Health Studies**. Soumis le 30 mars 2016. Financement demandé : 3 M€. Résultat : non financé. Note de 13/15.

2.3.2 En tant que coordinateur

- **2006-2008 : Programme SYSCOLAG** : Opérationnalisation et application du service de métadonnées pour l'aide à la localisation d'information au sein de la base de connaissances SYSCOLAG. ». Financement Région Languedoc-Roussillon. Montant : 143 000 €.

2.3.3 En tant que responsable de lots de travail

- **2015-2017 : GAPAM-sentinela**. Guyane française – Amapá – Amazonas – Malaria : **Site sentinelle transfrontalier de l'Observatoire Climat et Santé**. Responsable Tâche R2 Tâche R2 : Représenter, partager et intégrer des données et informations hétérogènes.

- **2012-2019 : Projet Equipex-GEOSUD : Infrastructure de données spatiales pour la diffusion d'images haute et très haute résolution pour les acteurs publics français.** Appel à projets «Equipements d'Excellence » du Programme Investissements d'Avenir (2011) : Responsable WP2 : mise en place de l'infrastructure logicielle de données spatiales et de traitements (WP2). Subvention ANR pour le WP2 : 2 M€

- **2013-2015 : EOPOWER : Earth Observation for Economic Empowerment** (grant N°603500). Call FP7 « Mobilising environmental knowledge for policy and society ». Responsable du WP3 : Resource Facility. Subvention UE : 110 000 €.

- **2011 : REFLECS** : **Outil de référencement des missions et programmes scientifiques du CNES**. Projet R&T du CNES. Expert auprès de la société Géomatys pour réaliser l'étude et les spécifications de l'apport des référentiels terminologique et spatial pour le service de catalogage REFLECS.

- **2009-2013 : GEONetCab : GEO Network for Capacity Building** (grant N°244172). Call FP7 ENV-2009-4.1.4.1: Action in the domain of Earth Observation to support Capacity Building in GEO. Responsable WP3 : Connecting and Building. Conception et développement d'une infrastructure de découverte et d'accès aux ressources de renforcement de capacités dans le domaine de l'observation de la terre. Subvention UE : 180 000 €

- **2007-2013 : CARTAM-SAT : Cartographie dynamique des Territoires Amazoniens: des Satellites aux Acteurs**. Programme de recherche - développement technologique dans la Région Guyane-Amazonie. Responsable du Work Package Interface homme-machine et plate-forme interopérable de transfert de contenus cartographiques et thématiques. Subvention : *non retrouvée*

- **2008-2011 : NatureSDI+ : Best Pratice Network for SDI in Nature Conservation** (FP7 call EContentplus). Leader de la tâche T3.2 portant sur la spécification d'un profil de métadonnées sur les données de conservation de la nature, participation à la spécification de l'architecture et des services de l'infrastructure spatiale du projet et implémentation des services interopérables français. Subvention UE : 130 000 €.

- **2002 – 2005** : **PADOUE**. Projet ANR – Masse de données. Conception et prototypage de composants de médiation et d'intégration pour les systèmes environnementaux. En charge de la tâche : gestion des métadonnées. Subvention : *non retrouvée*

- **2000-2005** : **ROSELT/OSS** (Réseau d'Observatoires pour le Suivi Ecologique à Long Terme). Projet FFEM, IRD, MEDD, MAE. Coordination de la conception, du développement et mise en place du système d'information ROSELT. Subvention : *non retrouvée*

2.4 Responsabilités collectives

2.4.1 Animations scientifiques

- **Conférence INFORSID** (INFormatique des ORganisation et Systèmes d'Information et de Décision)
    - **2011** : Membre du comité de programme, session Innovations en Systèmes d'information pour l'environnement
    - **2012** : Membre comité de programme, membre du comité d'organisation

- **2012** : co-organisateur de la session Innovations en Systèmes d'information pour l'environnement

- **2013** : co-organisateur de la session Innovations en Systèmes d'information et de connaissances spatio-temporelles

- **Conférence SAGEO** (Spatial Analysis and Geomatics) :

  - **2001** : membre comité d'organisation

  - **2008** : membre comité de programme et d'organisation

  - **2008** : organisateur de l'Atelier Temps et Espace

- **Axe Ontologie**, UMR ESPACE-DEV (piloté avec Isabelle Mougenot, 2015-2016)

2.4.2 Encadrement de mastère 2

Synthèse des encadrements d'étudiants
- Mastère 2 Informatique, spécialité Géomatique : 3
- Mastère professionnel spécialité Informatique : 3
- Mastère recherche spécialité Informatique : 6

- Encadrement du stage de **DESS IAO** de l'Université de Montpellier : **développement des interfaces de consultation et de saisie pour une application d'échange des informations scientifique. Application au Projet ROSELT**. Nicolas Moyroud. 2001

- Co-encadrement du stage DESS IAO de de l'Université de Montpellier : **développement d'un prototype de SIEL (Système d'Information sur l'Environnement Local) – Programme ROSELT /OSS**. Didier Leibovici. 2002

- Encadrement stage de **Mastère Pro en Informatique** de l'Université de Montpellier : **Création d'un logiciel pour l'extraction automatique de métadonnées géographiques**. Baranov Alexandre. 2006

- Co-encadrement du stage de **Mastère recherche en Informatique** de l'université de Montpellier : **Exploitation sémantique d'un service de catalogage**. Paul Boisson. 2006

- Co-encadrement du stage de **Mastère recherche en Informatique** de l'université de Montpellier : **Expansion de requêtes spatio-thématiques dans un service de catalogage**. Stéphane Clerc. 2006

- Co-encadrement du stage de **Mastère recherche en Informatique** de l'université de Montpellier : **Une nouvelle approche pour les systèmes d'aide à la décision environnementaux**. Redhouane Kissi. 2006

- Co-encadrement du stage de **Mastère Pro en informatique** de l'université de Montpellier : **Analyse et re-architecture en Java JEE d'une application web de gestion des données géographiques.** Guilhem Legal et Mehdi Sidhoum. 2007

- Co-encadrement du stage de **Master recherche en informatique** de l'université de Montpellier : **Appariement dynamique entre des ontologies métiers et des métadonnées spatio-temporelle : application au domaine environnementale**. Ahmed Benyahia. 2011

- Co-encadrement du stage de **Master recherche en informatique** de l'université de Montpellier : **Composant Sémantique pour l'amélioration de la recherche de données environnementales.** Hafida Sayah.2011

- Co-encadrement du stage de fin d'études, **Ecole navale de Brest**, spécialité Informatique : **Analyse des techniques de stockage et d'interrogation du web sémantique : Application à la médiation de métadonnées environnementales** Droz-Bartholet Albin et Chapeau Jérémy. 2011.

- Co-encadrement du stage de **Master recherche en informatique** de l'université de Montpellier : **Conception d'un modèle abstrait de métadonnées pour l'interconnexion de jeux de données géoréférencées.** Hatim Chahdi. 2013

- Co-encadrement du stage de **Mastère Pro en Informatique** de l'université de Montpellier : **Environnement web pour l'exploitation des métadonnées RDF. Application aux données issues de l'observation de la terre**. Nordine El Hassouni. 2014

- Co-encadrement du stage de **Mastère Pro en Informatique** de l'université de Montpellier : **Développement d'une application web de recherche à facettes sur les métadonnées au format RDF. Application au domaine de l'Observation de la Terre**. Jean Christophe Desconnets. 2014

- Co-encadrement du stage de **Mastère Informatique, spécialité Géomatique** de l'Université de Montpellier : **Plateforme pour l'accès et la représentation de données hétérogènes. Cas des données épidémiologique et spatiale.** Anass El Ghallab. 2016

2.4.3 Encadrement de doctorant et post-doctorant

Co-Encadrement de doctorant

- **Julien Barde (co-encadrement).** Thèse soutenue en novembre 2006 à l'école doctorale I2S, spécialité Informatique. Sujet : **Mutualisation de données et de connaissances pour la Gestion intégrée des Zones Côtières. Application au projet SYSCOLAG**.
    - Publications en tant que co-auteur : [5, 19, 44]

- **Yuan Lin (co-encadrement)**. Thèse soutenue en 2011 à l'école doctorale I2S, spécialité Informatique. Sujet : **Méthodologie et composants pour la mise en œuvre de workflows scientifiques**
    - Publications en tant que co-auteur : [15, 16]

- **Dominique Briand (35%)**. Ecole doctorale I2S, spécialité Informatique. Thèse en co-tutelle la FIOCRUZ (Brésil) démarrée en 2016, Sujet : **Interopérabilité des standards de métadonnées : mise en correspondance de standards pour faciliter l'analyse de données hétérogènes distribuées. Application dans le cadre du site sentinelle transfrontalier Guyane/Brésil consacré au paludisme**.

- **Eva Serrano (10%)**. Ecole doctorale I2S, spécialité Informatique. Thèse en co-tutelle avec UNAM (Université Autonome de Mexico, Mexique). Soutenance prévue en janvier 2017. Sujet : **Strategies for environmental data preprocessing: Application to water quality assessment of Mexican rivers.**
    - Publication en tant que co-auteur : [47]

Encadrement de post-doctorant

- **Marie Angélique Laporte**. Spécifications et prototypage d'un système à facettes pour les données d'Observation de la Terre. 2013, contrat d'environ 1 an. Projet **Equipex-GEOSUD**

- **Meriam Bayoudh.** Etude et développement d'un composant logiciel s'appuyant sur de la PLI à des fins d'évaluation et d'enrichissement d'une ontologie dans le domaine de l'Observation de la Terre. 2014, contrat d'environ 6 mois. projet **EOPOWER**

- **Yuan Lin**. Réflexions sur la formalisation de la description et l'adaptation des chaînes de traitements pour les données Observations de la Terre. 2013, contrat d'environ 6 mois. Projet **Equipex-GEOSUD**

- **Romain Boulet** : Analyse de graphes multiplexes. Application aux référentiels terminologiques en environnement. 2010-2011. Contrat d'environ un an. Projet **Cartam-Sat**.

- **Julien Barde**. Spécifications et prototypage d'un module de gestion des référentiels terminologique et spatial pour un service de catalogage: 2007. Contrat d'environ 6 mois. projet **SYSCOLAG**

2.4.4 Recrutement et encadrement de CDD dans les projets de recherche

- **Nicolas Moyroud** : Ingénieur d'étude, informaticien. 2002-2004. Projet **ROSELT/OSS**

- **Didier Leibovici** : Ingénieur de recherche, géomaticien. 2002-2004. Projet **ROSELT/OSS**

- **Bruno Granouillac** : Ingénieur d'étude, géomaticien. 2005-2007. Projet **SYSCOLAG**

- **Stéphane Clerc** : ingénieur d'étude, informaticien. 2007. Projet **SYSCOLAG**

- **Christelle Pierkot** : Ingénieur de recherche, géomaticien. 2009-2011. Projet **Cartam-Sat**

- **Dorian Ginane** : Ingénieur de recherche, géomaticien. 2008-2010. Projet FP7 **NatureSDIplus** FP7 **GEONetCab**

- **Anne Toulet** : ingénieur d'étude, informaticien. 2014. Projet FP7 **EOPOWER**.

- **Mathieu Kazmierski** : ingénieur de recherche, géomaticien. 2013-2014. Projet **Equipex-GEOSUD**

- **Benoît Hiroux** : Ingénieur d'étude, calcul intensif. 2016-2017. Projet **Equipex-GEOSUD**

2.5.4 Participation à jury de thèse

- Examinateur de la thèse de Julien Barde soutenue le 9 novembre 2006 : **Mutualisation de données et de connaissances pour la Gestion intégrée des Zones Côtières. Application au projet SYSCOLAG**.

2.4.5 Participation à des comités de suivi de thèse

- Membre du CST de la thèse de Yuan Lin (2009-2011). Université de Montpellier. Ecole doctorale I2S. Spécialité Informatique. Sujet : **Méthodologie et composants pour la mise en œuvre de workflows scientifiques.**

- Membre du CST de Eva Serrano (2015 – 2016). Université de Montpellier. Ecole doctorale I2S. Spécialité Informatique. Sujet : Strategies for environmental data preprocessing: **Application to water quality assessment of Mexican rivers.**

- Membre du CST de Mojdeh Soltan Mohammadi (2015 – 2018). Université de Montpellier. Ecole doctorale I2S. Spécialité Informatique. Sujet : **Supervision de comportements remarquables d'objets mobiles à partir de leurs trajectoires.**

2.4.6 Evaluation de projets scientifiques

- **Projets ANR** : Appel d'offres CORPUS et CE26 – Innovation

- **Projets SPIRALES** (Appel d'offres interne IRD)

2.4.7 Relecture d'articles scientifiques

- **Conférence SAGEO** (Spatial Analysis and Geomatics)

- **Conférence INFORSID** (INFormatique des ORganisation et Systèmes d'Information et de Décision)

- **Revue internationale de la géomatique**. Edition Hermès

- **International Journal of Applied Earth Observation and Geoinformation**. Edition Elsevier

3. Activités d'enseignement

3.1 Responsable de modules d'enseignement

- **Université de Montpellier : Mastère Géomatique** (environ 80ETD/an)
    - Co-responsable UE: Bases de données spatiales (FMIN 266) de 2009 à 2015
    - Co-responsable UE: Information, cartographie et web (GMIN 322) de 2011 à 2015
- **Université de Montpellier : Mastère Eau** (30 ETD)
    - Co-responsable UE: Pratiques des SIG, année 2010-2011
- **Université de la Réunion - Antanarivo** : **Mastère Territoire et Risques Naturels** (30 ETD)
    - Co-responsable UE: Bases de données spatiales, année 2010-2011

3.2 Encadrements

- **Université de Montpellier** :
    - **Master géomatique** : Encadrement de projets tutorés durant les années 2013-2014 et 2014-2015
    - **Tuteur pédagogique** pour les stages de fin d'études Master Géomatique de 2011 à 2014 : de 2 à 3 étudiants/an
    - **Examinateur ou rapporteur** dans les jurys du Mastère Informatique spécialité Géomatique
- **AgroParisTech :**
    - **Mastère SILAT** (Systèmes d'informations localisées pour l'aménagement des territoires) encadrement du micro-projet « Catalogage pour l'association SIG –LR » des Masters SILAT.
    - **Tuteur Pédagogique** pour les stages de fin d'études du Mastère SILAT
    - **Examinateur ou rapporteur** dans les jurys du Mastère SILAT

3.3 Autres implications en enseignement

- **AgroParisTech** (10 ETD/an)
    - Intervenant dans la session de formation « Administration de données localisées » (Maison de la télédétection)
    - Intervenant Cours et TP au Master SILAT dans la session « catalogage » (Maison de la télédétection)

4. Valorisation de la recherche

4.1 Développement logiciel

- **Initiateur, concepteur et chef de projet du logiciel libre MDweb** : outil de catalogage et de localisation de l'information environnementale.
- **Co-concepteur et chef de projet informatique de l'outil SIEL** (Système d'Information sur l'Environnement à l'Echelle Locale) : Outil d'analyse de spatio-temporelles environnementales pour l'évaluation de la désertification

4.2 Licences logicielles

- **2004** : Dépôt l'Agence de Protection des Programmes de l'outil MDweb et ses codes sources
- **2005** : Dépôt d'une Licence Libre CeCILL pour l'outil MDweb 1.x et mise en place d'une forge (arrêtée en 2016)
- **2006** : Dépôt l'Agence de Protection des Programmes de l'outil SIEL et ses codes sources
- **2008** : Dépôt d'une Licence Libre LGPL 3.0 pour l'outil MDweb 2.x et mise en place d'une forge (arrêtée en 2016).

4.3 Transfert de technologies

- **2009-2013** : Transfert du projet MDweb vers la société Geomatys. Ce transfert s'est appuyé sur deux conventions de recherche UMR ESPACE-DEV-Geomatys durant les années 2009-2011 et 2011-2013.
- **2009- 2011** : Agrément expert scientifique auprès du ministère de la recherche.

5. Liste des publications

Synthèse publications

- Revue internationale: 6
- Revue nationale: 4
- Chapitre d'ouvrage : 3
- Conférence internationale : 11
- Conférence nationale : 7
- Poster conférence internationale : 3
- Communications orales conférence internationale : 15

5.1 Thèse

[1] **Desconnets J.C.** (1994, Novembre) : Caractérisation et typologie des systèmes endoréiques en zone sahélienne. Degré carré de Niamey Hapex-Sahel. Doctorat 3eme cycle. 50Université de Montpellier II Sciences et Techniques du Languedoc. Ecole doctorale Mécanique, Génie Mécanique, Génie Civil.

5.2 Mastère

[2] **Desconnets J.C.** (2000, Septembre) : Développement d'un outil d'évaluation du risque sismique applicable au milieu construit. Mémoire de fin d'étude. DESS Informatique Appliquée aux Organisations. Université de Montpellier II, CNAM.

5.3 Revues internationales avec comité de lecture

[3] **Desconnets J.C.**, Mougenot I. & Chahdi H. (2017) : A methodology for effective Metadata Design in Earth observation. In *Developing Metadata Application Profiles* (pp. 65-97). IGI Global.

[4] **Desconnets J.C.,** Giuliani G., Guigoz Y., Lacroix P., Mlisa A., Noort M., Ray N., Searby N.D. (2017): GEOCAB Portal: a gateway for discovering and accessing capacity building resources in Earth Observation. International Journal of Applied Earth Observation and GeoInformation. Volume 54, February 2017, Pages 95–104.

[5] Barde, J., Edgington, D., & **Desconnets, J.C.** (2008). A Generic Approach to Manage Metadata Standards. OSGeo Journal, 3(1).

[6] Leibovici, D., Quillevere, G., & **Desconnets, J.C.** (2007). A method to classify ecoclimatic arid and semiarid zones in circum-Saharan Africa using monthly dynamics of multiple indicators. IEEE Transactions on Geoscience and Remote Sensing, 45(12), 4000-4007.

[7] Mazouni, N., Loubersac, L., Rey Valette, H., Libourel, T., Maurel, P., & **Desconnets, J.C.** (2006). Syscolag: a transdisciplinary and multi-stakeholder approach towards integrated coastal area management. An experiment in Languedoc-Roussillon (France). Vie et Milieu/Life & Environment, 56(4), 265-274.

[8] **Desconnets, J.C.**, Taupin, J. D., Lebel, T., & Leduc, C. (1997). Hydrology of the HAPEX-Sahel Central Super-Site: surface water drainage and aquifer recharge through the pool systems. Journal of Hydrology, 188, 155-178.

[9] **Desconnets, J.C.,** Vieux, B. E., Cappelaere, B., & Delclaux, F. (1996). A GIS for hydrological modelling in the semi‐arid, HAPEX‐Sahel experiment area of Niger, Africa. Transactions in GIS, 1(2), 82-94.

5.4 Conférences internationales avec comité de lecture

[10] Mougenot, I., **Desconnets, J.C.**, & Chahdi, H. (2015, Septembre). A DCAP to promote easy-to-use data for multiresolution and multitemporal satellite imagery analysis. In Proceedings of the 2015 International Conference on Dublin Core and Metadata Applications (pp. 10-19). Dublin Core Metadata Initiative.

[11] **Desconnets, J. C.**, Chahdi, H., & Mougenot, I. (2014, Novembre). Application profile for Earth Observation images. In Research Conference on Metadata and Semantics Research (pp. 68-82). Springer International Publishing.

[12] Loireau, M., Fargette, M., **Desconnets, J.C.**, Mougenot, I., & Libourel, T. (2014). Observatoire Scientifique en Appui à la GEstion du territoire (OSAGE): entre espaces, temps, milieux, sociétés et informatique. Actes de la 10ème conférence internationale annuelle Spatial Analysis and GEOmatics. Grenoble. France

[13] Kazmierski M., **Desconnets J.C.**, Guerrero B., Briand D. (2014). GEOSUD SDI : Accessing Earth Observation data collections with semantic-based services. Proceedings of the 17th AGILE Conference on Geographic Information Science, Connecting a Digital Europe through Location and Place, Castellon, Spain.

[14] Pierkot C., **Desconnets J.C.** & Libourel T. (2010, Juin) : Adaptation de la localisation des ressources à l'usage. Proceeding of 25th International Cartographic Conference, Paris. 2011.

[15] Libourel, T., Lin, Y., Mougenot, I., Pierkot, C., & **Desconnets, J. C.** (2010). A Platform Dedicated to Share and Mutualize Environmental Applications. In ICEIS (1) (pp. 50-57).

[16] Lin, Y., Pierkot, C., Mougenot, I., **Desconnets, J.C.**, & Libourel, T. (2010, Juin). A framework to assist environmental information processing. In International Conference on Enterprise Information Systems (pp. 76-89). Springer Berlin Heidelberg.

[17] Carlisle M., Green D.R., De Martino M., Albertino R ., **Desconnets J.C.**, Waver R. & Cabello M. (2010, January): INSPIRE and Nature-SDIplus: further progress towards a spatial data infrastructure (SDI) for nature conservation in the EU. Proceedings of the seventeenth annual IALE, UK Conference.

[18] **Desconnets, J.C**., Libourel T., Clerc, S., & Granouillac, B. (2007, May). Cataloguing for distribution of environmental resources. In AGILE'07: 10th International Conference on Geographic Information Science (p. 15). Aalborg University.

[19] Barde, J., Libourel T., Maurel, P., **Desconnets, J.C.,** Mazouni, N., & Loubersac, L. (2005, July). A metadata service for managing spatial resources of coastal areas. In CoastGIS'05: Defining and Building a Marine and Coastal Spatial Data Infrastructure.

[20] Boisson, P., Clerc, S., **Desconnets, J.C.**, & Libourel, T. (2006, Octobre). Using a semantic approach for a Cataloguing Service. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 1712-1722). Springer Berlin Heidelberg.

5.5 Chapitres d'ouvrage

[21] **Desconnets, J.C.** & Libourel, T. (2012) MDweb : Catalogage et localisation de ressources environnementales dans Développements logiciels en géomatique: innovations et mutualisation (eds : B. Bucher and F. Le Ber). Traité IGAT. Hermès-Lavoisier. Paris, France.

[22] **Desconnets, J.C.** & Libourel, T. (2013) MDweb : Cataloging and Locating Environmental Resources, in Innovative Software Development in GIS (eds B. Bucher and F. Le Ber), John Wiley & Sons, Inc., Hoboken, NJ, USA. doi: 10.1002/9781118561928.ch8

5.6 Revues nationales avec comité de lecture

[23] **Desconnets, J.C.**, & Kazmierski, M. (2015). Mutualiser des données spatiales et des traitements en environnement. Ingénierie des Systèmes d'Information, 20(3), 89-115.

[24] Loireau, M., Sghaier, M., Chouikhi, F., Fétoui, M., Leibovici, D. G., Debard, S., **Desconnets J.C.** & Khatra, N. B. (2015). SIEL: Integrated system for modeling and assessment of desertification risk. Revue des Sciences et Technologies de l'Information-Série ISI: Ingénierie des Systèmes d'Information, 20(3), 117-142.

[25] Loireau, M., Sghaier, M., Fétoui, M., Ba, M., Abdelrazik, M., d'Herbès, J. M., **Desconnets J.C.** & Delaître, É. (2007). Système d'information sur l'environnement à l'échelle locale (Siel) pour évaluer le risque de désertification: situations comparées circumsahariennes (réseau Roselt). Science et changements planétaires/Sécheresse, 18(4), 328-335.

[26] Loireau, M., **Desconnets, J.C.**, & d'Herbès, J. M (2005). Concepts et méthodes du SIEL-ROSELT/OSS. Collection Roselt/OSS, DS3.

5.7 Conférences nationales avec comité de lecture

[27] Hajalalaina, A. R., Hervé, D., Razafimandimby, J. P., Delaître, E., **Desconnets, J. C.,** & Libourel, T. (2015). Formalisation des chaînes de traitements de données spatiales satellitaires sur la forêt à Madagascar dans Transitions agraires au sud de Madagascar : résilience et viabilité, deux facettes de la conservation : actes du séminaire de synthèse du projet FPPSM (eds : Dominique H., Carrière S.)

[28] Kissi, R., Libourel T., & **Desconnets, J.C.** (2008, Mai). Intégration de données hétérogènes pour un SAD environnemental. In INFORSID'08: Atelier SIDE Systèmes d'Information et de Décision pour l'Environnement (p. 14).

[29] **Desconnets, J.C.**, Maurel, P., Valette, E., Libourel, T., & Tonneau, J. (2007). Excellence et innovation rurales. Outil Web de gestion des données et référentiel d'analyse de projets PER pour un développement territorial durable. Congrès joint du 47e ERSA (European Regional Science Association) et du 44e ASRDLF (Association de science régionale de langue francaise).

[30] **Desconnets, J. C.,** Libourel, T., & Clerc, S. (2007, Mai). Cataloguer pour diffuser les ressources environnementales. In INFORSID (Vol. 7, pp. 22-25).

[31] **Desconnets, J.C**., Moyroud, N., & Libourel, T. (2003, Juin). Méthodologie de mise en place d'observatoires virtuels via les métadonnées. In INFORSID (pp. 253-267).

[32] Libourel, T., **Desconnets, J. C.**, Maurel, P., Moyroud, N., & Passouant, M. (2003). Les métadonnées: pourquoi faire. Proceedings of Géoévénement, 2003.

[33] **Desconnets, J.C.,** Libourel T., Maurel, P., Miralles, A., & Passouant, M. (2001, Septembre). Proposition de structuration des métadonnées en géosciences: Spécificité de la communauté scientifique. In Journées Cassini'2001: Géomatique et espace rural (pp. 69-82).

5.6 Communications orales à des conférences internationales

Seules, les communications les plus représentatives de mes travaux de recherche sont mentionnées dans cette section.

[34] Maurel P., Faure J.F.,· Cantou J.P., **Desconnets J.C.**, Teisseire M., Mougenot I., Martignac C., Bappel E. (2015, Septembre) : The GEOSUD remote sensing data and services infrastructure. ISPRS Conference - RSDI (Remote Sensing Data Infrastructure) Workshop. 7-14 septembre 2015. La Grande Motte, France.

[35] Révillion C., Bouche D., Catry T., Padeau J., Guyard ·S., Kazmierski M., Faure J.F., · **Desconnets** J.C., Brou T., Sand A. (2015, Septembre) : SEAS-OI Spatial Data Infrastucture : A tool for the diffusion of satellite imagery in the South Western Indian Ocean. ISPRS Conference - RSDI (Remote Sensing Data Infrastructure) Workshop. 7-14 septembre 2015. La Grande Motte, France.

[36] Baghdadi N., Leroy M., Maurel P.,· Cherchali S.,· Stoll M., Faure J.F., **Desconnets J.C.**, Hagolle O., Gasperi J., Pacholczyk P. (2015, Septembre): The Theia Land Data Centre. ISPRS Conference - RSDI (Remote Sensing Data Infrastructure) Workshop. 7-14 septembre 2015. La Grande Motte, France.

[37] **Desconnets J.C** (2014, Janvier): EOPOWER Resource Facility and connection with othersresource facilities. Side Event CEOS / EOPOWER Capacity Development Resource Facility, GEO X ministerial summit. Genève

[38] **Desconnets J.C** (2014, Avril) : EOPOWER Resource Facility and connection with others resource facilities ; Capacity Building Side Event GEO Work Plan Symposium, Genève.

[39] Sutherlun J., **Desconnets J.C**., Noort M. (2014, Juin) : Mapping capacity Building activities for the use of space borne Earth Observation data : The EOPOWER/CEOS Resource facility. GLAC Conference. Paris, Juin 2014.

[40] **Desconnets J.C** : GEOCAB portal (2014, Novembre): GEO CApacity Building Portal for Earth Observation. GEO XI Plenary. Genève, Novembre 2014.

[41] **Desconnets J.C.**, Libourel T., Desruisseaux M. (2010, Mai) : Métamodélisation pour les services de catalogage et de localisation de ressources environnementales. Atelier SIDE, XXVIIIèeme Conférence INFORSID. 25-28 Mai, 2010. Marseille.

[42] **Desconnets J.C.**, Libourel T., Heurteaux V. (2009, Juillet) : Partage et mutualisation en environnement : des concepts à l'usage. L'expérience MDweb. OGRS 2009 (: International Opensource Geospatial Research Symposium), Nantes. 8-10 juillet 2009.

[43] **Desconnets J.C.**, Heurteaux V. (2008, septembre): MDweb 2.0 - A Java/JEE Metadata Catalog. Free Open Source Software For Geography (FOSS4G). Cape Town, South Africa.

[44] Barde J., Egdington D., **Desconnets J.C.** (2007, Septembre) : A generic approach to manage metadata standards. Free Open Source Software For Geography (FOSS4G), Victoria, Canada.

[45] **Desconnets J.C.** (2007, Avril) : MDweb, outil libre de catalogage et de localisation de l'information : Un composant pour les infrastructures de données spatiales. Salon du Géo-évènement 2007. Porte de Versailles. Paris.

[46] **Desconnets J.C.**, Loireau M., Leibovici D., Moyroud N., D'herbès J.M. (2003, Novembre) : Approches pour la constitution des systèmes d'information autour de la désertification dans la zone Circum-Saharienne - programme ROSELT/OSS. Conférence AfricaGIS. 3-8 Novembre, Dakar.

5.7 Session Poster à des conférences internationales

Seules, les posters les plus représentatifs de mes travaux de recherche sont mentionnés dans cette section.

[47] Serrano Balderas E.C.,Berti-Equille L., Armienta Hernandez M.A. and **Desconnets J.C.** (2016, Juillet): Water Quality Data Analytics . 8th International Congress on Environmental Modelling and Software in Toulouse, France, on July 10-14, 2016.

[48] **Desconnets J.C.,** Kazmierski K., Mazetti P. (2014, Avril): Architecture framework for Capacity building Databases Interoperability. 5th GEOBIA Conference. Thessaloniki, Greece.

[49] **Desconnets J.C**, Libourel T., Heurteaux V. (2009, Mai): Mutualisation et partage des données et des connaissances en environnement. Geoïde, Conférence annuelle scientifique. Vancouver, Canada. 27-29 Mai 2009.

6. Livrables de projets

Dans cette section, je liste les livrables de projets, des cinq dernières années, qui ont fortement contribués à mes développements méthodologiques ou qui en sont issus.

Projet Equipex-GEOSUD

- **Spécifications techniques et méthodologiques**

[L1] **Conception, développement et mise en place d'une plateforme de traitements d'images satellitaires dans un environnement de calcul haute performance**. WP2 – T2.3 : Gestion des solutions de calcul à distance. Janvier 2016. 56p.

[L2] **Cahier des Clauses Techniques Particulières : Conception, Développement et mise en place de l'infrastructure de données spatiales de l'Equipex-GEOSUD**. WP2 – T2.1 : Mise en place de l'infrastructure de données spatiales. Janvier 2014. 80p.

[L3] **Spécifications des critères pour la découverte et la consultation du catalogue d'images equipex GEOSUD**. WP2 – T2.1 : Mise en place de l'infrastructure de données spatiales. Aout 2013. 12p.

[L4] **Spécifications pour le service de thesaurus GEOSUD**. WP2 – T2.1 : Mise en place de l'infrastructure de données spatiales. Septembre 2013. 4p.

[L5] **Spécifications du modèle de métadonnées pour les images satellitaires GEOSUD**. Tâche 2.1. Juillet 2013. 35p.

[L6] **Formalisation de la description et l'adaptation des chaînes de traitements pour les données Observations de la Terre**. Tâche3.1. 2013. 14p.

- **Outils**

[L7] **Géoportail de l'infrastructure de données spatiales** de l'Equipex-GEOSUD. http://ids.equipex-geosud.fr/web/guest/catalog1 . Février 2016

[L8] **Outil de gestion et d'intégration des images satellaires multi capteurs**, http://wiki.equipex-geosud.fr/index.php/Geosud_thesaurus: http://ids.equipex-geosud.fr/constellation/ . février 2016

- **Modèles**

[L9] **Thésaurus SKOS des caractéristiques d'images et de traitements** Equipex-GEOSUD : http://wiki.equipex-geosud.fr/index.php/Geosud_thesaurus

[L10] **Profil d'application pour l'Observation de la Terre**. http://wiki.equipex-geosud.fr/index.php/Earth_Observation_Application_Profile_(EOAP)

Projet FP7 EOPOWER

[L11] D3.01: **Enhance resource facility for GEO web portal**. Mai, 2015.17p.

[L12] D3.11: **Ontology of Capacity Building Earth Application Domain**. Avril 2014. 28p.

[L13] D3.12: **Incorporating the ontology and modifying the search interfaces**. Novembre 2014. 10p.

[L14] D3.13: **Architectural Framework for a Capacity Building System of Systems**. Septembre 2013. 24p.

[L15] D3.31: **Methodological material to establish the CB resources catalogue**. Avril 2015. 11p.

[L16] D3.33: **Best Practices to reference and administrate the Capacity Building resources**. Avril 2015. 26p.

[L17] D3.34: **Training session to EO Conference**. Mai, 2015. 11p.

[L18] **Portail GEOCAB** pour la recherche et l'accès aux ressources de renforcement de capacités pour l'Observation de la Terre. http://www.geocab.org/

Projet REFLECS

[L19] Lot C. Etude ONTOLOGIE. Partie 4 : Spécifications fonctionnelles pour l'utilisation du thésaurus dans un outil de référencement. Avril 2011. 28p.

[L20] Lot D Spécifications. Spécifications fonctionnelles du module de recherche de BDMS. Mai, 2011. 37p.

[L21] Metadata specification for the NatureSDI+ project : Best pratices network for Nature Conservation in Europe. Grant Agreement N. ECP-2007-GEO-317007. Mai 2010. 50p.

7. Autres

7.1 Session de formation dans une conférence

[50] **Desconnets J.C.** et Lagarde P. (2011, Mai): Master Class Saisie pratique de métadonnées. Rencontres SIG-La-Lettre. Marne La Vallée. 18 Mai, 2011.

[51] **Desconnets J.C.** (2009, Juillet) : Atelier MDweb 2 : outil libre de catalogage et de localisation de l'information environnementale: Un composant pour les infrastructures de données spatiales. Conférence internationale OGRS (International Opensource Geospatial Research Symposium). Nantes. 8-10 juillet 2009.

8. Références bibliographiques

Il s'agit des références citées dans la section synthèse de mes travaux.

[Barde 06] Barde, J. (2006): Mutualisation de données et de connaissances pour la gestion intégrée des zones côtières. Application au projet Syscolag. Doctorat. Université Montpellier II, Ecole Doctorale Information, Structures, Systèmes.

[Bechhofer 09] Bechhofer, S. (2009). OWL: Web ontology language. In *Encyclopedia of Database Systems* (pp. 2008-2009). Springer US.

[Boulet 11] Boulet R. (2011, Octobre) : Introduction d'indices structuraux pour l'analyse de réseaux multiplexes. Application à l'analyse d'un thésaurus. MARAMI, Seconde conférence sur les Modèles et l'Analyse des Réseaux: Approches Mathématiques et Informatique. 19 au 21 octobre 2011. Grenoble.

[Berners-Lee 01] Berners-Lee, T., Hendler, J., & Lassila, O. (2001).The semantic web. Scientific American, 284 :35-43.

[Baziz 03] Baziz, M., Aussenac-Gilles, N., et Boughanem, M. (2003). Exploitation des liens sémantiques pour l'expansion de requêtes dans un système de recherche d'information. In INFORSID, pages 121-134

[Christian 08] Christian, E.J. (2008). GEOSS Architecture Principles and the GEOSS Clearinghouse. Systems Journal, IEEE, vol.2, n°.3, pp.333,337.

[Claramunt 98] Claramunt C., Coulondre S., Libourel T., 1998, Autour des méthodes orientées objet pour la conception des SIG , Revue Internationale de Géomatique, Vol. 7, n° 3-4, Hermes, pp 237-257.

[Claveau 04] Claveau, V. et Sébillot, P. (2004). Extension de requêtes par lien sémantique nom-verbe acquis sur corpus. In Proceedings of TALN 2004, pages 121-130.

[Clauset 04] Clauset A., Newman M. E. J. & Moore C. (2004) : Finding community structure in very large networks. Phys. Rev. E, 2004 ; vol.70, no.6.

[Coyle 08] Coyle K., & Baker T. (2008). *Guidelines for Dublin Core Application Profiles*. Retrieved June 2014, from http://dublincore.org/documents/profile-guidelines/index.shtml

[Craglia 07] Craglia, M. &Annoni, A. (2007). INSPIRE: An innovative approach to the development of spatial data infrastructures in Europe. Research and Theory in Advancing Spatial Data Infrastructure Concepts, 93-105.

[Egenhofer 95] Egenhofer, M. J., & Mark, D. M. (1995). Modelling conceptual neighbourhoods of topological line-region relations. *International journal of geographical information systems*, *9*(5), 555-565.

[Egenhofer 99] Egenhofer, M. (1999, October). Spatial information appliances: A next generation of geographic information systems. In *1st Brazilian workshop on geoinformatics, Campinas, Brazil*.

[Fralkes 92] Frakes, W. B., & Baeza-Yates, R. (1992). Information retrieval: data structures and algorithms.

[Friss-Christensen 07] Friss-Christensen A., Ostländer N., Lutz M., Bernard L. (2007): Designing Service Architectures for Distributed Geoprocessing: Challenges and Future Directions. Transaction in GIS, vol.11, N°6, p. 799-818.

[Haslhofer 10] Haslhofer B., Klas W. (2010): A survey of techniques for achieving metadata interoperability , *ACM Comput. Surv.*, vol. 42, n°2.

[Horrocks 04] Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., & Dean, M. (2004). SWRL: A semantic web rule language combining OWL and RuleML.*W3C Member submission*, *21*, 79.

[Lamb 01] Lamb, J. (2001). Sharing best methods and know-how for improving generation and use of metadata. *New Techniques and Technologies for Statistics and Exchange of Technology and Know-how*, 175-194.

[Lin 11] Lin Y. (2011, Décembre): Méthodologie et composants pour la mise en oeuvre de workflows scientifiques. Thèse de Doctorat. Université de Montpellier II. Ecole doctorale I2S.

[Maguire 05] Maguire D. J., Longley P. A. (2005): The emergence of geoportals and their role in spatial data infrastructures. *Computers, Environment and Urban Systems*, vol. 29, n°1, p. 3-14

[Martin 2004] Martin D., Burstein M., Hobbs J., Lassila O., McDermott D., McIlraith S., Narayanan S., Paolucci S., Parsia S., Payne T., Sirin T, Srinivasan N. & Sycara K (2004). OWL-S : Semantic Markup for Web Services. W3C, 22 November 2004.

[Miles 05] Miles A., Matthews B., Wilson B. & Brickley B. (2005) : Skos core : simple knowledge organisation for the web. In DCMI '05 : Proceedings of the 2005 international conférence on Dublin Core and metadata applications, pages 1-9. Dublin Core Metadata Initiative.

[Milgram 67] Milgram S. (1967) The small world problem. Psychology Today, 1967 ; 2.P,60-67

[Nilsson 08] Nilsson M., Baker T. & Johnston P., (2008). The Singapore Framework for Dublin Core Application Profiles. Retrieved from June 21, 2016 from http://dublincore.org/documents/singapore-framework/

[Nilsson 09] Nilsson, M., Miles, A. J., Johnston, P., & Enoksson, F. (2009). Formalizing Dublin Core Application Profiles–Description Set Profiles and Graph Constraints. *In Metadata and Semantics* (pp. 101-111). Springer US.

[ISO, 03] ISO (2003): Geographic Information Metadata, ISO 19115, International Organization for Standardization, Genève, Suisse, 2003.

[ISO 04] ISO (2004): 19106:2004. Geographic information — Profiles. International Organization for Standardization, Genève. Suisse

[OMG 04] OMG, (2004). Meta Object Facility (MOF) Core Specification OMG Available Specification Version 2.0. Retrieved June, 21, 2016 from http://doc.omg.org/formal/2006-01-01.pdf

[OMG 15] OMG, (2015). OMG Unified Modeling Language TM (OMG UML). Version 2.5. Retrieved June 21, 2016 from http://www.omg.org/spec/UML/2.5/PDF

[Pons 07] Pons P. (2007), Détection de communautés dans les grands graphes de terrain. PhD thesis.

[Powell 07] Powell A., Nilsson M., Naeve A., Johnston P., & Baker T., (2007). DCMI Abstract Model. DCMI Recommendation. Retrieved June 21, 2016 from http://dublincore.org/documents/abstract-model/

[Sowa 84] Sowa J.F (1984) : Conceptual graphs for a data base interface. BMJournal of Research and development, 1984.

[Uddin 07] Uddin M.N., Janecek P. (2007). Faceted classification in web information architecture: A framework for using semantic web tools. *Electronic Library*, vol. 25, n°2.

[Voorhees 94] Voorhees, E. M. (1994). Query expansion using lexical semantic relations. In SIGIR '94 : Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, pages 61-69.

[W3C 04] W3C OWL Working Group. OWL 2 : Web Ontology Language.W3C, February 2004.

[Zarazaga-Soria 03] Zarazaga-Soria, F. J., Torres, M. P., Nogueras-Iso, J., Lacasta, J., & Cantán, O. (2003, June). Integrating geographic and non-geographic data search services using metadata crosswalks. In *Proceedings of the 9th EC-GI&GIS Workshop: ESDI: Serving the User*.

Mémoire

Recherche d'informations spatio-temporelles : Application aux images satellitaires

1. Introduction

Contexte

Les images satellites sont devenues une source d'information indispensable pour aborder et analyser les problématiques environnementales, de manière rapide, répétée et fiable. La diversité des capteurs d'observation de la Terre, ainsi que leur nombre en constante augmentation, permettent actuellement d'envisager ces données comme une manne informationnelle sans précédent, à même d'apporter de nouveaux éclairages, par exemple, dans l'étude des dynamiques des écosystèmes, ou encore dans la détection et le suivi de changements d'occupation du sol, à des échelles de temps et d'espace très fines. Pour un utilisateur, le défi est alors de pouvoir découvrir, identifier, accéder puis intégrer à bon escient, l'information provenant de séries d'images multi-temporelles et multi-résolutions.

Si les outils et les méthodologies relatives au traitement des images satellites deviennent de plus en plus à la portée de la communauté scientifique environnementale notamment grâce à la mutualisation des ressources de calcul, à l'apparition d'outils open source et au renforcement de capacités ; il n'en va pas de même pour la découverte des données issues de l'observation de la Terre qui relève encore d'un parcours émaillé de multiples obstacles.

Ces obstacles tiennent, en premier lieu, à la grande diversité des capteurs et des images qui en sont issues. Dans son étude portant sur le dimensionnement d'une antenne de réception en Haïti (Faure et Caminade, 15) dénombre plus de 70 plateformes satellites opérationnelles qui répondent potentiellement à des besoins de suivi environnemental. A titre d'exemple, pour un capteur donné, une observation peut être déclinée en divers produits qui offriront des images de qualités diverses en relation avec leur précision radiométrique, géométrie ou temporelle. Par ailleurs, chaque opérateur qu'il soit industriel ou public appuie ses processus de production et de diffusion des images sur des spécifications qui sont propres à ses contraintes opérationnelles et/ou commerciales. Cela a pour conséquence de fournir des jeux de données très hétérogènes autant en ce qui concerne les formats de diffusion que les schémas et les nomenclatures sur lesquelles reposent leur description. Il est alors délicat pour un utilisateur d'appréhender cette diversité, d'en découvrir et d'en exploiter la richesse pour sélectionner les images satellites appropriées à son questionnement thématique. Dans ce contexte, il convient alors de faciliter la découverte de tels jeux de données à tout utilisateur, expert ou non, en télédétection.

Etat de l'art

Depuis déjà de nombreuses années, des initiatives émanant des communautés de l'observation de la Terre ou de la communauté environnementale sont venues, en partie, pallier ces difficultés. Elles ont porté sur la définition et la mise en œuvre d'un cadre opérationnel d'échanges des données d'observation de la Terre. Ces systèmes de partage, également appelés infrastructures de données spatiales (IDS), s'appuient sur l'adoption et l'implémentation de standards définissant, entre autres, les schémas de métadonnées ainsi que les méthodes (API) pour les interroger. Ces préconisations constituent le cadre sur lequel repose l'interopérabilité des données. La découverte de jeux de données hétérogènes est rendue possible grâce à un service de catalogage qui apporte une vue unifiée de ces différentes ressources à des fins de recherche et de consultation. D'autres services standards assurent

l'accès à la donnée (WCS[11], WFS[12]…) ou son analyse (WPS[13]) de manière interopérable. Les infrastructures de données GEOSUD, THEIA en France, INSPIRE, COPERNICUS à l'échelle européenne ou encore GEOSS à l'échelle mondiale en sont les réalisations les plus emblématiques. Malgré ces efforts de standardisation, l'efficacité et la précision des recherches sur de grands ensembles de données spatiales en général, et de produits d'observation de la terre en particulier demeurent encore un grand défi à relever.

(Gui et *al*., 13 ; Santoro et *al*., 12) relèvent les principales difficultés qui sont à l'origine de ce constat. Nous les abordons comme la conjonction de quatre principaux facteurs : l'ambiguïté sémantique introduite lors de la production des métadonnées, les lacunes des modèles de recherche utilisés, la difficulté pour les utilisateurs de construire des requêtes conformes aux modèles de recherche d'information utilisés et enfin l'interaction limitée entre la formulation de requêtes et les résultats retournés par des applications de recherche actuelles.

Verrous

Ambiguïté sémantique dans les métadonnées

Les informations (métadonnées) sur lesquelles s'appuient les moteurs de recherche introduisent des ambiguïtés sémantiques. Elles sont inhérentes aux diverses terminologies utilisées. Dans ce contexte, un terme peut avoir plusieurs significations ou un concept peut être porté par plusieurs termes. Ces deux cas de figure, connus sous le nom de polysémie pour le premier et synonymie pour le second, sont courants et induisent dans la recherche d'information le retour de résultats avec un rappel très faible. Par ailleurs, les données telles que les images satellitaires dont la production de métadonnées sont aux mains d'industriels sont décrites en utilisant des nomenclatures qui sont peu ou pas partagés avec la communauté d'utilisateurs. Ils font rarement l'objet d'une standardisation. De fait, ces nomenclatures auront du mal à être utilisées pour orienter une recherche d'information. A ce titre, on peut parler de fossé sémantique si l'on considère la distance qui sépare la représentation sémantique d'une donnée du point de vue de l'utilisateur de celle proposée par le producteur.

Modèles de recherche limités

Parallèlement et dans leur grande majorité, les systèmes de recherche sur lesquels s'appuient les services de découverte des infrastructures sont construits sur une approche « classique » de recherche d'information. Elle s'appuie sur la combinaison de deux stratégies de recherche : l'approche booléenne ou modèle d'appariement lexical strict (Manning et *al*., 08) qui peut être combinée à une indexation de type vectorielle (Salton et *al*., 75). Cette dernière associe la décomposition des termes représentatifs d'un document et d'une requête dans un espace vectoriel afin de produire une mesure de similarité entre les termes de ces deux composantes. Elle offre ainsi une mesure de la pertinence d'un résultat. Les principaux outils open source actuels, comme Apache-Lucene[14], Apache-Solr[15] ou encore

[11] WCS (Web Coverage Service) est une spécification de l'Open Geospatial Consortium qui propose des opérations standard d'accès aux données matricielles

[12] WFS (Web Feature Service) est une spécification de l'Open Geospatial Consortium qui propose des opérations standard d'accès aux données vectorielles

[13] WPS (Web Processing Service) est une spécification de l'Open Geospatial Consortium qui propose des opérations standard d'accès aux traitements sur les données spatiales

[14] Apache Lucene est une bibliothèque open source écrite en Java qui permet d'indexer et de chercher du texte : https://lucene.apache.org/

[15] Apache Solr est une plateforme logicielle de moteur de recherche s'appuyant sur la bibliothèque de recherche Apache Lucene : http://lucene.apache.org/solr/

ElasticSearch[16] implémentent ces deux modèles et les ont mis, ces dernières années, à portée de la communauté environnementale. Leur utilisation s'est traduite par une approche de recherche qui s'affranchit d'une formalisation (booléenne) de la requête pour proposer à l'utilisateur de formuler sa requête de manière non structurée : libre. Elle pourra être associée à une expression booléenne. On parle de recherche plein texte ou *full text search*. La recherche plein texte est sans aucun doute plus simple pour l'utilisateur mais ne constitue plus un bon filtre sur les données recherchées. Par conséquence, elle produit un grand nombre de résultats. La classification des résultats par niveau de pertinence vient pallier ces limites. Du fait de l'ambiguïté intrinsèque du langage naturel, elles ne permettent pas d'assurer de manière concomitante une forte précision[17] et un rappel[18] élevé sur les résultats retournés (Tunkelang, 09). De fait, la précision et le rappel des résultats retournés sont fortement dépendants de la précision sémantique des termes sur lesquelles reposent les descriptions d'une part, et de la concomitance de la sémantique employée par l'utilisateur pour construire ses requêtes avec celui du producteur de données, d'autre part. Dans notre contexte, ces deux conditions sont rarement réunies.

Formulation de requêtes et IHM

Enfin, l'utilisation de ces approches « classiques » induit pour les utilisateurs des difficultés dans la formulation de leur requête. Ces difficultés résident d'une part dans l'usage des opérateurs logiques (AND, OR) et de leur connotation qui peut être source de confusion (Boubekeur, 08). Leur connotation dans le langage naturel est l'inverse de la signification des opérateurs logiques. Elles sont également associées à la complexité des formulaires proposés par les applications de recherche. Dans ces dernières, l'utilisateur doit, en aveugle, paramétrer une requête en utilisant plusieurs critères *Quoi ? Quand ? Où ?* et *Qui ?* sans pour autant connaître la pertinence de sa requête vis à vis de la collection de données interrogées. La plupart du temps, il va se retrouver face à un moteur de recherche « muet » ou au contraire trop « bavard ». D'une manière générale, ces approches guident faiblement l'utilisateur dans le choix des termes potentiellement pertinents pour formuler une requête. Enfin, elles n'offrent pas d'interaction entre la phase de formulation de la requête et l'obtention des résultats de la requête, interaction qui permettrait par une formulation successive d'aboutir à des résultats plus précis et moins nombreux.

Ces limites, inhérentes autant au modèle de recherche qu'à la « rudesse » cognitive des interfaces Homme Machine (IHM), sont d'autant plus prégnantes dans la recherche de données spatiales. En effet, dans ce type de recherche, en l'absence d'informations sur le contenu de la donnée, comme cela peut être le cas pour une image satellitaire, le critère *Où ?* revêt un intérêt tout particulier. Il constitue la plupart du temps, en association avec le critère temporel *Quand ?* le premier critère utilisé et le plus discriminant. Le choix puis la saisie de l'emprise spatiale devient alors déterminante dans la pertinence de la recherche. Encore aujourd'hui, l'essentiel des applications de recherche de données géographiques demande la saisie manuelle ou propose un composant cartographique qui permet de déterminer une emprise rectangulaire. Cela est loin d'être satisfaisant car cela s'avère fastidieux et source d'erreur. Cette saisie vient compliquer la formulation déjà délicate d'une requête. Enfin, les requêtes spatiales reposant sur l'interrogation de l'empreinte spatiale (sa géométrie) peuvent s'avérer coûteuses en temps car elles font appel à des opérateurs topologiques (intersection, inclusion,

[16] ElasticSearch est un moteur de recherche open source et commercial www.elastic.co

[17] Précision. Cette notion est définie en section 2.2.1 Définitions et principes

[18] Rappel : Cette notion est définie en section 2.2.1 Définitions et principes

voisinage) dont la complexité de calcul peut être élevée. Les temps de réponse à de telles requêtes peuvent devenir rédhibitoires pour un utilisateur.

Améliorations des approches de recherche

De nombreux travaux sont venus améliorer ces approches « classiques » soient en travaillant sur l'amélioration de la mesure de similarité entre documents et requêtes comme les diverses approches probabilistes comme celles proposées par (Roberston et Spark, 76) ou encore le modèle de langue proposé par (Ponte et *al*., 98). D'autres ont porté sur l'amélioration du modèle d'indexation vectoriel de Salton. Il s'agit de construire une indexation tenant compte de la sémantique portée par les termes. Cette approche permet de traiter les problèmes de pertinence liés à la synonymie ou la polysémie introduites dans les descriptions. C'est l'objet des propositions autour du modèle d'indexation à sémantique latente (Deerwester et *al*., 90). De nombreux travaux portent également sur la reformulation de requêtes. Elle consiste à partir d'une requête initiale formulée par l'utilisateur, à construire une requête qui répond mieux à son besoin, soit par réinjection de pertinence (Rocchio, 71 ; Buckley et *al*., 94) soit en s'appuyant sur des ressources termino-ontologiques telles que les thésaurus ou les ontologies (Moldovan, 00 ; Baziz et *al*., 03).

Les approches de reformulation ont également fait l'objet de nombreux travaux et implémentations dans notre domaine (Lutz et Klien, 06 ; Shvaiko et *al*., 10 ; Santoro et *al*., 12, Gui et *al*., 13). Elles ont permis notamment de prolonger et de coupler les expansions basées sur des ressources terminologiques à des référentiels spatiaux (Boisson et *al*., 06) afin de reformuler sur le contenu de la donnée (le critère Quoi ?) mais également sur les objets d'intérêt. De même, des travaux récents (Barragáns-Martínez, et *al*., 10) ont exploré l'utilisation des folksonymies venant des réseaux sociaux pour apporter des recommandations aux utilisateurs ou désambiguïser les termes en les mettant en correspondance avec des classifications existantes (Vilches-Blázquez et Corcho, 09). Ces diverses approches utilisant les ressources du web 2.0 ont été implémentées dans le contexte d'accès aux produits d'observation de la Terre (Vaccari et *al*., 12). Elles viennent se heurter aux pratiques et à l'expertise des utilisateurs qui ne sont pas forcément enclins à s'investir dans la reformulation de requêtes ou dans l'amélioration *a posteriori* de l'annotation d'une ressource. Elles se heurtent également au rôle des fournisseurs de données qui sont en usage dans le domaine du spatial. Ce sont généralement eux qui structurent, peuplent et gèrent les catalogues d'images. Elles posent également le problème du contrôle de la qualité de ces annotations (Stvilia, et *al*., 08).

Navigation et recherche à facettes

Enfin, d'autres travaux, qui sont particulièrement proches de nos préoccupations ont porté leur effort sur l'interactivité des interfaces graphiques de recherche. Un des objectifs est de guider l'utilisateur dans la formulation de sa requête puis dans l'exploration des résultats. Il s'agit également de remédier à la frustration et aux difficultés qui peuvent résulter d'une recherche classique et mettre ainsi à portée une recherche d'information dans un domaine non maitrisé (Hearst et *al*., 02). Les travaux autour des requêtes dynamiques (Shneiderman, 94) et de la recherche orientée vue (Ahlberg et Shneiderman, 94), *view-based search*, ont ouvert la voie à cette nouvelle méthodologie d'accès à l'information. Ces travaux ont donné naissance à ce que l'on appelle aujourd'hui la navigation à facettes ou recherche à facettes *faceted search* (Yee et *al*., 2003). Ces approches sont aujourd'hui largement utilisées sur le web (Atkisson, 05) par les sites de vente en ligne (e.g Amazon, Ebay) ainsi que par le monde académique (Labarre, 04). Basée sur l'accès à l'information à travers une classification par facettes (Ranganthan, 67) qui correspond à autant de dimensions discriminantes

d'une collection de données, la navigation par facettes guide l'utilisateur grâce à la possibilité d'une sélection incrémentale. Elle permet ainsi d'envisager une construction progressive de la requête. La simultanéité assurée entre la sélection des facettes et les résultats correspondants permet de parvenir progressivement à la sélection des jeux de données recherchés et d'en assurer également la découverte. Plus qu'une stratégie de recherche d'information, l'utilisation d'une recherche facettes au sein des applications de recherche apporte une méthodologie d'accès à l'information qui combine à la fois les techniques de recherche d'information, tire partie des représentations partagées de la connaissance par l'utilisation de systèmes de classification. Elle prend également en considération l'importance de l'interactivité nécessaire aux utilisateurs, via les IHM, pour améliorer leur expérience dans le processus de découverte et d'accès à l'information dont ils ont besoin.

Objectifs

Les travaux que nous présentons dans ce mémoire s'attachent à réutiliser et adapter cette méthodologie d'accès à l'information dans le contexte de la recherche d'information à destination des acteurs du suivi de l'environnement. En effet, elle semble bien adaptée pour relever le défi de la découverte d'information dans de grands ensembles de jeux de données hétérogènes et distribués dont le public cible est tout autant hétérogène en terme d'expertise. Je m'attacherai plus particulièrement à proposer l'adaptation de cette méthodologie et sa mise en œuvre dans le contexte des catalogues d'images satellitaires. Ce contexte offre un cadre qui permet d'utiliser pleinement le potentiel de la recherche à facettes. Et cela, à plusieurs titres: la description d'une image ne proposant pas d'information sur la nature de son contenu, son choix repose essentiellement sur l'analyse de ses caractéristiques et son contexte d'acquisition dont la sémantique n'est pas toujours à portée de tous les utilisateurs. Une grande majorité d'applications de recherche d'images reposent encore aujourd'hui sur une vision experte du domaine de la télédétection. Elles rendent pour cela délicat le processus de recherche et de découverte d'images. Enfin, issues d'un processus de production industriel, les images sont toujours accompagnées d'une description structurée (métadonnées) qui peut être traitée, au sein des architectures existantes, de manière automatique en vue d'élaborer une indexation adaptée à nos préoccupations. Si le sujet traité dans ce mémoire ne permet d'offrir un panorama de nos contributions passées et actuelles, elle a malgré tout l'intérêt de remobiliser et de tirer partie des travaux menés précédemment autour de l'accès à l'information puis ceux réalisés plus récemment autour de l'interopérabilité des métadonnées.

Orienté vers l'utilisateur, l'objectif des travaux est de rendre les processus de découverte et de sélection des images satellites sémantiquement et ergonomiquement « accessibles » aux utilisateurs potentiels (acteurs publics, scientifiques). Cet objectif est doublé d'une volonté de proposer une méthodologie qui puisse être mise en œuvre dans un contexte opérationnel de découverte dans des catalogues de plusieurs centaines de milliers d'images. Ces travaux empruntent les méthodes et les outils des domaines de la recherche d'information (RI), ceux de la représentation des connaissances et enfin ceux qui étudient les représentations visuelles des données.

En s'appuyant sur les classifications partagées au sein du web sémantique, il s'agit de traiter les hétérogénéités sémantiques présentes dans les métadonnées d'images, de réduire également le fossé sémantique que véhiculent les nomenclatures utilisées par les industriels. Enfin, il s'agit d'enrichir les descriptions par du contenu issues de classifications toponymiques, environnementales afin d'offrir de nouvelles dimensions de recherche (facettes) faciles d'appropriation pour les utilisateurs. L'adaptation de la recherche à facettes dans le domaine de l'imagerie satellitaire demande donc de porter la réflexion d'une telle stratégie, initiée dans le domaine des bases de données documentaires, vers le

domaine de l'information spatio-temporelle. Pour cela, il s'agit de prendre en compte à la fois les spécificités de l'information spatio-temporelle, son cycle de vie et celui de ses métadonnées, les pratiques des utilisateurs et les architectures dans lesquelles nous pouvons mettre en œuvre une telle approche.

Plan du manuscrit

Après avoir rappelé les spécificités des données spatiales, de leurs métadonnées et montrer de quelle manière leur hétérogénéité intrinsèque n'est pas compatible avec le besoin de découverte d'information des utilisateurs, les principes de la recherche d'information seront présentés et discutés. Les travaux propres à la communauté environnementale seront également rappelés. La méthodologie d'accès à l'information que constituent la navigation et la recherche à facettes sera analysée aux regards de nos préoccupations. La méthodologie visant à enrichir, adapter les métadonnées d'images en vue de proposer un environnement de découverte et d'exploration des catalogues d'images sémantiquement et ergonomiquement accessibles sera présentée. Elle consiste principalement à proposer une indexation guidée par les ressources terminologiques et spatiales sur laquelle viendra s'appuyer une navigation par facettes. Nous illustrerons la faisabilité et la pertinence de nos propositions en présentant leur mise en oeuvre dans le contexte du projet d'envergure nationale: GEOSUD[19]. Nous conclurons ce mémoire en remettant en perspective l'apport de ces travaux vis à vis des besoins de la communauté des utilisateurs de la sphère environnementale. Des perspectives seront énoncées.

2. Etat de l'art

A la croisée de plusieurs domaines tels que la recherche d'information, l'ingénierie des connaissances, la télédétection et les infrastructures de données spatiales, les travaux présentés dans ce mémoire nécessitent, avant d'en examiner les aspects fédérateurs, de définir les bases de ces disciplines et les termes consacrés de façon à préciser notre discours. Aussi, nous introduisons dans ce chapitre les notions relatives à l'information spatio-temporelle puis nous précisons la spécificité de données issues de l'observation de la Terre, à savoir les images satellitaires. L'image satellitaire est prise ici comme un cas particulier de données spatio-temporelles avec laquelle nous envisageons d'étudier un système de recherche d'information adapté à la communauté environnementale. A cet effet, nous exposons les spécificités de l'image satellitaire en relation avec les besoins applicatifs des utilisateurs de notre communauté et les contraintes qui sont inhérentes à leur consultation au sein des infrastructures de données spatiales. Nous introduisons ensuite les notions relatives à la recherche d'information. Nous précisons les avancées en la matière pour la recherche d'information en environnement. Enfin, les notions autour de la recherche à facettes que nous souhaitons adapter à notre contexte sont présentées de manière détaillée. Les notions autour de l'organisation et la formalisation de la connaissance sont introduites au travers de la présentation des systèmes de classification à facettes. Nous aborderons de manière plus consistante ces aspects dans le chapitre suivant. Ils constituent les traits d'union entre la communauté des producteurs d'images satellitaires, les applications de recherche et les besoins des utilisateurs.

[19] GEOSUD : projet financé par le Programme Investissements d'Avenir (2011) vise à développer une infrastructure nationale de données satellitaires accessibles gratuitement par la communauté scientifique et les acteurs publics : www.equipex-geosud.fr

2.1 Spécificité de l'information spatio-temporelle en environnement

2.1.1 Généralités

Notre domaine d'application, l'environnement, est relatif aux disciplines et aux approches qui se rattachent à l'étude de l'environnement. Ces études sont rendues possibles grâce à l'acquisition puis l'analyse de données spatio-temporelles qui en représentent les différentes composantes, à savoir les dimensions sociétale, biologique, climatique et physique. Une donnée ou une information est appelée « géographique » si elle fait référence à la description d'objets, d'évènements spatialement référencés par rapport à la surface de la Terre (Livre Blanc, 98). Dans ce mémoire, la notion de donnée ou d'information spatio-temporelle est privilégiée bien qu'elle soit identique à celle de l'information géographique. Elle vient la compléter pour signifier l'importance de la dimension temporelle. Comme nous le discuterons plus tard, en recherche d'information, les dimensions spatiales et temporelles sont rarement indissociables. La figure 1 rend compte de la nature composite de l'information géographique. Elle introduit la notion d'entité, ou objet d'intérêt. Cette notion correspond à une abstraction du monde réel. On la retrouve dans les travaux de standardisation de l'information géographique proposée par l'ISO (ISO, 02a ; ISO, 05), et repris par l'OGC, sur lesquels sont construits toutes les travaux de standardisation. Ainsi, l'information géographique peut être caractérisée par un ensemble de propriétés (Figure 2.1):

- **Une ou plusieurs propriétés thématiques** qui relèvent de la nature de l'objet ou du phénomène représenté (réseau hydrographique, unité administrative, pollution atmosphérique…),
- **Une ou plusieurs propriétés spatiales** qui permettent de définir et positionner l'objet sur la surface terrestre. La donnée est alors associée à un référentiel spatial (e.g système de coordonnées, système de projection),
- **Une ou plusieurs propriétés temporelles** qui permettent de définir l'instant ou la période, rattachés à un référentiel temporel donné, pendant lesquels a été observé un objet ou un phénomène.

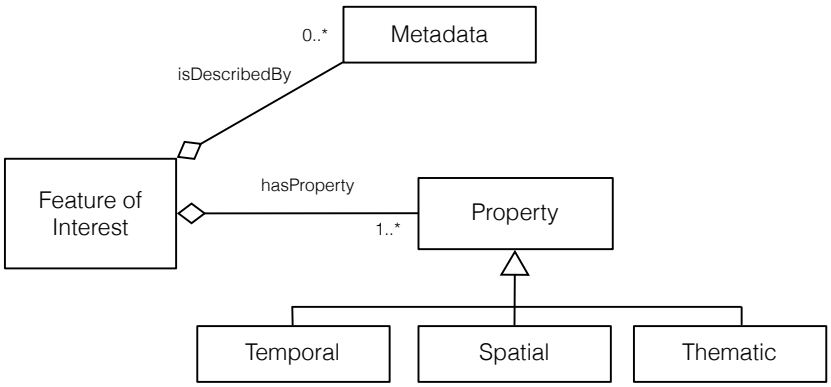


Figure 2.1 : Nature composite de l'information spatio-temporelle (Diagramme de classe UML)

Parmi ces données, on distingue des données dites " de référence " qui constituent le socle pour la représentation d'un territoire (cadastre, unité administrative, hydrographie) ou le suivi de phénomènes naturels. Elles sont généralement produites par des organismes nationaux ou internationaux spécialisés. Elles seront par la suite exploitées par des organismes de recherche ou des acteurs de la

société civile pour produire de nouvelles données dites “métier”, fruits de nouvelles acquisition et/ou de traitements métier (Desconnets et *al*., 01 ; Gayte et *al*., 97).

Ces deux ensembles de données sont toujours issus de protocoles d’acquisition variés (observation *in situ*, aéroportée, satellitaire). Ils nécessitent des phases complexes d’instrumentation (pré-traitements, enrichissement et validation avec des données *in situ*) avant de pouvoir être exploitables pour des applications en environnement, comme c’est le cas pour les données issues de l’observation de la Terre.

2.1.2 Les images satellitaires

2.1.2.1 Caractéristiques des images

La notion d’image satellitaire est associée à celle de 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. » (Pouget, 05). L’image satellitaire est donc la principale résultante du processus de télédétection. Les plateformes utilisées en télédétection sont diverses (ballons, avions et surtout satellites).

L’observation de la Terre est réalisée de manière passive ou active. La télédétection est dite passive lorsque la source illuminant la cible est le soleil. C’est le cas de la majorité des satellites utilisés à des fins de suivi environnemental comme les constellations de satellites SPOT (Satellite Pour L’Observation de la Terre), PLEIADES, LANDSAT (Land Satellite), Sentinel-2, CBERS (China-Brazil Earth Resources Satellite) ou encore Rapide Eye. Dans un moindre part mais de manière de plus en plus fréquente, la télédétection dite active est également utilisée pour aborder l’observation de la Terre dans des zones où les conditions ne permettent pas l’acquisition d’images par des capteurs passifs. C’est le cas des plateformes de type LIDAR (*LIght Detection And Ranging*), RADAR (*RAdio Detection And Ranging*). Dans ce mémoire, nous appuierons essentiellement nos travaux et leurs justifications sur la recherche d’information dans les images issues de la télédétection passive. Nous donnons ci-dessous (figure 2.2) deux exemples d’images acquises par la plateforme SPOT6.

Figure 2.2a : Images panchromatique, à gauche, et multi spectrale de la région de Montpellier acquises le 18 avril 2017 par le satellite SPOT6, à droite.

Une image est une représentation discrète d'une portion de la Terre. Une image est un tableau de pixels (Figure 2.2b) ou matrice organisée de manière régulière en L lignes et de P colonnes (image panchromatique) ou un jeu de matrices de pixels (image multi spectrale). Chaque pixel contient une valeur numérique (compte numérique ou valeur radiométrique) qui est la moyenne du rayonnement électromagnétique renvoyé par les éléments constituant la surface couverte par le pixel. Elle peut être codée entre 8 à 16 bits. Du type d'encodage choisi (8,10 ou 16 bits) dépendra la taille en octets du fichier stockant l'image. Selon les fournisseurs, l'image peut être distribuée soit sous forme de fichiers « images » où un fichier correspond à un canal (les images LANDSAT TM, par exemple) ou soit compilé dans un seul et même fichier (les images SPOT multi spectrales). Afin d'en diminuer la taille, les images sont diffusées sous différents formats. Ils permettent la compression des données : GeoTiff ou JPG2000 sont parmi les plus répandus. En fonction de la résolution spatiale, du format (donc de la compression), de l'encodage et de la quantité d'information contenue, la taille d'une image peut atteindre 50 Mo pour une image MR (Landsat TM) à 1.5 Go pour une image THR au format GEOTIFF comme une image PLEIADES panchromatique à 0,5 m de résolution spatiale. Elle n'occupe plus que 500 Mo lorsqu'elle est stockée au format JPG2000 qui utilise une transformée en ondelettes pour améliorer la compression. Enfin, à une image est rattachée à des données descriptives. Elles donnent des informations sur la structure, les conditions d'acquisition nécessaires pour pourvoir lire et traiter l'image. Il s'agit des métadonnées de l'image. Nous revenons dans le détail sur les métadonnées d'images plus bas en section 2.2.2.

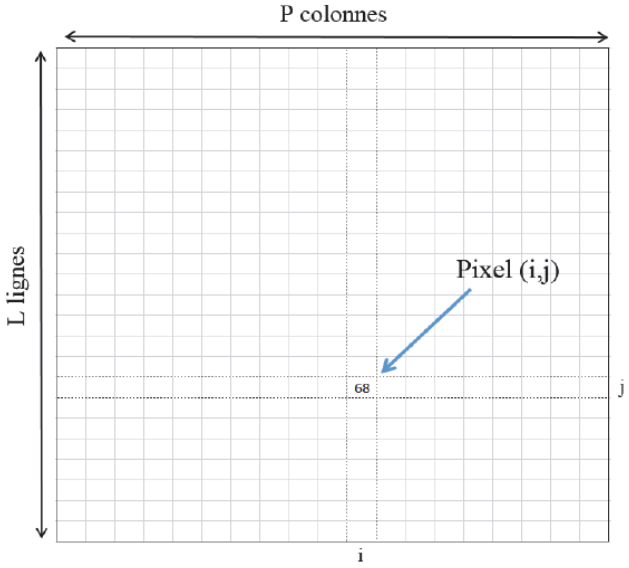


Figure 2.2b : Structure de stockage des valeurs de pixels d'une image. Le pixel (i,j) contient la valeur numérique 68 (Baghdadi et Zribi, 17).

Un instrument de télédétection, et donc l'image qui en résulte, se caractérisent par trois types de résolution :

- **La résolution spectrale.** Elle correspond aux bandes de longueurs d'onde auxquelles les instruments sont sensibles (SPOT panchromatique dans la bande visible 510-730 nm, SPOT XS canal vert 500-590 nm, canal rouge 610-680 nm et canal proche infrarouge 790-890 nm). La résolution spectrale est à mettre en relation avec la signature spectrale[20] des objets que l'on souhaite observer,

[20]Signature spectrale : La signature spectrale d'une surface est constituée de valeurs caractéristiques de sa réflectance à différentes longueurs

identifier ou quantifier. A la résolution spectrale d'un instrument correspondra une gamme d'objets perceptibles. Pour cela, la résolution spectrale d'un capteur ne peut être évaluée que dans une perspective applicative. Le choix d'un ou de plusieurs capteurs pour aborder l'analyse d'objets d'intérêt dépendra de la capacité avec laquelle les bandes spectrales, leurs largeurs, les valeurs sur lesquelles elles sont centrées permettent leur distinction.

- **La résolution spatiale** correspondant à la surface élémentaire d'échantillonnage observée par l'instrument. A une altitude donnée, un observateur ou un capteur perçoit le sol suivant un cône de visibilité (Figure 2.3). Cette surface élémentaire correspond au pixel. La surface observée dépend donc fortement de l'altitude de la plateforme. Dans le détail, chaque capteur possède son propre cône de visibilité. Il est effectivement associé aux caractéristiques du détecteur. La figure 2.3 donne l'exemple d'une acquisition d'image via un capteur à barrettes. La dimension minimale de la projection au sol d'un détecteur via le système optique du capteur correspond à la résolution spatiale. De cette résolution spatiale dépend la perception des objets. Pour pouvoir être détecté, un objet au sol doit avoir une dimension égale ou supérieure à la taille du pixel. A titre d'exemple, la résolution spatiale est de 1.5 m pour la bande panchromatique de SPOT6, 6.5 m pour les bandes XS de SPOT6, 10 m pour les bandes multi spectrales de Sentinel-2, 30 m pour LANDSAT Thematic Mapper et de 1 km pour NOAA-AVHRR (*National Oceanographic and Atmopsheric Administration, Advanced Very Hight Resolution Radiometer*). Sans qu'il y ait un réel consensus, les résolutions spatiales sont classées en plage de valeurs. On distingue quatre grandes plages de valeurs :

- **Basse résolution (BR)** correspondant à des résolutions supérieures à 200 mètres,
- **Moyenne résolution (MR)** correspondant à des résolutions supérieures à 30 mètres et inférieures à 200 mètres,
- **Haute résolution (HR)** correspondant à des résolutions supérieures à 2 mètres et inférieures à 30 mètres et enfin,
- **Très haute résolution (THR)** correspondant à des résolutions inférieures à 2 mètres.

La largeur de la bande balayée, le long d'un parcours correspondant à la trace du satellite, définit la fauchée au sol. Elle est de 60 km pour SPOT, 185 km pour Landsat TM, de 290 km pour Sentinel-2 et de 2800 km pour NOAA-AVHRR.

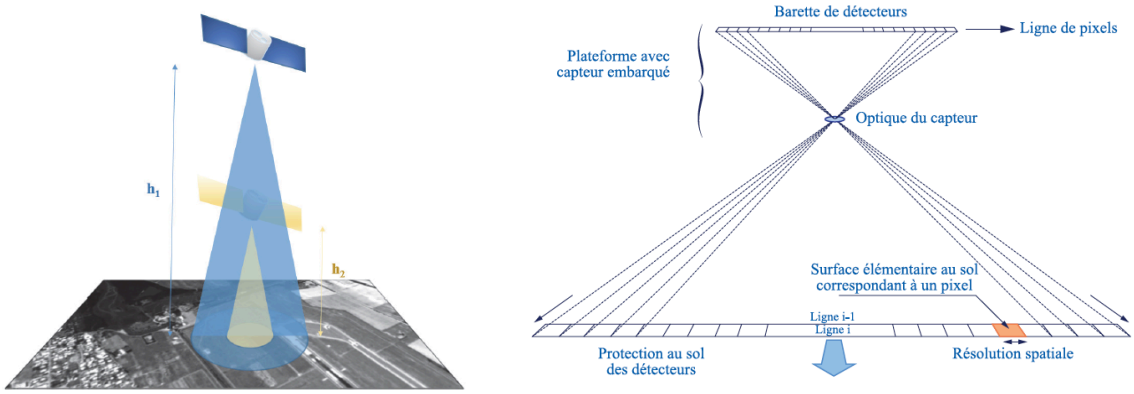


Figure 2.3 : a) cône de visibilité d'un capteur optique embarqué à deux altitudes différentes b) acquisition de l'image via un capteur à barettes (Baghdadi et Zribi, 17).

d'ondes.

- **La résolution temporelle ou répétitivité** correspond à la période entre deux acquisitions de la même portion d'espace à la surface de la Terre. Cette résolution ne dépend pas du capteur mais de l'orbite et du mode de manœuvre du satellite. De manière standard (sans programmation), la résolution temporelle du satellite SPOT est de 26 jours, 16 jours pour LANDSAT TM et 14.5 jours pour NOAA-AVHRR. La nouvelle Constellation européenne Sentinel-2A et 2B, quant elle, assurera à terme une répétitivité de 5 jours.

- **La résolution radiométrique** correspond à la sensibilité d'un capteur. Elle traduit sa capacité à distinguer au sein d'une même bande spectrale les différences dans l'énergie électromagnétique renvoyée par les surfaces élémentaires du sol. Ainsi un capteur avec une résolution radiométrique fine enregistrera un nombre d'intensité plus important. L'histogramme des valeurs de pixels d'une scène s'étalera sur d'autant plus de valeurs que sa résolution est fine.

2.1.2.2 Cycle de vie de l'image satellitaire

Nous reprenons la décomposition du cycle de vie des images satellitaires telle que proposée par le projet CARTAM-SAT[21]. Trois grandes étapes sont identifiées et mises en relation avec les problématiques technologiques, scientifiques ou informatiques :

- **Etape 1 : du satellite à l'image** : Cette étape correspond à la sollicitation des technologies d'observation (instrument, plateforme), de réception (antenne, décodeur) et de pré-traitements de la télémesure (chaîne de traitement de la télémesure). Elle vise à produire une image sous un format interprétable par un humain et utilisable par une machine avec les outils *ad hoc*. De manière concomitante, les métadonnées afférentes aux caractéristiques de l'image, à sa production sont générées à ce stade. Elles accompagnent les différents produits issus de cette étape. Dans la majorité des cas, c'est l'industriel qui a conçu et lancé la plateforme, ou ses partenaires qui conduisent les opérations destinées à produire l'image. Dans un contexte où les résolutions spatiale et temporelle sont de plus en plus faibles, les problématiques d'optimisation de traitements de grands volumes d'images surgissent à ce stade. L'objectif est alors de pouvoir mettre à disposition les images acquises et pré traitées dans des délais les plus courts. Nous noterons également qu'à ce stade, l'image ne porte pas d'information sur les objets ou les phénomènes observés. C'est l'objet de l'étape suivante. L'image est une matrice de comptes numériques ou de valeurs géophysiques. A ce stade, elle a pu être corrigée des effets de l'atmosphère (calibration optique). Elle a également pu être corrigée en géométrie afin qu'elle puisse être positionnée a minima dans un système de référence spatial planimétrique ou, c'est plus souvent le cas, elle est corrigée des déformations du relief. On parle alors d'ortho rectification. Cette dernière opération nécessite de posséder les modèles numériques de terrain et des points de contrôle au sol.
- **Etape 2 : de l'image à la connaissance** : Cette étape porte sur l'interprétation des valeurs géophysiques contenues dans l'image en vue d'en extraire la connaissance utile à la gestion d'une ressource, d'un territoire ou encore au suivi des changements. Cette interprétation fait appel à des méthodologies de traitements du signal qui assurent des opérations (segmentation, désenuagement, détection de changement) de préparation de l'image afin de rendre possible

[21]CARTAM-SAT : Cartographie dynamique des Territoires Amazoniens : des Satellites aux Acteurs. Projet par le Fond Européen de Développement Régional (FEDER).

l'étape d'interprétation à proprement parlé. C'est au cours de cette phase que le télédétecteur ou le géomaticien intervient et relie les valeurs de l'image à la connaissance experte du domaine étudié (agriculture, littoral, foresterie...) pour en produire une représentation cartographique.

- **Etape 3 : de la connaissance à l'acteur** : Cette étape correspond à la diffusion des cartographies issues de l'interprétation de l'image vers l'utilisateur final. Cette diffusion est réalisée au travers de systèmes d'information orientés vers le partage de l'information, comme les infrastructures de données spatiales. Les cartographies sont visualisables, téléchargeables à partir de services web géographiques standardisés. Certaines infrastructures proposent des services de traitements qui permettent de produire les représentations cartographiques à la demande. Les contenus diffusés par de tels systèmes peuvent être uniquement ou également les images qui ont servi à l'élaboration des produits cartographiques. C'est l'objet de nombreux projets et plateformes d'industriels de l'espace, d'initiatives nationales comme GEOSUD, ou encore d'initiatives transnationales comme Copernicus[22] et son projet phare Sentinel-2. Dans ce contexte où les initiatives et les produits sont nombreux, l'enjeu pour l'utilisateur est de parvenir à identifier, qualifier et accéder aux images ou produits en adéquation avec son besoin applicatif. Pour les concepteurs des systèmes de diffusion, l'enjeu est de faciliter l'accès à un large panel d'images et de produits cartographiques.

Cette vision scientifique du cycle de vie de l'image est intéressante à plusieurs titres. Elle permet, d'une part, de mettre en lumière, qu'au cours de son cycle de vie, l'image est entre les mains de divers acteurs (les industriels, les spécialistes de la télédétection, les géomaticiens, les acteurs du suivi de l'environnement) pour être finalement diffusée sous diverses formes : les images elles mêmes à différents niveaux de traitements ou les produits cartographiques qui sont le résultat de leur interprétation. Il est également important de retenir qu'à l'issue de leur production (étape 1), les images ne portent pas de propriétés thématiques (cf. figure 2.1) relatives aux objets d'intérêts observés mais sont caractérisées par les propriétés intrinsèques à la structure de données (propriétés de la matrice et des valeurs) ainsi qu'aux propriétés relatives aux conditions dans lesquelles elles ont été acquises (ensoleillement, angle de prise de vue, tangage, roulis, positionnement du satellite).

La problématique que nous étudions dans ce mémoire porte sur les données et les informations produites lors de la première étape du cycle de vie (du satellite à l'image) et lors de la troisième étape du cycle de vie (de la connaissance à l'acteur). Toutefois, nous nous sommes concentrés sur la recherche d'information sur les images non interprétées (issues de l'étape 1). En effet, il nous semble qu'elle recèle un verrou peu traité à ce jour : la mise à disposition d'éléments de découverte sémantiquement accessibles à des utilisateurs peu familiers de la télédétection à partir d'informations issues d'une sphère de spécialistes (les industriels de l'espace et les télédétecteurs).

2.1.2.3 Accès aux images satellitaires

Depuis déjà plusieurs années, nous pouvons faire le constat que la diffusion des images satellitaires s'est largement répandue. C'est certainement sous les effets conjugués des avancées technologiques en matière d'observation, des capacités toujours plus performantes des architectures

[22] Copernicus : connu précédemment comme le programme GMES (Global Monitoring For Environment and Security) http://www.copernicus.eu/

matérielles, des nouvelles approches méthodologiques en télédétection ou encore de la demande de plus en plus pressante de la société pour une meilleure gestion de l'environnement. Cela n'implique pas pour autant une exploitation toute aussi répandue des produits de l'observation de la Terre par la communauté environnementale. (Maurel et *al.*, 16) ont notamment bâti des initiatives, comme GEOSUD, pour que l'imagerie satellitaire soit plus à la portée des gestionnaires de l'environnement.

Aujourd'hui, les moyens et les médias de diffusion de l'imagerie satellitaires sont variés. Elles émanent pour une grande part de l'initiative des producteurs et distributeurs d'images et de plus en plus d'initiatives nationales, transnationales ou internationales. Ces dernières sont portées conjointement par les utilisateurs finaux, scientifiques et acteurs de l'environnement, en partenariat avec la sphère industrielle qui est à l'origine des investissements dans les systèmes d'observation.

Concernant les systèmes de diffusion implémentés par les producteurs d'images, nous citerons quelques uns qui permettent d'accéder à l'imagerie HR et THR comme le site de l'USGS[23] qui propose depuis plusieurs dizaines d'années l'accès gratuit aux images issues des plateformes LANDSAT : https://landsat.usgs.gov/landsat-data-access, celui de l'INPE[24] au Brésil qui propose le téléchargement des images acquises par la plateforme CBERS : http://www.dgi.inpe.br/CDSR/. En France, c'est AIRBUS Defense and Space qui diffuse les images acquises par les constellations SPOT et PLEIADES : http://www.intelligence-airbusds.com/fr/5114-parcourir-et-commander. A l'échelle européenne l'ESA, quant à elle, a mis en place un portail d'accès à l'imagerie issues des constellations Sentinel-1 et Sentinel-2 : https://scihub.copernicus.eu/dhus/#/home. Afin de répartir les charges d'accès à ces portails, des points d'accès nationaux sont également prévus[25]. La figure 2.4 présente la page de recherche des images dans le catalogue LANDSAT et CBERS.

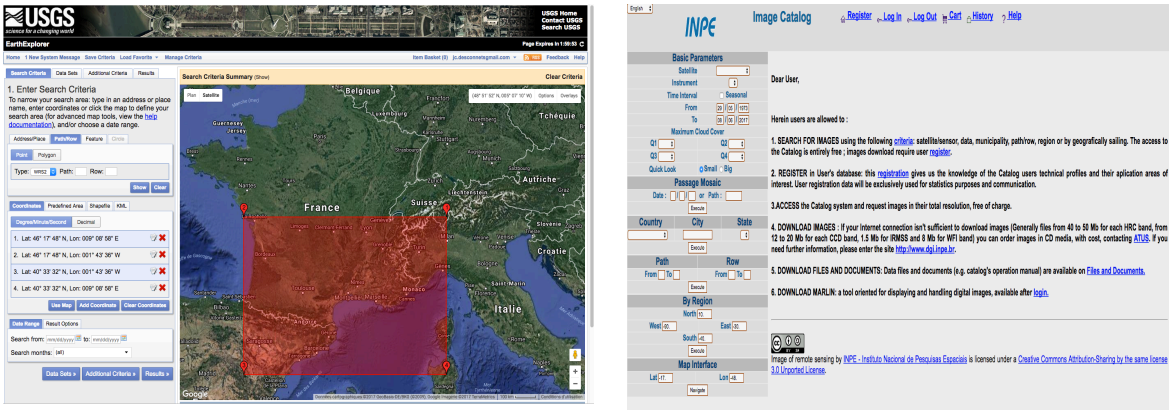


Figure 2.4 : Exemple de système de diffusion des images satellitaires par les producteurs de données. Portail d'accès aux images LANDSAT (à gauche) et portail d'accès aux images CBERS (à droite).

Les initiatives émanant des communautés d'utilisateurs sont également nombreuses. Elles se situent à différents niveaux d'organisation, aux échelles nationales comme le projet GEOSUD et THEIA auquel nous participons, à l'échelle européenne comme les initiatives expérimentales GENESI-DEC, GENESI-DR (Cossu et *al.*, 10), EuroGEOSS (Vaccari et *al.*, 12), ou plus

[23] USGS : United States Geological Survey

[24] INPE : Instituto Nacional de Pesquisas Espaciais (Institut national des études spatiales)

[25] En France, les images Sentinel-2 sont accessibles via un système de diffusion géré par le CNES : https://peps.cnes.fr/rocket/

institutionnelles comme les systèmes de diffusion rattachés au programme Copernicus, comme DIAS[26]. Enfin, il existe également des initiatives de diffusion à l'échelle internationale comme le système GEOSS[27] qui émane du GEO[28]. Ce dernier a pour objectif d'agréger l'imagerie satellitaire et les produits cartographiques pour servir les besoins en information en lien avec de grands enjeux sociétaux comme la Santé, la Biodiversité, le Climat. Des déclinaisons existent à l'échelle continentale : AFriGEOSS pour le continent Africain ou encore AmericaGEOSS pour le continent Américain.

Contrairement aux systèmes de diffusion émanant de la sphère des producteurs d'images, les systèmes de diffusion des utilisateurs ont été développés avec l'objectif de fournir un accès à des images issues de différents plateformes satellites afin de couvrir, notamment des besoins applicatifs qui demandent l'utilisation d'images présentant des résolutions spectrale, spatiale ou temporelle complémentaires. Comme nous l'avons présenté et souligné dans (Mougenot et *al*., 15) dans lequel sont présentés quelques cas de scénarii d'utilisation de l'imagerie HR et THR, les applications environnementales nécessitent souvent la combinaison d'images venant de divers capteurs. Les exemples du suivi des changements d'occupation du sol autour des agglomérations, encore appelée suivi de l'artificialisation (Dupuy et *al*., 12) ou encore les travaux sur la détection des bas fonds dans les forêts tropicales malgaches (Hajalalaina et *al*., 13), illustrent ces besoins. Ils mettent également en lumière la nécessité de pouvoir accéder sur un même système de diffusion à des produits d'observation de la Terre présentant des caractéristiques complémentaires.

De la nécessité de proposer des images d'origines diverses, ces systèmes reposent sur les principes d'interopérabilité qui assurent au moins des services de découverte interopérables. Ils restent néanmoins difficiles d'accès à des non spécialistes tant du point de vue de leur ergonomie que du point de vue de la sémantique sur laquelle l'utilisateur doit s'appuyer pour rechercher des images. Par contre, dans leur grande majorité, les systèmes de diffusion des producteurs d'images se concentrent sur les images qu'ils produisent. Ils les diffusent selon des spécifications qui leurs sont propres tant pour les formats des images que ceux des métadonnées.

2.2 Métadonnées pour l'information spatio-temporelle

2.2.1 Généralités

La nature composée de l'information (thématique, spatiale et temporelle) et la multiplicité des producteurs rendent indispensable la description détaillée de l'information spatio-temporelle tout au long de son cycle de vie et dans toutes ses composantes. Si la notion de métadonnée est préexistante au web, l'intérêt porté et son utilisation sont désormais largement répandus dans le domaine des applications en environnement. Elles sont abondamment exploitées pour assurer la découverte, le partage, la réutilisation de l'information et la gestion des données. Comme (Lamb, 2001) le propose dans sa classification, au moins trois catégories d'éléments de métadonnées sont utiles aux besoins de gestion dans les infrastructures de données spatiales:

[26] DIAS : Copernicus Data and Information Services http://www.copernicus.eu/news/upcoming-copernicus-data-and-information-access-services-dias

[27] GEOSS : Global Earth Observation System of Systems : https://www.earthobservations.org/geoss.php

[28] GEO : Group on Earth Observation : https://www.earthobservations.org/

- **Les métadonnées « descriptives»** portant sur le contenu de la donnée et ses caractéristiques techniques. Les données spatio-temporelles, quant à elles, nécessitent des éléments spécifiques pour décrire le mode de représentation (matriciel, vectoriel), le niveau de représentation des objets (échelle/résolution). Elles fournissent également des éléments qui permettent de positionner la donnée dans l'espace et dans le temps, au travers notamment la description, au moins, d'un référentiel spatial.
- **Les métadonnées « administratives »** apportant l'information sur le cycle de vie de la donnée et sa conservation pérenne. Ces métadonnées seront dédiées plus spécifiquement à la gestion et à la préservation des données,
- **Les métadonnées « structurelles »** décrivent les schémas sur lesquels sont construites les données et les relations hiérarchique et d'agrégation qui les lient.

Une quatrième catégorie que l'on appellera **métadonnées « qualitatives »** est également nécessaire pour évaluer la qualité de l'information, et donc son adéquation avec l'usage envisagé. Il s'agit notamment d'informations décrivant la généalogie, la précision (géométrique ou temporelle) ou bien encore le niveau de complétude. Enfin, il est à noter que le plus souvent, au sein des systèmes d'information géographique, les métadonnées sont séparées de la donnée d'origine. C'est le cas pour l'imagerie satellitaire. Cette séparation autorise l'accès aux métadonnées sans accéder à la donnée elle-même (Weibel et *al*., 98).

2.2.2 Métadonnées pour les images satellitaires

Chaque producteur d'images qu'il soit privé ou public s'appuie sur des spécifications pour implémenter les opérations de transformation des flux de télémesure reçus par une antenne de réception en images dans un format lisible par les outils *ad hoc*. Ces spécifications sont implémentées dans les terminaux de production et fournissent un ensemble d'informations pour assurer les corrections radiométriques, géométriques. Elles permettent de réaliser des traitements métier sur les valeurs géophysiques des images. Ces spécifications sont propres à chaque producteur d'images et à chaque famille de plateformes. Par exemple, AIRBUS Defense and Space, qui opère depuis plusieurs années les plateformes SPOT et PLEIADES, repose la production des images et des métadonnées associées sur la spécification DIMAP[29] . Elle définit l'ensemble des exigences et des formats de production des images issues des deux plateformes. Les spécifications des images acquises par la constellation Sentinel-1 et 2 sont rédigées et diffusées par l'ESA (ESA, 15), celles permettant la production d'images LANDSAT sont quant elle, produites et publiées par l'USGS (USGS, 15).

Les images sont accompagnées de métadonnées sous deux formes : dans l'entête des fichiers qui contiennent les données. On parle alors de métadonnées « internes ». Elles peuvent être séparées de la donnée. On parle alors de métadonnées « externes ». Dans les premiers, on retrouve les informations qui permettent aux logiciels SIG d'afficher à minima l'image dans un *viewer*. Nous donnons un extrait de ces métadonnées d'entête dans le listing 2.1. Elles sont extraites d'une image multi spectrale SPOT6 grâce à l'utilitaire *gdalinfo*.

[29] DIMAP : Digital Image Map

```
 [...]
Coordinate System is:
PROJCS["2154 RGF93 / Lambert-93 (FR.)",
    GEOGCS["RGF93 (FR.) [4171]",
[...]
Corner Coordinates:
Upper Left  (231096.000, 6895506.000) (3d24'49.80"W, 48d59'16.79"N)
Lower Left  (231096.000, 6830004.000) (3d20'31.30"W, 48d24' 2.75"N)
Upper Right (297432.000, 6895506.000) (2d30'34.73"W, 49d 1'58.77"N)
Lower Right (297432.000, 6830004.000) (2d26'52.43"W, 48d26'42.99"N)
Center      (264264.000, 6862755.000) (2d55'42.08"W, 48d43' 3.52"N)
[...]
```

Listing 2.1 : Extrait des entêtes du fichier GeoTIFF d'une image SPOT6 (formaté par *gdalinfo*). L'extrait fournit une partie de la description du système de projection ainsi que les coordonnées des quatre coins de l'image en projection cartographique puis en coordonnées géographiques.

Les fichiers de métadonnées associés aux images reprennent et complètent la description des images stockées dans les entêtes en apportant des informations sur les conditions d'acquisition. Elles décrivent entre autres les paramètres de prise de vue (incidence du satellite, angle avec le soleil, roulis, tangage, position du satellite), les conditions environnementales (ensoleillement, nuage,..) lors de la prise de vue, autant d'informations qui seront utiles aux pré-traitements et post-traitements de l'image. Enfin, d'autres informations décrivent la généalogie des traitements qui ont permis la transformation de la télémesure en image. Ils fournissent également une information sur son niveau de traitement. Ce dernier correspond aux corrections apportées à l'image diffusée. Il donc renseigne sur le niveau de qualité de l'image en géométrie et en radiométrie. Pour une grande part, les métadonnées sont diffusées sous format XML ou autre format texte propriétaire.

De fait, les métadonnées issues des chaînes de production présentent de fortes hétérogénéités. Nous énumérons les principales et positionnons notre analyse sur la classification proposée par (Haslhofer et Klas, 10). En effet, dans leur revue portant sur les techniques d'interopérabilité sur les métadonnées, (Haslhofer et Klas, 10) posent un panorama exhaustif de la nature des hétérogénéités possibles. Ils les expriment à travers deux axes : leur nature (structurelle, sémantique) et le niveau d'abstraction (métamodèle, schéma, instance) auxquelles ces hétérogénéités apparaissent. Sans revenir en détail sur la classification proposée, l'analyse des métadonnées des différents producteurs laisse apparaître que les hétérogénéités rencontrées sont de nature structurelle, sémantique, et cela aux différents niveaux d'abstraction. Nous les décrivons à partir de deux exemples qui présentent les métadonnées relatives à la description de l'emprise spatiale (Listing 2.2) et celles relatives à la description du niveau de qualité de l'image (Listing 2.3). Ces extraits sont issus du fichier de métadonnées qui accompagne l'image ortho rectifiée venant successivement de la plateforme LANDSAT 8, Sentinel-2, SPOT6 et Rapid Eye. Ces derniers sont représentatifs de la diversité des métadonnées que nous pouvons rencontrer.

Le listing 2.2 permet de mettre en lumière les incompatibilités structurelles qui sont rencontrées au niveau des schémas de métadonnées. Elles sont de trois types. Elles portent sur les conflits de nommage (*naming conflict*), les correspondances *n-a*ire entre éléments de schémas (*Multilateral correspondences*) et enfin l'étendue du domaine couvert par le schéma (*domain coverage*). L'élément de métadonnées décrivant la coordonnée en latitude du coin haut gauche : `CORNER_UL_LAT_PRODUCT`

(Landsat 8) et `re:topleft/re:latitude` (Rapid Eye) illustre un conflit de nommage au niveau du schéma. L'élément de métadonnées `LOWER_CORNER` (Sentinel-2) qui décrit les coordonnées en latitude et en longitude contient les informations fournit par deux éléments des schémas Rapid Eye (`re:bottomRight/re:latitude` et `re:bottomRight/re:longitude`), DIMAP (`Vertex/LON` et `Vertex/LAT`) et LANDSAT (`CORNER_LR_LAT_PRODUCT` et `CORNER_LR_LON_PRODUCT`). Ce dernier exemple illustre la correspondance *n-a*ire entre éléments de schéma. Enfin, les incompatibilités associées à l'étendue du domaine couvert par le schéma de métadonnées (*domain coverage*) sont nombreuses. Elles portent sur les différentes caractéristiques de l'image et ses conditions d'acquisition. Elles révèlent la spécificité du système d'observation (la plateforme et les instruments de mesure) utilisé. Elles seront particulièrement limitantes lors des opérations de corrections et d'interprétation des images mais auront peu d'impacts pour couvrir les besoins de découverte des images.

```
[…]
     DATUM=WGS84
[…]
    CORNER_UL_LAT_PRODUCT = 44.24600
    CORNER_UL_LON_PRODUCT = 3.46341
    CORNER_UR_LAT_PRODUCT = 44.19801
    CORNER_UR_LON_PRODUCT = 6.34258
    CORNER_LL_LAT_PRODUCT = 42.14164
    CORNER_LL_LON_PRODUCT = 3.44776
    CORNER_LR_LAT_PRODUCT = 42.09704
    CORNER_LR_LON_PRODUCT = 6.22990
[..]
```

a) Extrait du fichier de métadonnées d'une image Landsat 8

```
<Tile_Geocoding metadataLevel="Brief">
<HORIZONTAL_CS_NAME>WGS84 / UTM zone
 30N</HORIZONTAL_CS_NAME>
<HORIZONTAL_CS_CODE>EPSG:32630</HORIZONTAL_CS_CODE>
[…]
<Area_Of_Interest>
 <Bbox>
  <LOWER_CORNER>43.654 2.835</LOWER_CORNER>
  <UPPER_CORNER>45.455 4.564</UPPER_CORNER>
</Bbox>
</Area_Of_Interest>
[…]
</Tile_Geocoding>
```

b) Extrait du fichier de métadonnées d'une image Sentinel-2

```
 […]
<Dataset_Extent>
<EXTENT_TYPE>Bounding_Box</EXTENT_TYPE>
          <Vertex>
            <LON>-4.75847111077</LON>
            <LAT>48.1004504667</LAT>
<COL>1</COL>
            <ROW>1</ROW>
          </Vertex>
          <Vertex>
            <LON>-3.93738143032</LON>
            <LAT>48.1517535887</LAT>
<COL>10227</COL>
            <ROW>1</ROW>
          </Vertex>
[…]
</Dataset_Extent>
```

c) Extrait du fichier de métadonnées d'une image SPOT6

```
<re:geographicLocation>
  <re:topLeft>
    <re:latitude>44.5041</re:latitude>
    <re:longitude>4.1538/re:longitude>
</re:topLeft>
<re:topRight>
   <re:latitude>44.366771</re:latitude>
   <re:longitude>5.102924</re:longitude>
</re:topRight>
<re:bottomRight>
   <re:latitude>43.19993</re:latitude>
   <re:longitude>4.69121</re:longitude>
</re:bottomRight>
<re:bottomLeft>
   <re:latitude>43.334701</re:latitude>
   <re:longitude>3.760414</re:longitude>
</re:bottomLeft>
</re:geographicLocation>
```

d) Extrait du fichier de métadonnées d'une image Rapid Eye

Listing 2.2 : Extrait des fichiers de métadonnées de 4 images acquises par quatre plateformes. Elles décrivent l'enveloppe spatiale de l'image. Remarque : le système de référence spatial auquel sont associées les coordonnées géographiques des emprises n'est pas systématiquement explicité. C'est le cas, par exemple pour les métadonnées d'une image SPOT6 ou Rapid Eye.

Le listing 2.3, quant à lui, met en évidence les incompatibilités sémantiques au niveau des schémas et des valeurs de métadonnées. Ils sont autant de verrous pour mettre en œuvre une recherche sur des images satellitaires issues de plusieurs plateformes. Deux incompatibilités sémantiques majeures sont à considérer. La première porte sur ce que (Haslhofer et Klas, 10) appellent le conflit de représentation (*representation conflict*), et cela au niveau des valeurs de métadonnées. La seconde porte au niveau du schéma de métadonnées et concerne les ambiguïtés terminologiques des éléments de métadonnées (*terminologies mismatches*). Le fragment de métadonnées (listing 2.3) décrit le niveau de traitements de l'image, et de manière implicite sa qualité géométrique et radiométrique. Elle fait partie des informations qui servent à un utilisateur pour définir l'adéquation entre son besoin applicatif et les images qui lui sont proposées. Le conflit de représentation, que nous appellerons dans la suite du mémoire « fossé » ou « décalage » sémantique entre la vision « producteur » et « utilisateur », porte

sur les valeurs proposées par les éléments de métadonnées `DATA_TYPE`, `PROCESSING_LEVEL` ou encore `eop:productType`. En effet, pour décrire un même type de produit ou niveau de traitement, chaque producteur d'images utilise sa propre nomenclature. Quand une image ortho rectifiée LANDSAT est caractérisée par le codification `L1TP`, l'image Sentinel-2 a un niveau de traitement qui est codifié par la valeur `Level-1C`, la valeur `ORTHO` est associée à l'image SPOT6, la valeur `L3A` est quant elle associée au niveau de traitement d'une image Rapid Eye. Ces incompatibilités au niveau valeurs apportent d'une part des ambiguïtés sémantiques car elles sont toutes synonymes et d'autre part sont porteuses d'un décalage sémantique pour l'utilisateur. Ces différentes codifications sont incompréhensibles pour les utilisateurs non avertis. D'autres hétérogénéités sémantiques, que nous détaillerons pas ici, relèvent des incompatibilités d'encodage des valeurs (e.g. date, géométrie) ou encore de l'utilisation de différentes unité de mesures.

<table>
<tr><td>

```
GROUP = PRODUCT_METADATA
  DATA_TYPE = "L1TP"
  […]
END_GROUP = PRODUCT_METADATA
```

</td><td>

```
<Product_Info>
  […]
  <PROCESSING_LEVEL>Level-
  1C</PROCESSING_LEVEL>
  <PRODUCT_TYPE>S2MSI1C</PRODUCT_TYPE>
  […]
</Product_Info>
```

</td></tr>
<tr><td>a) Extrait du fichier de métadonnées d'une image Landsat 8</td><td>b) Extrait du fichier de métadonnées d'une image Sentinel-2</td></tr>
<tr><td>

```
</Product_Info>
<Product_Settings>
<PROCESSING_LEVEL>ORTHO</PROCESSING_LEVEL>
[…]
</Product_Settings>
```

</td><td>

```
<re:EarthObservationMetaData>
<eop:productType>L3A</eop:productType>
[…]
</re:EarthObservationMetaData>
```

</td></tr>
<tr><td>c) Extrait du fichier de métadonnées d'une image SPOT6</td><td>d) Extrait du fichier de métadonnées d'une image Rapid Eye</td></tr>
</table>

Listing 2.3 : Extrait des fichiers de métadonnées de 4 images acquises par quatre plateformes. Elles décrivent le type de produit proposé ou niveau de correction apportée à l'image. Nous parlons de niveau de traitements. Ces métadonnées décrivent le même niveau de traitements tout en utilisant une nomenclature propre à leur système de production. Dans ce cas, il s'agit d'images qui ont été corrigées en géométrie des déformations du relief, correction géométrique également appelée ortho rectification.

Les systèmes de diffusion des images, comme tout autre système de diffusion de données, repose sur l'utilisation des métadonnées pour fournir à minima un service de découverte qui en est le point d'entrée. Dans un contexte de mise à disposition d'images issues de différents dispositifs d'observation, comme au sein d'une infrastructure de données spatiales, la forte hétérogénéité des métadonnées qu'elle soit structurelle ou sémantique est un réel verrou. Il est de nature à limiter les capacités de découverte et d'accès à des catalogues multi capteurs. Néanmoins, certains producteurs institutionnels (Union européenne, gouvernement Américain) ont produit des efforts pour harmoniser les métadonnées et les formats de diffusion des images. Cela leur permet d'implémenter ainsi des sites portant sur la recherche d'images multi capteurs pour lesquelles ils maîtrisent le cycle de vie. Ce sont par exemple, les efforts entrepris par l'USGS et la NASA pour diffuser des images multi capteurs sur le portail EarthExplorer[30] ou encore l'effort produit par l'ESA qui offre au sein de son portail de données Earth online[31] un catalogue multi capteurs. Dans un cas comme dans l'autre, cette harmonisation est bâtie sur les préconisations des agences de standardisation attachées à ces institutions. Quant l'USGS harmonise avec le standard de métadonnées FGDC (FGDC, 98), l'ESA

[30] EarthExplorer : https://earthexplorer.usgs.gov/

[31] Earth Online : https://earth.esa.int/web/guest/data-access/online-archives

propose de le faire sur la base des recommandations produites dans la directive européenne INSPIRE[32] (European Commission Joint Research Centre, 13).

2.2.3 Standardisation des métadonnées

Différents standards de métadonnées, généralistes ou dédiés à une discipline joue un rôle essentiel pour faciliter le partage et la gestion des données dans des architectures décentralisées comme le sont les infrastructures de données. Un standard de métadonnées définit explicitement, et de manière prescriptive, les éléments de métadonnées et leurs propriétés. Il fournit les informations nécessaires à sa mise en oeuvre. Ils sont le socle de l'interopérabilité pour l'échange de métadonnées entre les différents centres de données. Pour ce qui concerne les données spatio-temporelles, nous pouvons citer, en premier lieu, les standards ISO 19115 (ISO, 03) pour l'information géographique. Le standard ISO 19115-2 (ISO, 09) est dédié aux images satellitaires. La spécialisation de la norme O & M (Observation and Measurement) proposée par (Gaspéri et *al*., 12) fournit des éléments de métadonnées supplémentaire pour décrire plus précisément les instruments de mesure et les conditions d'acquisition des images satellitaires.

Plusieurs de nos récents travaux (Desconnets et *al*., 14 ; Mougenot et *al*., 15 ; Desconnets et *al*., 17b) décrivent les différents standards de métadonnées. Le lecteur intéressé pourra s'y reporter. Leur rôle et l'étendue du périmètre fonctionnel pour lequel ils ont été conçus sont décrits en détail.

La standardisation traite les hétérogénéités au niveau du schéma de métadonnées. Elle propose en partie une harmonisation des valeurs à travers des énumérations ou liste de valeurs prédéfinies. De nombreux éléments de métadonnées sont laissés au libre choix de la communauté qui les met en application. Les conflits de représentation constatés (cf. listing 2.3) sont, de ce fait, partiellement couverts par la standardisation. Le fossé sémantique mis en lumière sur certains éléments de métadonnées demeure un important verrou si l'on se met dans une perspective de recherche d'information sur des catalogues multi capteurs à destination d'utilisateurs peu familiers de la télédétection.

2.3 Recherche d'information spatio-temporelle

Nos travaux font appel à des notions et des techniques du domaine de la recherche d'information (ou RI). Sans avoir la prétention de présenter le domaine de la recherche d'information dans toutes ces composantes, nous nous focalisons sur les approches de RI qui sont en relation avec notre domaine d'étude. Nous présentons les grands principes de la RI et mettons en relief les problématiques soulevées par les modèles mathématiques de recherche actuels, leur limites dans un contexte de recherche d'information sur des collections de données spatio-temporelles. Après avoir rappelé les différents travaux qui ont été entrepris dans le domaine de l'information environnementale, et plus particulièrement autour de l'utilisation des thésaurus et des ontologies comme support à la reformulation de requêtes, nous présentons la recherche à facettes. Nous posons un regard sur cette approche en la considérant non pas comme un nouveau modèle de recherche mais comme une méthodologie orientée utilisateur. En effet, cette dernière emprunte des techniques au domaine de la recherche d'information comme à celui de l'interaction Homme/Machine pour proposer une nouvelle

[32] INSPIRE : Infrastructure for Spatial INformation in Europe

approche d'interrogation et d'exploration des collections de documents. Elle nous semble particulièrement adaptée à notre contexte d'étude et aux besoins de la communauté environnementale.

2.3.1 Définitions et principes

La recherche d'information (RI) traite de la représentation, du stockage, de l'organisation et de l'accès à l'information. Le but d'un système SRI (Système de Recherche d'Information) est de retrouver, parmi une collection de documents préalablement stockés, ceux qui répondent au besoin de l'utilisateur exprimé sous forme la d'une requête. Un SRI se concentre sur la représentation des documents et des requêtes et leur mise en correspondance. Le fonctionnement d'un SRI consiste à un processus de recherche, également appelé processus en U (Belkin et *al*., 92). Nous le présentons en figure 2.5. Ce processus peut être décomposé en quatre sous processus :

- **Le processus d'indexation** : généralement automatisé, l'indexation consiste à analyser des documents en vue d'en extraire les mots clés ou les groupes de mots clés représentatifs du contenu informationnel d'un document. Elle repose sur les étapes suivantes : l'extraction des termes, la réduction du langage et la pondération des termes. Il est à noter que le processus d'indexation peut être appuyé par un thésaurus. Il permet de contrôler, normaliser, voire compléter les termes représentant le document dans le SRI,
- **Processus d'interprétation des requêtes**. Formulé, par l'utilisateur, dans un langage de requêtes qui peut être le langage naturel, un langage à base de mots clés ou le langage booléen, la requête est transformée en une représentation équivalente à celle du document en suivant les mêmes étapes,
- **Le processus de recherche** consiste alors à mettre en correspondance et à calculer le degré de correspondance des représentations des documents et de la requête. Les documents qui correspondent au mieux à la requête, ou documents dits pertinents, sont alors retournés à l'utilisateur, dans une liste ordonnée par ordre décroissant de degré de pertinence lorsque le système le permet,
- **Le processus de reformulation** : Afin d'améliorer les résultats de la recherche, le système peut être doté d'un mécanisme d'amélioration et de raffinement de la requête par reformulation. Nous abordons en détail cet aspect plus bas dans cette section.

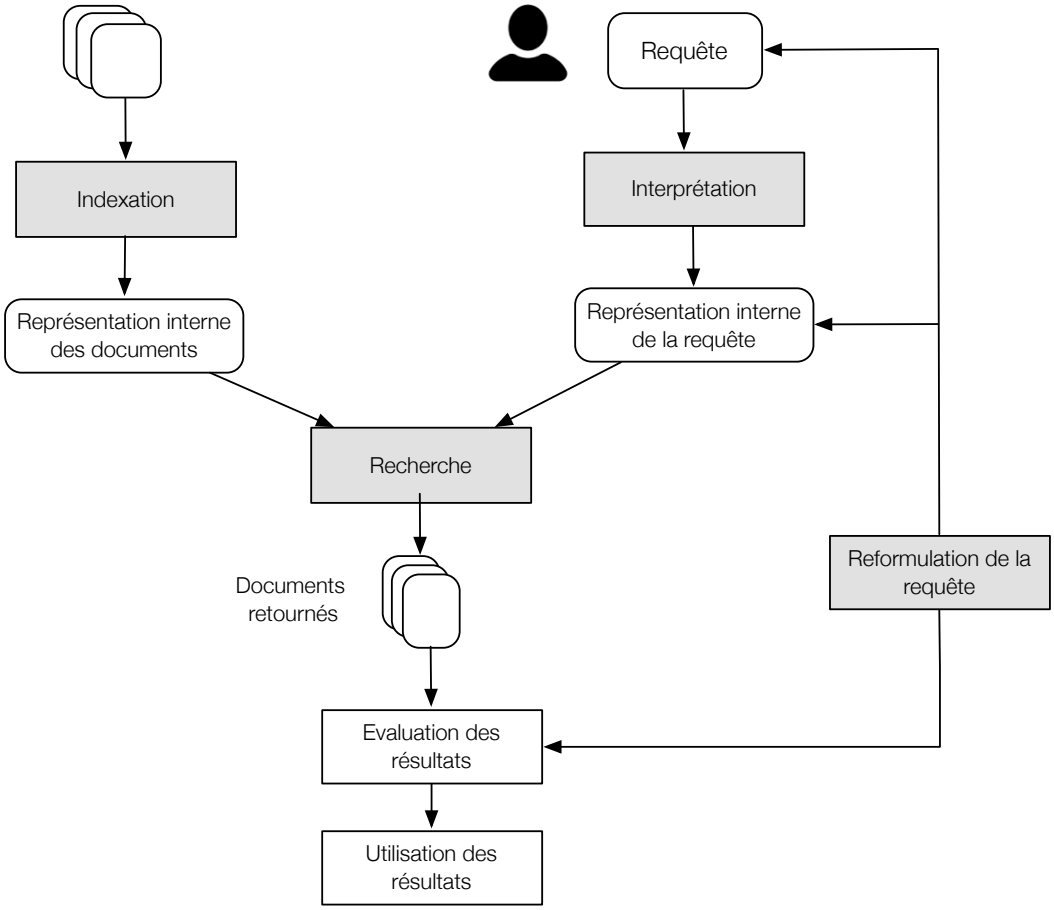


Figure 2.5: Processus de recherche en U.

Notions de précision et de rappel

La mesure de l'efficacité d'un SRI évalue sa capacité à retourner à un utilisateur, selon une requête donnée, le plus grand nombre de documents pertinents. Cette efficacité est traduite par deux notions (Cleverdon et *al*., 66) qui permet d'en faire une évaluation qualitative : la précision (*precision*) et le rappel (*recall*). Elles sont schématisées en figure 2.6. La première correspond au nombre de documents pertinents restitués par le système par rapport au nombre total de documents retournés. La seconde correspond au nombre de documents pertinents restitués par le système par rapport au nombre total de documents pertinents. (Chiaramella et Mulhem, 07) considère la première comme la pertinence du système et la seconde comme la pertinence de l'utilisateur. L'évaluation d'un SRI, à travers la mesure du rappel et de la précision, confronte donc le point de vue de l'utilisateur à celui du SRI. C'est une dimension clé du domaine : elle place l'utilisateur au centre du processus de recherche.

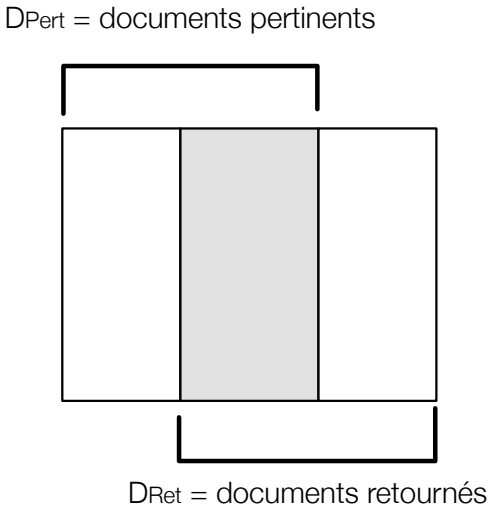


Figure 2.6: Notion de Précision et de Rappel

Notion de pertinence

De fait, la notion de pertinence et sa mesure sont donc au cœur de la recherche d'information. C'est une notion difficile appréhender car complexe et multi forme. Elle a fait l'objet de nombreux travaux et attentions (Spärk et Willet, 97). Elle diffère selon la perspective considérée, celle de la conception du SRI ou celle de son usage et les diverses situations rencontrées (librairie, web, base de données). Concrètement, dans les SRI, l'évaluation de la pertinence d'un document par rapport à une requête est basée sur la mesure de similarité entre le document et la requête. Plus la requête et le document ont de mots en commun, plus le document est considéré comme étant pertinent. A chaque modèle d'appariement document-requête utilisé correspond un calcul de pertinence. Bâti dans les premières heures de la RI sur le modèle booléen, l'amélioration des modèles d'appariement et le calcul de la pertinence ont contribué à faire avancer le domaine et a proposé des approches variées qui viennent remédier aux limites des modèles ensemblistes. Pour mémoire, les modèles de recherche sont classés dans trois grandes familles : les modèles booléens ou modèles ensemblistes, les modèles vectoriels ou algébriques et enfin les modèles probabilistes. Nous nous attarderons en particulier sur les deux premiers qui aujourd'hui implémentent la majorité des SRI des applications de recherche d'information en environnement.

Modèle Booléen

Pour le modèle booléen, le document est représenté par un ensemble de termes. La requête est représentée par un ensemble de mots clés reliés par des opérateurs booléens (AND, OR et NOT). L'appariement requête-document est strict et se base sur des opérations ensemblistes utilisant les opérateurs AND, OR ou NOT. Bien que simple de mise en œuvre, ce modèle présente plusieurs inconvénients : l'appariement est strict et ne permet de classer les résultats qu'en deux catégories : pertinent, non pertinent sans pouvoir les ordonner ; l'élaboration de requêtes à base d'expressions booléennes n'est pas accessible à un large public. Par ailleurs, l'appariement est lexical. Il est réalisé sur les termes et non sur le sens des termes. Le modèle booléen étendu proposé par (Salton et *al.*, 83) apporte à cet effet une pondération des termes dans le corpus de documents, et permet, entre autres, de

traiter le problème de classement des résultats. Ceux sont ces modèles, proche du langage de requêtes SQL utilisés dans les bases de données relationnelles, qui ont servi à bâtir les premiers systèmes de recherche de l'information pour l'environnement.

Modèle vectoriel (Vector Space Model) de Salton

Afin de s'affranchir des limites de l'approche booléenne, le modèle vectoriel proposé par (Salton et *al*., 75) représente un document dans un espace vectoriel composé de tous les termes de la collection de document. Les coordonnées d'un vecteur document représentent le poids des termes correspondants. Une requête est également représentée sous la forme d'un vecteur de termes défini dans le même espace vectoriel que le document. Cette décomposition permet d'introduire, et de calculer, la notion de mesure de similarité requête-document comme base d'évaluation de la pertinence système. Cette mesure de similarité correspond au cosinus de l'angle existant entre le vecteur de document et celui de la requête. Elle est également connu comme l'indice tf*idf[33]. Cet indice autorise le classement des résultats d'une recherche quand son évaluation numérique est rendue possible par le SRI. Cela permet de faire « remonter » des documents jugés les plus similaires à la requête dans des applications de recherche de type texte libre (*fulltext search),* recherches qui peuvent potentiellement retourner de nombreux résultats.

De part sa simplicité conceptuelle et sa facilité de mise en œuvre, ce modèle a été et est toujours largement implémenté dans des SRI ou librairies permettant d'en implémenter. Nous pensons en particulier aux outils diffusés depuis plusieurs années par la communauté du logiciel libre comme Apache Lucene ou Apache SolR, plus récemment la suite ELK[34] (ElasticSearch, Logstach, Kibana) qui proposent le VSM comme le modèle de recherche par défaut. Les deux premiers constituent des composants de recherche privilégiés pour l'implémentation d'applications de recherche d'information spatiale dans le domaine de l'environnement, comme par exemple l'outil de catalogage des données géographiques GeoNetwork[35].

Indexation sémantique

Les deux approches décrites plus haut, que nous qualifierons de « classiques », ne tiennent pas compte du sens des mots. Un même mot peut désigner différents concepts, et donc exprimer différents sens (polysémie), et différents mots peuvent avoir une même signification (synonymie). L'appariement lexical ne considère pas ces aspects. Ainsi, un SRI peut retourner des résultats non pertinents pour un utilisateur même si le résultat de l'appariement a pu être jugé pertinent par le système au regard du nombre de termes partagé entre requête et document.

Les travaux autour de l'indexation sémantique se sont attachés à remédier à ces problèmes en utilisant des index conceptuels ou sémantiques au lieu de simples mots clés. De tels index portent sur la sémantique des termes retenus. Ils sont construits soit à partir des concepts explicites des textes eux-mêmes (indexation conceptuelle), soit à partir de la sémantique latente des textes des documents

[33] tf*idf : Term Frequency-Inverse Document Frequency : http://www.tfidf.com/

[34] ELK, pour ElasticSearch, Logstach, Kibana, ou Elastic Stack est une suite logicielle open source et commerciale qui permet d'implémenter des applications de recherche, d'analyse et de visualisation de données : www.elastic.co

[35] Geonetwork : outil de catalogage open source pour les données spatiales : geonetwork-opensource.org

(indexation par sémantique latente (Deerwester et *al*., 90), soit à partir de la sémantique explicitée au sein de la connaissance organisée dans les dictionnaires, thésaurus ou ontologies (indexation sémantique). Nous verrons dans notre section proposition que notre démarche d'indexation tire partie, comme l'indexation sémantique, des référentiels issus de notre domaine pour contrôler, voire enrichir la sémantique des termes indexés.

Reformulation de requêtes

Par ailleurs, l'ambiguïté intrinsèque au langage naturel ainsi que la difficulté d'un utilisateur à exprimer ses besoins par une requête à partir de quelques termes amène une difficulté supplémentaire au retour de documents pertinents. L'approche envisagée consiste alors à aider l'utilisateur à reformuler sa requête pour qu'elle soit davantage en adéquation avec son besoin.

Un premier type de reformulation est constitué des méthodes globales. Elles permettent d'étendre une requête quels que soient les résultats retournés. Elles sont désignées comme des méthodes d'expansion de requêtes et opèrent en amont du processus d'appariement dans les SRI (Baziz et *al*., 03). La forme la plus commune d'expansion de requête est l'analyse globale qui utilise un thesaurus ou une ontologie (Voorhees, 94; Baziz et *al*., 07). Pour chaque terme, la requête peut être automatiquement étendue avec des mots du thésaurus synonymes ou liés au terme. Le système peut ainsi apparier la requête à des documents pertinents qui ne contiennent aucun des mots de la requête originale. Nos travaux précédents ont étendu la technique d'expansion par analyse globale à la recherche de données spatio-temporelles à la fois sur leurs composantes thématique et spatiale. Nous détaillerons ces travaux dans les sections qui suivent.

Le deuxième type de reformulation, appelées méthodes locales, ajuste la requête initiale à partir des premiers documents retournés. Les méthodes locales s'appuient sur la technique de réinjection de pertinence (*relevance feedback*) (Rocchio, 71 ; Harman, 92 ; Buckley et *al*., 94). L'idée de la réinjection de pertinence est de faire participer l'utilisateur dans le processus de recherche de sorte à améliorer l'ensemble final de résultats. La reformulation a pour base les documents manuellement jugés pertinents, ou non pertinents, par l'utilisateur ainsi que des méthodes de retour de pertinence aveugle ou implicite (*pseudo relevance feedback*) lorsque les premiers documents des résultats sont automatiquement supposés pertinents.

Comme l'indiquent (Chamralia et Muhler, 07) dans leur revue sur l'historique de la RI, démarrés sur des approches centrées système, les travaux du domaine de la RI se sont déplacés progressivement vers des approches centrées utilisateur et cela pour prendre en compte les éléments de contexte nécessaire à l'adaptation de la recherche à l'utilisateur (*contextual R*I). Suite à l'avènement du web, l'augmentation des utilisateurs et de leur diversité, la multiplicité des informations à rechercher participent à cette nouvelle orientation car il s'avère indispensable de tenir compte des situations diverses pour utiliser l'approche la plus appropriée. Ces nouvelles approches ont, entre autre, réinvesti les approches de reformulation globales ou locales précédentes en tirant partie des capacités offertes notamment par les développements autour des techniques de fouille de données ou celles issues du web sémantique.

2.3.2 Problématiques de recherche d'information liées aux données spatio-temporelles

Sur la représentation des données spatio-temporelles dans les SRI

Les objectifs et la nature de l'information à rechercher dans le domaine environnemental diffèrent en plusieurs points avec ceux des communautés des sciences de l'information, du domaine documentaire et du traitement automatique des langues naturelles. En effet à son origine, la RI s'est avant tout attachée à la recherche d'information issue de ressources d'information textuelles, non structurées ou semi structurées dans lesquelles le document est considéré comme l'unité d'indexation. Une des problématiques est alors la représentation d'un document au sein d'un SRI. L'enjeu consiste, pour l'essentiel, à assurer au sein de ces collections l'analyse de son texte afin d'en extraire les termes les plus représentatifs de son contenu.

Comme nous l'avons mentionné précédemment section 2.1 de l'état de l'art, l'information spatio-temporelle est une information structurée. Elle est accompagnée le plus souvent de ces métadonnées qui en fournissent son identification, ses caractéristiques ainsi que des éléments de contexte. Par ailleurs, outre sa dimension thématique, les dimensions spatiale et temporelle (cf. figure 2.1) constituent des éléments indispensables à sa représentation dans un SRI.

La structuration des métadonnées qui a fait l'objet de standardisation dans le domaine de l'information géographique permet assez facilement de mettre en correspondance la sémantique des éléments qu'elles portent avec leurs valeurs. De fait, cela facilite la représentation d'un jeu de données dans un SRI. Par contre, cette sémantique est avant tout celle partagée par les producteurs des données. Pour cela, et comme le constat fait par (Smeulders et *al*., 00) dans le contexte de la recherche d'information multimédia, le premier obstacle de la représentation des données spatio-temporelles au sein d'un SRI est la résolution du fossé sémantique existant entre les caractéristiques et leur interprétation symbolique par l'utilisateur. Ce constat a été discuté au paragraphe 2.2.2.

Pour les dimensions spatiale et temporelle, la problématique porte d'une part sur l'adoption de référentiels de positionnement ou de nommage qui est essentielle pour assurer une indexation et des interrogations uniformes sur ces dimensions. Elle nécessite également de prévoir une structure d'indexation adaptée qui permette d'interroger efficacement cette dimension avec les opérateurs topologiques *ad hoc*. Enfin, les informations spatio-temporelles peuvent être vues à travers différents niveaux de granularité (Desconnets et *al*., 03). La représentation et la recherche peuvent donc être envisagées à ces différents niveaux. Dans la réalité, le niveau de recherche s'avère être le plus souvent être celui du jeu de données (*dataset*) ou celui de la collection (*dataseries*).

Sur les systèmes de recherche d'information spatio-temporelles

Ce n'est que tardivement que la communauté environnementale s'est saisie des travaux émanant du domaine de la RI. Les problématiques d'accès à l'information environnementale se sont avant tout posées au travers de la structuration et de l'interrogation de l'information. Ainsi, les principaux enjeux traités ont porté sur l'optimisation des structures de stockage afin de favoriser l'interrogation, l'analyse de l'information au sein de systèmes de persistance des données tels que les bases de données relationnelles ou encore les entrepôts de données multi dimensionnels de type SOLAP[36] (Rivest et *al*.,

[36] *SOLAP : Spatial On-Line Analytical Processing*

05). Les langages d'interrogation alors utilisés viennent de l'algèbre relationnelle (SQL et ses dérivés) et les requêtes sont construites en grande partie sur les opérateurs booléens.

Ce n'est que récemment, fin des années 2000, que le besoin de partage et l'effort de mutualisation d'information consenti par la communauté environnementale a fait émerger de nouveaux systèmes d'information, les infrastructures de données spatiales (Desconnets et Kazmierski, 15). Ces systèmes de partage requièrent un composant qui assure la recherche et la localisation de l'information qui par nature est répartie chez les différents fournisseurs de données. Si toutefois, une partie des infrastructures ont bâti leur service de découverte en y intégrant un composant dédié à l'indexation des données partagées, d'autres restent sur des SRI issus des SGBD. Pour les premiers, ils sont bâtis avec des outils « sur étagère » comme *Geonetwork* et ses déclinaisons françaises *Geosource*[37], *Prodige*[38]. Ils peuvent également utiliser des librairies ou composants issus du monde de la RI comme ceux cités plus haut. Initialement construits pour assurer la recherche sur des informations textuelles, ces outils se sont dotés récemment de capacités d'indexation et de recherche sur des informations spatialement localisées en proposant l'encodage de coordonnées géographiques au sein des index. Plus récemment encore, les librairies comme le *Spatial Module* d'*Apache-Lucene* 4.0[39] proposent la prise en charge de l'indexation et la recherche sur les différentes primitives géométriques comme le point, la ligne ou le polygone et les opérateurs métriques et topologiques associés. D'autres contributions concernent l'adaptation des structures sur lesquelles est indexée la dimension spatiale de l'information. En effet, les requêtes reposant sur l'interrogation de l'empreinte spatiale d'un jeu de données peuvent s'avérer coûteuses en temps car elles font appel à des opérateurs topologiques (intersection, inclusion, voisinage) dont la complexité de calcul est élevée. (Heurteaux, 16) propose une extension au moteur *ElasticSearch* en lui associant un index spatial de type RTree (Rectangle Tree) de (Guttman, 84) qui est l'équivalent d'un arbre balancé dans lequel les données sont organisées autour de la notion de rectangle englobant. Ils reposent sur l'implémentation de la librairie *JTS* (*Java Topology Suite*). Toutes les requêtes spatiales sont alors déléguées à cet index spatial qui optimise les requêtes formulées à partir d'une géométrie.

Sur la formulation des requêtes

La formulation d'une requête par un utilisateur sur un SRI est rarement triviale. Le choix des termes ou de l'expression requiert une connaissance du domaine et de la terminologie qui est généralement celle employée par les concepteurs du SRI. Ce qui est bien évidemment beaucoup demandé à un utilisateur. L'ajout de composant apportant des listes prédéfinies ou l'auto complétion du formulaire de saisie traite en partie le problème. Les termes proposés, par nature, ne définissent pas précisément, ou trop peu, le besoin soit par manque de richesse, soit parce qu'ils véhiculent une ambiguïté sémantique. Par ailleurs, la mise en place d'une requête portant sur plusieurs critères peut être délicate, voire déroutante. Les difficultés pour formuler une requête sont également associées à la complexité des formulaires proposés par les applications de recherche. Dans ces dernières, l'utilisateur doit, en aveugle, paramétrer une requête en utilisant au minimum le critère *Quoi ?* associé à d'autres

[37] Géosource : outil libre de catalogage des données géographiques pour les acteurs publics français : http://www.geosource.fr/

[38] Prodige : Plate-forme Régionale pour Organiser et Diffuser l'Information GEographique : https://adullact.net/projects/prodige/

[39] Spatial module Apache lucene 4.0 : https://lucene.apache.org/core/4_0_0/spatial/

critères comme *Quand ? Où ?* et *Comment?* Cela est d'autant plus vrai lorsque cette recherche porte sur des données spatio-temporelles où les conditions d'acquisition, de traitements ou le média de distribution, par exemple, revêtent un caractère particulièrement discriminant dans l'appréciation de l'adéquation avec le besoin. Ce sont autant de difficultés auxquelles l'utilisateur doit se confronter. La plupart du temps, il va se retrouver face à un moteur de recherche « muet » ou au contraire trop « bavard ». D'une manière générale, ces approches guident faiblement l'utilisateur dans le choix des termes potentiellement pertinents pour formuler une requête. Enfin, elles n'offrent pas d'interaction entre la phase de formulation de la requête et l'obtention des résultats de la requête, interaction qui permettrait par une formulation successive d'aboutir à des résultats plus précis et moins nombreux.

Ces limites, inhérentes autant au modèle de recherche qu'à la « rudesse » cognitive des interfaces Homme Machine (IHM), sont d'autant plus prégnantes lorsque l'on souhaite formuler une partie de la requête sur la dimension spatiale de l'information. En effet, dans ce type de recherche, en l'absence d'informations sur le contenu de la donnée, comme cela peut être le cas pour une image satellitaire, le critère Où ? revêt un intérêt tout particulier. Il constitue la plupart du temps, en association avec le critère temporel Quand ? le premier critère utilisé et le plus discriminant. Le choix puis la saisie de l'emprise spatiale devient alors déterminante dans la pertinence de la recherche.

Les approches issues des pratiques du web ont essayé, en proposant une recherche appelée *fulltext search* de rendre la liberté d'expression à l'utilisateur en misant sur les capacités de classement du système (cf. indice de similarité évoqué plus haut). Il n'en demeure pas moins que de telles recherches privilégient la quantité de résultats sur leur qualité. Elles doivent nécessairement pouvoir traduire syntaxiquement et sémantiquement l'expression textuelle pour reconnaître les termes relatifs aux dimensions spatiale et temporelle (un lieu, un instant, une période) puis les traiter de sorte à bien baser les interrogations sur les dimensions concernées. L'apparition de services web offrant l'accès à des bases de données toponymiques tels que Geonames[40] ont permis à certains de moteur de recherche d'information *fulltext* comme le moteur de recherche RESTO[41] de mettre en correspondance les noms de lieu et les enveloppes spatiales qui délimitent les données spatio-temporelles.

2.3.3 Stratégies pour la recherche d'information spatio-temporelle

A l'utilisation des mêmes approches de recherche émergent les mêmes limites que celles identifiées dans le domaine de la RI orientée document. Des difficultés supplémentaires, liées à la fois à la nature spatiale de l'information mais également à son mode de production apparaissent. La sémantique relative au contenu des données, est, en plus d'être ambiguë, pauvre et peu précise, voire inexistante. Les divers travaux entrepris ces dernières années ont été réalisés dans le cadre de la recherche d'information spatio-temporelle dans le contexte de la mise en place des infrastructures de données spatiales. Les travaux les plus significatifs de notre communauté reprennent et étendent les approches de reformulation de requêtes proposées dans la RI pour les adapter aux données spatio-temporelles et aux architectures des systèmes de partage mises en oeuvre dans le domaine environnemental.

Une synthèse de nos travaux effectués, entre les années 2006 et 2011, sur cette problématique est présentée dans la première partie du mémoire (section 2.2.1 Activités autour de l'interopérabilité des

[40] Geonames : base de données géographiques mondiale : http://www.geonames.org/

[41] RESTO : moteur de recherche pour l'imagerie satellitaire https://github.com/jjrom/resto

systèmes d'information en environnement) Deux démarches ont été entreprises. Elles ont comme point commun l'aide à la formulation de requêtes avec l'appui d'une ressource terminologique et/ou spatiale. La première des pistes a exploré les stratégies d'expansion de requêtes à partir de référentiels terminologique et spatial, la seconde a porté sur l'appui de la représentation visuelle d'un référentiel terminologique comme outil de découverte et de sélection du vocabulaire spécifique à un domaine.

Aide à la formulation de requêtes dans des architectures distribuées et hétérogènes

De nombreux autres travaux, nous en citons quelques uns : (Lutz et Klien, 06 ; Shvaiko et *al*., 10 ; Santoro et *al*., 12, Gui et *al*., 13) ont investi la même problématique sur laquelle nous nous sommes penchées mais dans un contexte de médiation à plus grande échelle où la quantité et la diversité des sources d'information, d'utilisateurs doivent être pris en compte. Dans ce contexte, (Santoro et *al*., 12 ; Gui et *al*., 13) relèvent les mêmes difficultés pour la recherche d'information spatio-temporelle que celles énoncées plus haut. Les solutions apportées sont diverses mais celles s'appuyant sur des ressources onto-terminologiques sont privilégiées avec un investissement plus ou moins important de l'utilisateur. Nous en analysons trois d'entre elles.

Comme nous l'avons proposé avec l'outil MDweb (Desconnets et *al*., 07 ; Desconnets et Libourel, 12), (Shvaiko et *al*., 10) propose d'étendre l'outil open source GeoNetwork, en y associant un composant d'expansion de requêtes. Le composant s'appuie sur l'alignement des termes de la requête initiale sur les concepts d'une ontologie facettée, c'est à dire présentant un domaine de connaissances par plusieurs sous-arbres. Il s'agit d'étendre les termes de la requête à ses synonymes. (Santoro et *al*., 12) propose deux approches : une approche centrée utilisateur qui tire partie de l'historique des recherches pour proposer une reformulation par expansion. Un module *Recommanded Module* est dédié à ces expansions en assurant l'alignement entre termes issus des métadonnées et ceux de l'historique d'autre part. Il autorise également l'utilisateur à enrichir les annotations présentes dans les métadonnées. Il est alors délicat d'en garantir la qualité. Il n'y a pas de contrôle *a posteriori*. La deuxième approche privilégie l'utilisation de ressources termino-ontologiques standardisées issues de la communauté des producteurs de données au format RDF SKOS (W3C, 04). Un composant tiers, le médiateur, qui est en charge de la transformation syntaxique des métadonnées est complété par un dispositif, *Discovery Augmented Component (DAC)*. Il propose une reformulation des requêtes par expansion automatique ou assisté par l'utilisateur. Cette approche a l'avantage, en déléguant la reformulation à un composant tiers, de permettre son extensibilité. L'ajout de nouveaux référentiels, par alignement sur les précédents, permet d'étendre les champs de connaissance proposés ou d'étendre le multi-linguisme. Par contre, cette approche n'est pas aussi précise que celle qu'un producteur pourrait entreprendre. L'alignement automatique entre valeurs de métadonnées et référentiel terminologique que requiert la délégation de ces opérations à un composant tiers peut engendrer des expansions erronées. Enfin, les reformulations proposées par le DAC sont limitées au critère *Quoi ?* certes essentiel mais insuffisant.

Dans un contexte équivalent à celui des travaux de (Santoro et *al*., 12) mais avec un souci plus prononcé sur le niveau de performances attendu (Gui et *al*., 13) propose un composant de médiation qui s'appuie sur des ressources termino-ontologiques. Ses éléments décrivent aussi bien le contenu des données que les services web qui permettent d'y accéder. Cela permet d'envisager la reformulation des requêtes d'un utilisateur en tenant compte à la fois du domaine et des contraintes d'accès aux données (les services web). A cette reformulation est associée un calcul de similarité qui permet d'évaluer la qualité des appariements entre résultats retournés et requête initiale. Il fournit ainsi à

l'utilisateur un bon moyen de classer les résultats en fonction de leur pertinence.

Limites

Ces différentes approches de reformulation de requêtes bien qu'elles améliorent les précédentes ne sont pas entièrement satisfaisantes. Tout d'abord, elles demandent à l'utilisateur, les unes comme les autres, un investissement non négligeable pour reformuler une requête sans garantie de retourner des résultats plus proches de ses besoins. Par ailleurs, elles ne traitent pas les ambiguïtés sémantiques introduites lors de la description des jeux de données même si les expansions proposées sur les hiérarchies de concepts contribuent à les diminuer.

Enfin, l'appui à la formulation de requêtes sur la dimension spatiale a été en partie traitée par l'expansion spatiale proposée par les travaux de (Boisson et *al.*, 06). Néanmoins, elle s'est avérée délicate à mettre en œuvre pour des catalogues portant sur de grands territoires ou de grands ensembles de données. Et par là, elle a posé du problème de passage à l'échelle tant dans la mise en place du référentiel spatial que par le coup des requêtes. Ce qui est réellement problématique dans des SRI mis en place au sein d'architectures décentralisées.

2.3.4 Recherche à facettes

L'acceptation d'un SRI repose en partie sur la qualité de l'interaction entre système et utilisateur. Dans notre domaine, la majeure partie des SRI se limitent à des modalités d'interaction qui ne vont pas au delà de la boucle requête/réponse ou l'inclusion d'une boucle de reformulation de requête (par réinjection de pertinence ou par expansion). Comme le souligne (Charmaralia et Mulher, 07), le domaine de la RI a connu de grandes avancées ces dernières années mais les efforts consentis ont rarement porté sur l'interactivité Homme Machine et cela malgré des travaux précurseurs sur des modèles cognitifs de la RI interactive (Spink et Saracevic, 70 ; Ingwersen, 99). Par contre, une communauté de chercheurs a investi un champ de recherche proche : la visualisation d'information. Elle vise à étudier les représentations visuelles ou vues sur les données pour les rendre explorables, analysables. Elle a notamment pour objectif de proposer et concevoir des systèmes de visualisation interactifs permettant l'analyse visuelle de données en exploitant au mieux les capacités cognitives et visuelles de l'humain (Jacko et Sears, 03 ; Card et *al.*, 99). La recherche à facettes peut être vue comme le résultat de la rencontre de la RI et de la visualisation d'information interactive auquel les sciences de l'information ont apporté leur contribution, notamment au travers des réflexions posées sur les systèmes de classification de la connaissance. Après avoir défini les notions relatives à cette méthodologie d'accès à l'information orientée utilisateur, nous détaillons les principes de la recherche à facettes en nous appuyant sur quelques exemples de mise en œuvre. Nous concluons cette section en montrant en quoi cette approche est en adéquation autant avec la problématique de recherche d'informations spatio-temporelles qu'avec la communauté des utilisateurs concernée par ce besoin d'information.

2.3.4.1 Définitions et principes

Notion de Facette

Dans son utilisation courante, le terme facette désigne par métaphore un aspect d'une réalité complexe qui est envisagée sous un angle particulier : facette d'une personnalité, d'un problème. Ce

terme implique l'appartenance à un objet concret ou à une abstraction. Il peut être également perçu comme une caractéristique contrastant fortement avec les autres caractéristiques (Maniez, 99). Finalement, et cela fait écho aux primitives conceptuelles que nous manipulons, (Maniez, 99) étend sa réflexion sur le sens du terme facette pour le généraliser : « Ces caractéristiques contrastées peuvent être aussi bien celles d'un individu [..] que celles d'une classe d'individus [..]. Dans ce dernier cas de figure, une facette est un trait commun à tous les éléments de la classe. Elle constitue une catégorie de jugement, un critère caractéristique de chaque individu [...] ». Comme le souligne de nombreux auteurs, la notion de facette apparaît comme l'apport théorique le plus important du siècle dernier en science de l'information. C'est à partir des travaux du bibliothécaire indien Raganathan, qui a crée la classification à colonnes, que la notion de facette acquiert tout son sens dans le monde documentaire. Dans ce contexte, Une facette peut être définie comme l'une des dimensions d'analyse d'un sujet dans une classification multi-dimensionnelle (Desfriches-Doria, 12). Nous serions tenté de rajouter qu'une facette peut être également considérée comme un point de vue particulier sur un sujet, un domaine ou une classe d'objets.

Classifications à facettes

La classification à facettes est une alternative au modèle rigide des classifications bibliographiques qui sont énumératives et organisées sous forme hiérarchique, comme la *Dewey Decimal Classification* (DDC) ou encore la *Library of Congress Classification* (LCC). En effet, ces dernières n'autorisent pas facilement la combinaison de termes provenant de différentes parties d'un schéma de classification (Spiteri, 98). Issu des propositions de classifications à colonnes (CC) de (Raganathan 33), le modèle de classification à facettes permet d'exprimer sous la forme d'une combinaison personnalisée d'indice-concept des sujets complexes ou composés. Raganathan, 67 la met à jour et la complète pour proposer des principes pour la construction de schémas de classification à facettes, qui sont repris au sein du modèle PMEST (Personnalité, Matière, Energie, Espace, Temps). Nous donnons un exemple, extrait de (Maniez, 99), qui permet de différencier ce modèle de classification du modèle hiérarchique. Soit le sujet « prévention des maladies du riz » indexé en DDC puis en CC. L'indice DDC le plus proche est *633 189 8*. Il est dérivé par spécifications successives de l'indice de base *633* :

633 =céréales

633.18 = riz

633.189 = maladies du riz

633.189.8 = maladies du riz d'origine virale

Il est impossible d'ajouter à l'indice DDC la notion de prévention.

L'indice CC est : EJ,381;421:5: . Il est décomposé en quatre indices élémentaires traduisant chacun une facette:

EJ = agriculture : facette principale

381 = riz : facette : facette Personnalité

421 = maladie virale : facette Matière

5 = éradication : facette Energie

L'indice CC pourra être reformulé pour prendre en compte la notion de prévention avec la facette de type Energie.

La classification à facettes est obtenue par le biais de l'analyse par facettes, une technique détaillée notamment dans (La Barre, 10 ; Spiteri, 98 ; Denton, 03). D'une manière générale, il s'agit d'une technique en deux étapes ; la première correspond à l'analyse, qui consiste à déconstruire un sujet en ses diverses composantes, et la seconde, la synthèse, qui concerne la génération d'un indice significatif intégrant les différents composants du sujet en fonction de règles syntaxiques (Hudon et Mustafa El Hadi, 10). Pour cela, (Beghtol, 08) considère que la classification à facettes correspond à une approche analytico-synthétique qui s'apparente à un dispositif syntaxique et non sémantique.

Dans les années 2000, la communauté s'intéressant à la recherche d'information sur le web s'approprie cette notion et fait émerger une vision fonctionnelle orientée utilisateur pour mettre la classification à facettes au service de l'exploration et l'accès à l'information. Pour (Denton, 03) la classification à facettes est «, et que l'utilisateur peut utiliser en cherchant ou en parcourant pour trouver ce dont il a besoin ». Son usage, dans un contexte de recherche d'information sur le web, se comprend car comme l'énonce, (Brougthon, 06), sa structure est compatible avec une interface utilisateur. En effet les points d'accès multiples, les différentes dimensions du corpus, en favorisent la navigation. Enfin, la recherche est facilitée par le filtrage progressif basé sur des critères multiples (les facettes). Avant de définir la recherche à facettes ou recherche facettée, en anglais *faceted search,* nous évoquons ces prédécesseurs que sont la recherche paramétrique (*parametric search)* et la navigation à facettes ou *faceted navigation*. Cette progression apporte une meilleure compréhension des faiblesses comblées par une recherche facettée.

Recherche paramétrique

Une interface de recherche paramétrique est une interface de recherche booléenne pour une collection de produits, de documents pour laquelle un certain nombre de caractéristiques (facettes) sont proposées comme contraintes à la recherche. Elles permettent aux utilisateurs de formuler des requêtes en spécifiant visuellement ces caractéristiques sur leurs valeurs proposées (Neal, 97). La requête correspond alors aux valeurs sélectionnées dans une seule facette combinées à l'aide d'un OR alors que les valeurs associées à différentes facettes sont combinées à l'aide d'un AND logique. Le SRI répond en retournant les objets de la collection qui satisfont ces contraintes. Les valeurs des caractéristiques peuvent être une énumération, une hiérarchie de valeurs ou encore des valeurs numériques. La figure 2.7, tirée de (Tunkelang, 09) schématise une interface paramétrique pour la recherche de vins.

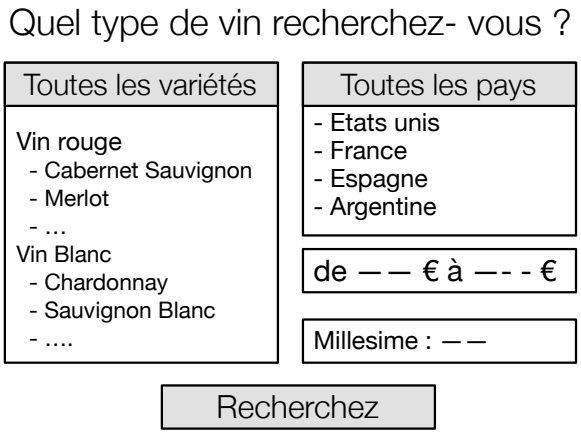


Figure 2.7: Interface d'une recherche paramétrique sur une collection de vins d'après (Tunkelang, 09)

La recherche paramétrique est une forme de recherche booléenne dans laquelle sont posées des contraintes de recherche (les facettes) qui évitent à l'utilisateur de baser sa recherche sur des expressions textuelles libres. Si cette recherche offre une meilleure expressivité, elle pose malgré tout certains des problèmes déjà évoqués. Les utilisateurs ont du mal à formuler leurs requêtes. Ils peuvent être alors confrontés au « silence » du SRI ou au contraire à un « bruit » tel qu'il est délicat de trouver l'information pertinente.

Navigation à facettes

La navigation à facettes ou navigation facettée comble le manque de la recherche paramétrique en guidant l'utilisateur dans l'élaboration de sa requête. En effet, la recherche paramétrique exige que l'utilisateur exprime son besoin par une requête qui est construite en une seule fois, en sélectionnant toutes les facettes d'intérêt. En revanche, la navigation facettée permet à l'utilisateur d'élaborer une requête progressivement, en voyant l'effet de chaque choix dans une facette sur les choix disponibles dans d'autres facettes. La figure 2.8 donne l'exemple du filtrage progressif, en une séquence de 3 étapes, pour aboutir à la sélection d'un vin inférieur à 10€. Cette sélection progressive est rendue possible par le changement d'états et donc de choix des facettes en fonction du choix précédent. Dans notre exemple, la contrainte du budget inférieur à 10€ élimine les vins français. D'autres sélections dans la facette « variétés » éliminent les possibilités dans les facettes non sélectionnées. Par exemple, la sélection de l'Espagne en tant que région élimine le Sauvignon Blanc comme option pour la facette variétale.

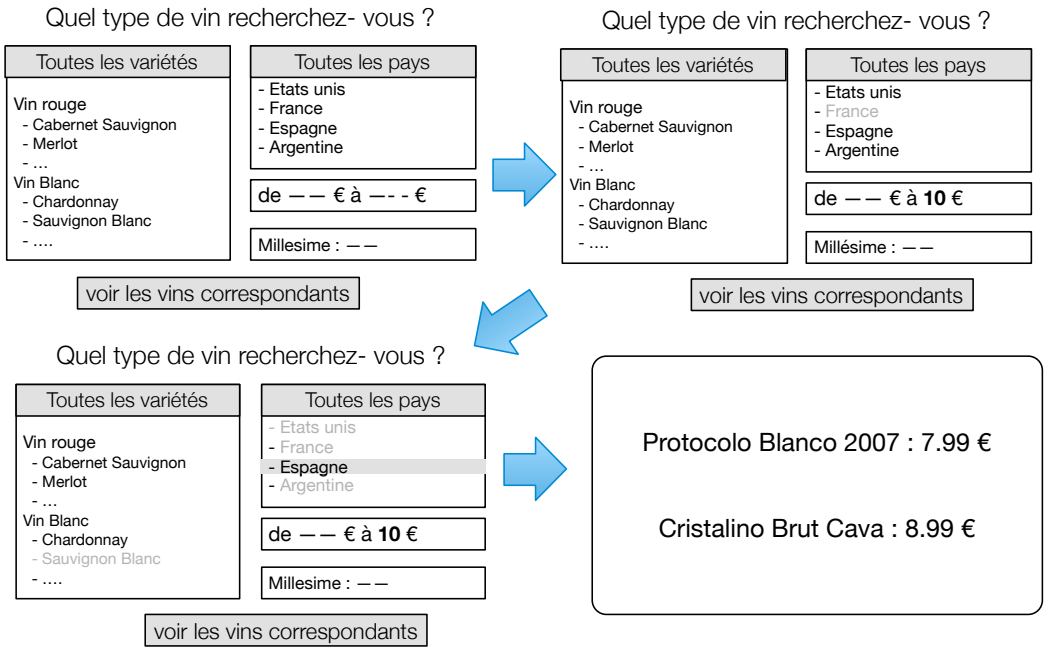


Figure 2.8: Schéma d'une recherche par navigation facettée. Les numéros associés aux différentes interfaces correspond à la séquence qui mène à la sélection finale des deux vins d'après (Tunkelang, 09)

L'utilisateur ne voit que les vins avec les caractéristiques souhaitées. Finalement, la navigation facettée offre une expérience de raffinement interactif par l'élaboration progressive de la requête. Du point de vue de l'utilisateur, elle élimine les "impasses" qui peuvent résulter de la sélection de combinaisons insatisfaisantes entre les facettes. Cependant, toutes sortes de facettes ne se prêtent pas aux interfaces de navigation à facettes. Par exemple, dans l'interface représentée ci-dessus, l'utilisateur pourrait encore arriver à une impasse en choisissant des vins pour moins de 1 €. Une interface utilisateur plus sophistiquée pourrait éviter cette voie sans issue, par exemple, en divisant les valeurs numériques en plages discrètes. En général, la navigation à facettes n'a de sens que pour les facettes dont les valeurs peuvent être présentées au moyen de listes de choix.

Recherche à facettes

La différence entre la notion de navigation à facettes et la recherche à facettes peut paraître subtile. (Tunkelang, 09) définit ainsi la recherche à facettes comme la combinaison d'une navigation à facettes avec une recherche plein texte (*fulltext*) ». Si les recherches paramétrique et de navigation à facettes supposent que les représentations des documents soient structurées à partir d'un système de classification à facettes, la recherche à facettes autorise la recherche sur des représentations semi-structurées. Effectivement, le plus souvent, une collection de documents, de produits ou de données est représentée par une combinaison de texte non structuré et d'attributs structurés ou de métadonnées. Ce sont les métadonnées qui fournissent les caractéristiques d'un document dont les valeurs sont conformes à un système de classification à facettes. La navigation à facettes se fera sur le contenu structuré quand la recherche plein texte est faite sur le contenu non structuré des représentations des documents. Généralement, la recherche plein texte fait appel à des contenus portés par le descriptif d'un document (résumé, titre) qui vient compléter ses caractéristiques. La figure 2.9 donne un exemple de recherche à facettes sur des produits d'actifs réseau. Le bandeau du haut propose un champ de saisie libre. Le bandeau gauche propose les différentes facettes à sélectionner. Généralement, la

recherche est initiée par la saisie d'un terme ou d'une expression dans le champ de saisie. Sur la base des résultats reçus, la sélection progressive des valeurs de facettes permet de raffiner la recherche.

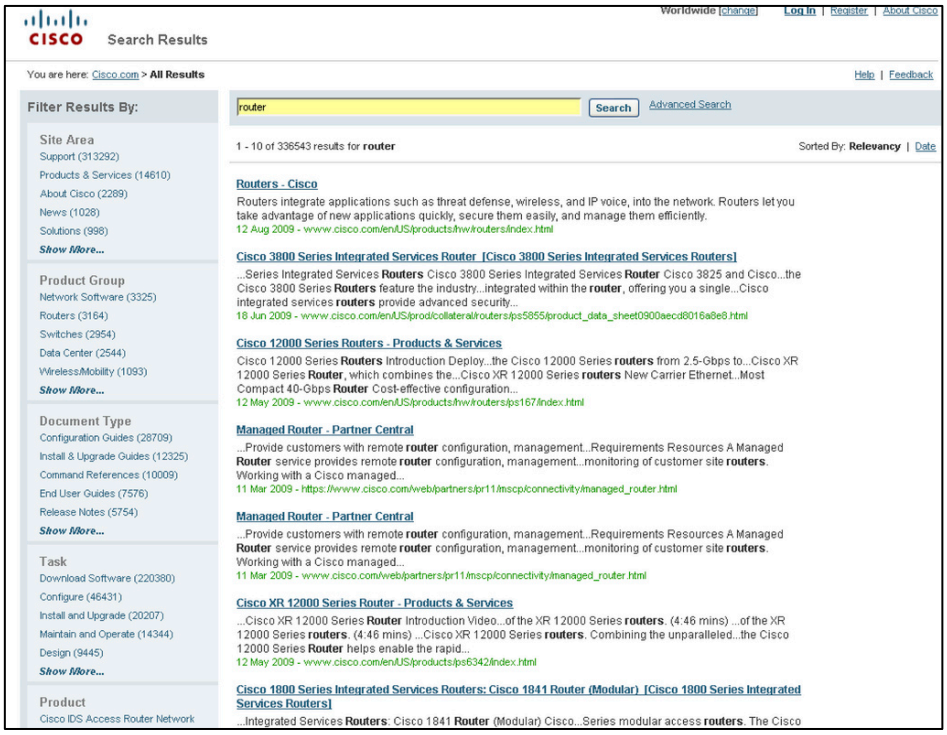


Figure 2.9: exemple de recherche à facettes. Site du fournisseur d'actifs réseau CISCO (d'après le site de Morville[42])

Dans une telle approche, l'interactivité des IHM est prépondérante. Elle nécessite un traitement particulièrement efficace des requêtes par le SRI pour que le choix d'une valeur de facette soit traduit simultanément par la mise à jour conjointe des résultats et des valeurs de facettes correspondant aux caractéristiques des résultats retournés. La rapidité de traitement des requêtes et la mise à jour des IHM deviennent des conditions *sine qua none* à la mise en œuvre de la recherche à facettes.

Les réflexions issues du domaine de la visualisation d'information apportent aux IHM toute l'interactivité nécessaire à l'amélioration de l'expertise des utilisateurs. Il est en particulier rendu possible de formuler des requêtes de manière progressive. L'interaction proposée dans de telles interfaces offre le *feedback* (nombre de documents associés à chaque facette, uniquement les facettes sur lequel des documents sont associés) nécessaire à l'utilisateur pour permettre ce filtrage successif et proposer une sélection fine des documents recherchés. La RI vient rendre possible l'interactivité recherchée en fournissant les techniques d'indexation et d'interrogation efficaces pour retourner les résultats de manière quasi instantanée. Enfin, la recherche à facettes tire également partie des systèmes de classification de connaissances pour proposer une décomposition multi points de vue d'un domaine complexe comme nous pouvons le rencontrer dans les sciences de l'environnement.

[42] http://www.flickr.com/photos/morville/collections/72157603785835882/.

2.3.4.2 Verrous pour la mise en œuvre au sein des SRI

Depuis déjà plusieurs années, la mise en œuvre de recherche à facettes est sortie de la sphère académique pour faire l'objet d'applications commerciales ou open source qui la mette ainsi à portée de nombreux développeurs d'applications de recherche. Néanmoins, lors de la conception de telles applications, divers points doivent faire l'objet d'une attention particulière. (Tunkelang, 09) en liste un certain nombre en distinguant ceux relatifs à l'indexation et l'efficacité du SRI et ceux relatifs à l'utilisabilité des IHM de l'application de recherche. Sans passer en revue l'ensemble de ces points, nous relevons que les verrous liés au passage à l'échelle des SRI et à leur efficacité sont aujourd'hui levés et cela suite à l'évolution des architectures matérielles et des logiciels disponibles dans le sphère open source ou commerciale. Nous citerons par exemple les travaux du concepteur de l'outil open source Apache SolR Yonik Seeley [43] qui propose des structures d'index et des techniques d'interrogation suffisamment efficaces pour que de tels SRI puissent passer l'échelle.

Les métadonnées facettées

Par contre, d'autres restent toujours d'actualité et sont au cœur de nos préoccupations. En premier lieu et contrairement aux approches classiques en RI, la recherche à facettes doit être organisée autour de contenus structurés ou semi structurés rattachés à une classification multi dimensionnelle du domaine traité. Construire de telles organisations sur des contenus non structurés reste délicat et peu satisfaisant. Comme le relève (Hearst et *al*., 06), les techniques de classification de contenu non structuré issues du domaine de la fouille de données, comme le regroupement (*clustering*), produisent de la confusion lors de la recherche car le regroupement de termes n'est pas prédictible et mélange des termes ayant différents niveaux de granularité. Aussi, la mise en place d'un index à partir d'une métadonnée facettée demeure une solution à privilégier car elle permet de maîtriser la stabilité et la cohérence sémantique des dimensions de recherche proposées à l'utilisateur. (Yee et *al*., 03) introduit la notion de métadonnées facettée (*faceted metadata*). Il met l'accent sur la nécessité d'organiser les contenus sur laquelle la recherche doit être mise en œuvre. Cette notion met à la fois en avant le besoin de structurer les différentes caractéristiques d'un contenu, d'un produit par un ensemble d'éléments de métadonnées. Il met également l'accent sur le besoin de définir les éléments de métadonnées pour qu'ils puissent exprimer les différentes dimensions de ce contenu. Les concepts issus de la classification à facettes sont alors utilisés pour fournir les valeurs des éléments de métadonnées pour chacune des dimensions représentées par les métadonnées. Le diagramme de classe en figure 2.10 formalise cette notion qui met en relation le schéma d'organisation des contenus à celui de la connaissance formalisée par une classification à facettes. Nous précisons cette vision dans notre contexte en remplaçant la notion de document par celle de jeu de données qui est plus appropriée.

[43] https://yonik.wordpress.com/2008/11/25/solr-faceted-search-performance-improvements/

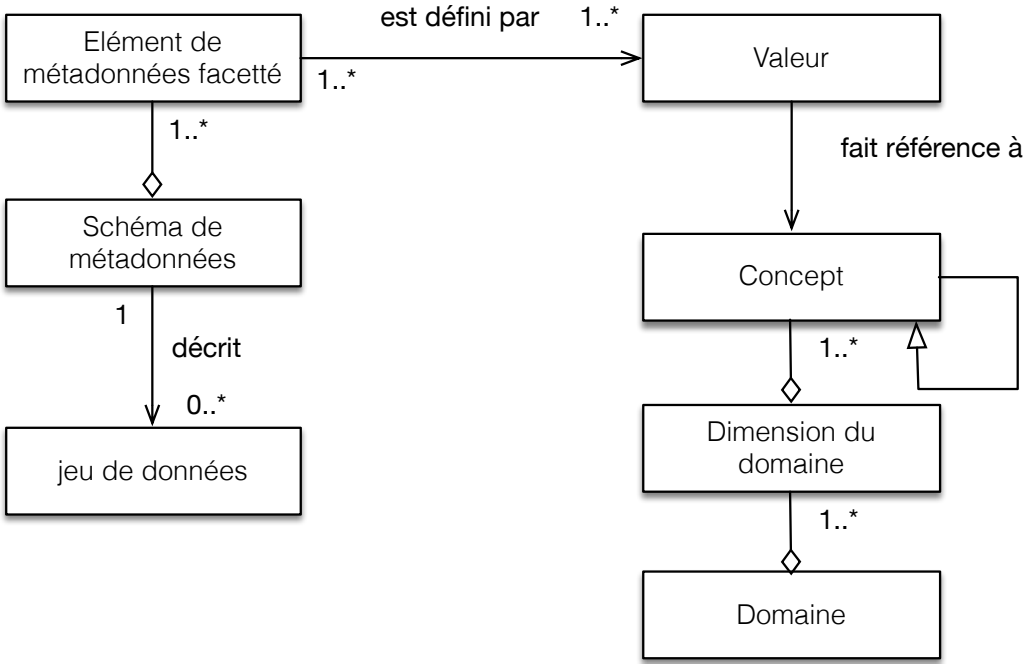


Figure 2.10: Diagramme de classe UML formalisant la notion de métadonnées facettées

Adaptation du vocabulaire à l'utilisateur

L'adaptation du vocabulaire fait partie des verrous traités par la recherche à facettes mais continue à en constituer un enjeu non négligeable. Comme le souligne (Furnas et *al*., 87) : « les mots clés qui sont choisis par les concepteurs de SRI sont souvent éloignés de ceux des utilisateurs ». Si la navigation par facettes atténue ces décalages sémantiques en guidant les utilisateurs par une découverte du vocabulaire et une construction progressive de leur requête, il n'en demeure pas moins que le vocabulaire mis à disposition n'est pas orienté utilisateur. Des travaux portant sur l'adaptation du vocabulaire au point de vue de l'utilisateur existent. Nous les avons évoqués précédemment autour des techniques de reformulation de requêtes. Ces techniques pourraient être mise en œuvre parallèlement à une navigation par facettes mais elles en diminueraient alors l'intérêt en réintroduisant une formulation de requête en « aveugle ». Ce verrou, et c'est un des éléments des propositions présentées dans ce mémoire, est l'objet de nos travaux. En effet, nous préconisons la construction d'une classification orientée utilisateur. Pour cela, le paradigme d'Observation nous semble suffisamment général et générique pour permettre de produire des classifications adaptées à une large palette d'applications de recherche sur les données spatio-temporelles en environnement.

Ergonomie des IHM

D'autres points touchant à l'utilisabilité des interfaces de recherche, comme la quantité d'information assimilable pour un utilisateur, les liens visuels entre les éléments de raffinement et les résultats qui leurs sont associés, requiert également une attention particulière. Nous abordons point par point ces aspects, dans la section recherche à facettes sur les données spatio-temporelles de notre chapitre proposition.

2.3.5 Intérêts et axes de réflexion pour l'adaptation à la recherche d'information spatio-temporelle

Comme l'a montré (Yee et *al*., 03), par son interactivité et sa capacité à guider l'utilisateur dans un domaine de connaissance non totalement maitrisé, la recherche à facettes emporte l'adhésion d'utilisateurs peu familiers, voire réfractaires aux applications de recherche. Les applications de recherche du e-commerce ont depuis déjà plusieurs années démontré leur efficacité. Les précurseurs en la matière ont été eBay puis Amazon. Aujourd'hui, cette approche de recherche est très répandue sur les sites de vente en ligne car la recherche de produits s'y prête particulièrement. Il est effectivement aisé d'organiser la description de produits en la structurant autour de métadonnées facettées. Les facettes consistent alors à présenter les grandes caractéristiques des produits à partir d'une nomenclature comprise par tous, au moins quand il s'agit de produits de consommation courante (disque, vêtement, chaussure, livre).

Notre contexte applicatif peut facilement être comparé à celui du e-commerce. En effet, les jeux de données à rechercher sont semblables à des produits manufacturés. Ils sont avant tout décrits à partir de leurs caractéristiques plutôt que par l'information qu'ils véhiculent. Par ailleurs, ils sont par construction accompagnés de métadonnées qui décrivent tout ou partie de leurs caractéristiques. La recherche sur les différentes caractéristiques des images sont autant de dimensions sur lesquelles l'utilisateur pourra être guidé pour aboutir aux jeux de données recherchés.

Pour mettre en œuvre, de manière efficace, une recherche à facettes sur les données spatio-temporelles, des travaux touchant à l'interopérabilité des métadonnées, aux choix des dimensions de recherche ainsi qu'à la mise en relation des référentiels terminologiques existants doivent être menés. En effet, les hétérogénéités structurelles et sémantiques des métadonnées qui accompagnent les jeux de données demandent à être considérées et traitées. En particulier, l'hétérogénéité sémantique des métadonnées qui s'observe autant au niveau des schémas que des valeurs, le fossé sémantique existant entre les nomenclatures « producteurs » et celles des « utilisateurs » demandent à être réduites afin de pouvoir constituer un index de recherche uniforme et cohérent d'un point de vue structurel et sémantique. La réduction du fossé fait également partie des efforts à entreprendre si l'on veut rendre sémantiquement accessible les dimensions de recherche à un utilisateur peu familier du domaine. Pour cela, notre souhait est de tirer parti des nombreux référentiels de valeurs : thématiques, spatiaux pour proposer une classification à facettes spécifique à notre domaine.

3. Propositions autour d'un système de recherche à facettes pour les images satellitaires

3.1 Introduction

Les travaux que nous avons mené de 2006 à 2012 (Boisson et *al*., 06 ; Desconnets et *al*., 07 ; Sayah, 11) avaient déjà investi les questions posées autour de l'aide à apporter à un utilisateur dans ses activités de recherche d'information. Leurs mises en œuvre dans deux services de catalogage, SYSCOLAG[44] et REFLECS[45], n'ont répondu que partiellement aux attentes des usagers.

Depuis lors, plusieurs avancées méthodologiques et technologiques ont permis d'envisager un traitement plus satisfaisant de la recherche d'information environnementale tant du point de vue des annotations sémantiques à partir des métadonnées que du point de vue de l'utilisabilité des applications de recherche. Tout d'abord, la maturité grandissante du web sémantique a été déterminante. Nous pensons plus particulièrement au langage d'organisation de la connaissance SKOS (W3C, 04). Nous pensons également aux initiatives de partage de connaissances qui permettent aujourd'hui d'avoir accès, via des services web, à de nombreuses bases terminologiques et toponymiques dans des domaines très variés. D'autre part, la richesse fonctionnelle, l'implémentation de plusieurs stratégies de recherche et la performance des moteurs de recherche open source actuels permettent d'envisager des applications de recherche supportant une interaction rapide, voire immédiate entre exécution des requêtes et affichage des résultats, et cela sur de très grandes collections de données. Enfin, les travaux menés autour de l'interconnexion des catalogues de données (Desconnets et *al*., 14, Mougenot et *al*., 15 ; Desconnets et *al*., 17b), nous permettent aujourd'hui d'appuyer la recherche d'information sur des jeux de métadonnées interopérables au sein d'architectures distribuées et ainsi envisager leur traitement automatique par lot.

Notre proposition vient se nourrir de ces diverses avancées et s'est concentrée sur deux points. Le premier point consiste à définir les différentes dimensions de recherche (ou facettes) et une classification *adhoc* sur laquelle viendra s'appuyer la recherche d'images satellitaires. Le deuxième point met l'accent sur l'indexation des descriptions des données spatio-temporelles à partir de standards de métadonnées. Il s'agit d'utiliser les référentiels terminologiques « métier », les référentiels toponymiques pour traiter à la fois les hétérogénéités sémantiques inhérentes aux origines diverses des métadonnées et de réduire le décalage entre le vocabulaire « utilisateur » et « producteur ».

L'exemple de la nomenclature décrivant le niveau de traitements d'une image illustre ce décalage. Lorsque les images ortho rectifiées produites par AIRBUS sont définies par le terme ORTHO, celles produites par l'USGS sont définies par le terme L1T. Ces deux nomenclatures ne sont pas de prime à bord accessibles à un utilisateur peu familier du domaine. La diversité des nomenclatures, une par producteur, en réduit d'autant l'accessibilité. Les référentiels « utilisateurs » viennent concrétiser la classification à facettes envisagée. La finalité est de proposer une IHM qui assure la découverte, l'exploration des catalogues d'images en fournissant une navigation intuitive, dynamique et ouverte à différents points de vue (les facettes). Il s'agit de permettre à l'utilisateur de s'affranchir, lorsque cela est possible, de la maîtrise du domaine sur lequel porte les données et de pallier aussi les frustrations générées par le côté, par trop restrictif, des recherches booléennes.

---

[44] SYSCOLAG : Programme Systèmes Côtiers et Lagunaires

[45] REFLECS : Catalogue de références des missions spatiales du CNES (REFerence catalogue of Long term CNES Scientific data)

Après avoir rappelé les profils des utilisateurs cibles qui accèdent aux infrastructures de données spatiales, le postulat de recherche de l'information spatio-temporelle est posé. Nous présentons ensuite le cœur de notre proposition : la classification à facettes relative à la recherche d'images satellitaires. Nous rappelons ensuite l'apport des profils d'application comme support interopérable à la structuration des métadonnées sur lequel nous envisageons de poser les différentes dimensions de recherche. Nous décrirons ensuite les mécanismes d'alignement envisagés sur les thésaurus pour traiter les « décalages » sémantiques au sein des métadonnées d'une part, leur enrichissement par l'ajout de toponymies d'autre part. Enfin, nous détaillerons la mise en œuvre de cette approche en nous appuyant sur l'exemple de l'infrastructure de données spatiales GEOSUD. Les propositions portant sur la classification à facettes ainsi que les approches d'indexation guidées par les thésaurus sont des travaux originaux. Ils n'ont pas encore fait l'objet de publications scientifiques.

3.2 Contexte

3.2.1 Les utilisateurs cibles

Un des objectifs affichés des infrastructures de données spatiales est d'élargir la diffusion des données au plus grand nombre d'utilisateurs en vue de mieux répondre aux politiques publiques en lien avec la gestion de l'environnement. Pour cela, les applications de découverte de ces plateformes, qui en sont le point d'entrée, sont destinées à être manipulées par un public varié, de culture et d'expertise toutes aussi variées. En lien avec les objectifs affichés, la majorité des utilisateurs des infrastructures de données spatiales est issue de la sphère publique (Georis-Creuseveau, 14), et du milieu associatif et dans un moindre part de la sphère privée et du citoyen. Aussi, nous nous concentrons sur les besoins en information des acteurs publics. En réalité, cette dénomination recouvre une variété d'utilisateurs que l'on peut séparer en deux grands groupes :

- *Ceux issus de la recherche académique et de l'enseignement supérieur*. Ils mènent des études sur des problématiques environnementales ou sur celles relatives au traitement de l'image. Ils sont à mêmes d'appréhender toute ou partie de la complexité de la donnée pour en extraire les informations utiles à leur besoin,
- *Les agents des services de l'état et des collectivités locales ou territoriales*. Ils répondent à des missions demandant, par exemple, la réalisation de produits cartographiques en vue de suivre et contrôler l'application de décisions gouvernementales ou assurer la gestion locale de l'environnement (e.g. gestion des déchets, biodiversité). Dans leur grande majorité, ces utilisateurs ont une faible expertise en matière de télédétection.

3.2.2 Postulats pour la recherche d'images satellitaires

Nous n'avons pas formellement mis en relation une stratégie de recherche spécifique à chaque type d'utilisateur identifié mais l'expérience accumulée au cours des divers projets menés (SYSCOLAG, ROSELT, NatureSDI+, GEONetCab, EOPOWER, GEOSUD, THEIA) permet de donner quelques pistes sur le cheminement d'un utilisateur en quête de données spatio-temporelles. Quatre critères principaux sont utilisés pour construire une recherche sur des données spatio-temporelles. Nous distinguons dans l'ordre :

- ✓ **Le critère Où ?** porte sur l'étendue spatiale de la donnée. Ce critère peut être exprimé par un nom de lieu, un rectangle englobant ou une géométrie quelconque. Dans le cas d'une

image, l'étendue spatiale est le plus souvent exprimée par son enveloppe spatiale (ou rectangle englobant) ;

- ✓ **Le critère Quand ?** porte sur la date de création de la donnée ou son étendue temporelle. Ce critère correspond à l'instant ou la période d'acquisition ou de création de la donnée. L'étendue temporelle correspond à une période. Concernant une image, l'étendue temporelle correspond au temps, très court, qui a été nécessaire à l'acquisition de l'image ;

- ✓ **Le critère Quoi ?** porte sur le contenu de la donnée. Comme précisé dans l'état de l'art, la description d'une image ne porte pas sur la nature de la surface terrestre observée mais sur ses caractéristiques et ses conditions d'acquisition. Aussi, la formulation du critère Quoi ? ne peut pas être exprimée à travers un champ thématique (biodiversité, foresterie, littoral), comme cela peut être le cas dans des catalogues portant sur des données élaborées ;

- ✓ **Le critère Comment ?** porte sur les conditions d'acquisition ou de traitements selon que l'on adresse ce critère sur des données observées ou dérivées après application d'un traitement « métier ». Les conditions d'acquisition ou de traitements peuvent être très nombreuses. Elles requièrent le plus souvent une expertise technique pour appréhender leur signification et les mettre en lien avec le besoin d'information recherchée.

Il ressort de cette analyse et de notre expérience que les critères Où et Quand sont discriminants et ne sont pas liés à une quelconque expertise du domaine pour être utilisés. Ce qui n'est pas le cas des critères Quoi et Comment. Pour cela, il nous a semblé pertinent de distinguer deux grands scénarii de recherche, relatifs à cette dichotomie :

- *Un recherche non experte* pour laquelle il est nécessaire de faciliter la formulation des critères les plus discriminants et qui ne demande pas d'expertise métier (Où ? et Quand ?). En effet, ces critères sont avant tout associés à l'objet de l'étude (son lieu et la période ou l'étendue temporelle pertinente pour l'étude). On pourra également mettre à portée de l'utilisateur certaines caractéristiques de l'image qui permettent de juger de son adéquation vis à vis de l'usage envisagé, comme par exemple, la précision et la qualité géométrique relative à la taille du pixel et au niveau de production de l'image.

- *Une recherche experte* dans laquelle d'autres caractéristiques de l'image pourront être utilisées pour préciser la recherche telles que la plateforme, l'instrument de mesure, la qualité radiométrique, les longueurs d'ondes, …

La figure 3.1 positionne les différents critères de recherche selon deux axes : le temps passé à formuler une requête, qui est en lien avec l'expertise de l'utilisateur et le niveau d'expression ou profondeur de la recherche. Nous faisons apparaître les propriétés pour chaque critère les caractéristiques correspondantes aux images satellitaires.

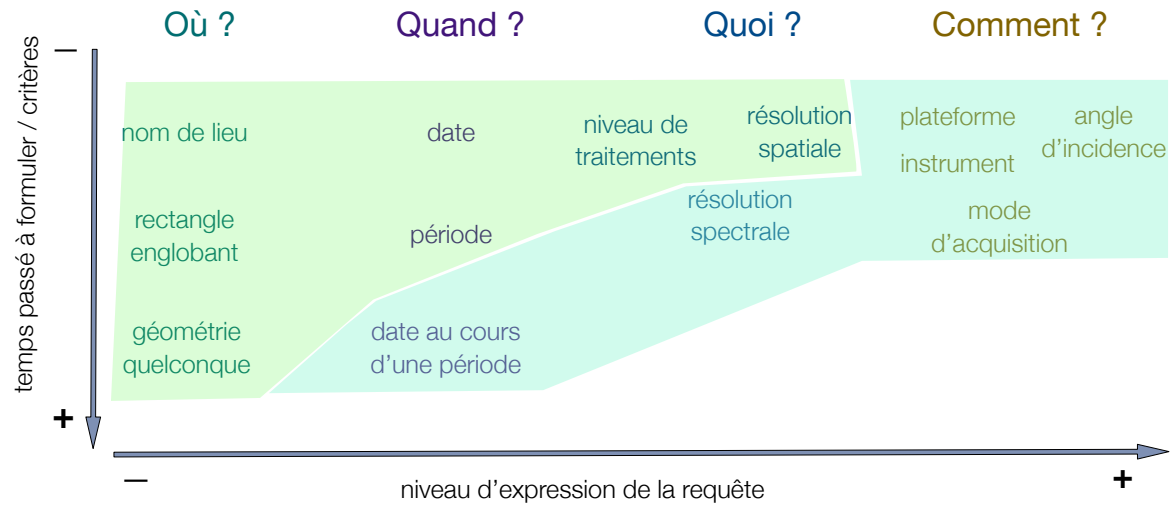


Figure 3.1: Les principaux critères de recherche et leur relation avec le niveau d'expression de la requête (abscisse) et le temps passé pour la formuler (ordonnée). L'enveloppe de couleur vert clair englobe les critères utilisés par des utilisateurs néophytes. L'enveloppe de couleur vert foncé englobe des critères plus experts.

3.3 Classification à facettes pour la recherche d'images satellitaires

3.3.1 Classification à colonnes

La classification à colonnes proposée par (Ranganathan, 33) vient remédier aux limites d'une classification hiérarchique ou énumérative. Ainsi, elle peut répondre aux besoins de classification sur des domaines complexes, interdisciplinaires. En effet, la classification à colonnes propose de naviguer dans un corpus de connaissance en utilisant plusieurs dimensions (plusieurs points de vue) quand une classification hiérarchique repose sur un point de vue qui se concrétise par un chemin d'accès unique à ce corpus. Les colonnes, ou facettes, correspondent ainsi aux différentes dimensions du corpus de connaissances. Ils sont autant de chemins pour parcourir les termes ou les concepts d'un domaine. Dans le domaine de la gestion documentaire, plusieurs travaux s'en sont inspirés notamment ceux de (Austin, 84) qui a mis au point un système d'indexation PRECIS utilisé par British National Library. Bien avant, (Aitchinson et *al*., 69) l'a utilisé pour développer les premiers thésaurus dans le domaine de l'ingénierie et des sciences (TEST). (Mills et Broughton, 77) quant à eux, ont mis à jour et amélioré un des systèmes classification utilisés par les libraires, *Bliss Bibliographic Classification*, en y introduisant les principes de la classification à facettes.

Le modèle PMEST

Dans la 3ème édition de la classification à colonnes, (Ranganathan, 67) rattache ses colonnes à des concepts de très haut niveau. Il propose cinq catégories fondamentales pour guider la création de schémas de classification : *Personnality Matter Energy Space* and *Time* : Personnalité, Matière, Energie complétée par le Temps et Espace. On parle également de la formule ou du modèle PMEST. Il s'agit de fournir un cadre pour généraliser l'application d'une classification multi dimensionnelle à divers domaines ou disciplines. PMEST peut alors être vu comme un modèle d'organisation d'un corpus de connaissance. En ingénierie de l'information, nous parlerions de méta modèle. Tous les concepts d'un domaine sont organisés autour des cinq catégories. Si le fort degré d'abstraction de cette

proposition soumet son utilisation à une forte interprétation de la part du concepteur, il ressort des écrits de (Raganathan, 52), et de la bibliographie (Brougthon, 06 ; Kumar, 74) parcourue, des lignes directrices pour interpréter et utiliser ce modèle. La facette *Personnality* (P) est définie comme l'essence même de la discipline. Ses concepts clés peuvent être rattachés à cette dimension. Relèvent, par exemple, de la facette *Personnality* : en médecine, les organes; en bibliothéconomie, les types de bibliothèques; en art, les styles. La facette *Matter* (M) se réfère aux caractéristiques du domaine et vient compléter la dimension *Personnality*. La facette *Energy* (E) inclut ce que l'on peut rattacher aux activités et aux actions liées au domaine. En d'autres termes, la facette *Energy* se réfère à l'action et aux facteurs qui sont responsables de l'action. *Energy* représente ainsi les aspects dynamiques rattachés au domaine. Les facettes *Time* (T) et *Space* (S) viennent compléter les 3 précédentes et précisent respectivement les dimensions géographiques et temporelles associées aux concepts du domaine lorsque cela est pertinent. Le tableau ci dessous éclaire, à travers quelques exemples rencontrés dans la littérature, l'utilisation du modèle PMEST. Chaque catégorie peut ensuite être déclinée en un ensemble de concepts ou facettes. Un concept peut être également décrit à travers ses propriétés. C'est d'ailleurs cette dernière déclinaison qui est le plus souvent rencontrée sur le web (Ellis et Vasconcelos, 00).

Tableau 3.1 : Exemples de catégories fondamentales pour la classification à facettes selon le modèle PMEST (Maniez, 99)

Catégorie fondamentale	**Personnalité**	**Matière**	**Energie**
Médecine	poumon	tuberculose	traitement thérapeutiques
Agriculture	riz	maladie virale	éradication
Enseignement	aveugles	jeu de rôles	éducation

Bien que ce modèle d'organisation ne soit pas aisé à manipuler, il constitue néanmoins un cadre de conception et un point de départ. Notre approche de classification vise à adapter ses grands principes aux données spatio-temporelles et plus particulièrement à celui des images satellitaires. Aussi, notre objectif est de transposer le modèle PMEST à notre domaine en nous appuyant sur une vision centrée utilisateur. In fine, il s'agit, d'identifier les catégories fondamentales puis de les décliner en *n* catégories et sous-catégories sur lesquelles il sera pertinent de construire une classification adaptée à la découverte et à la navigation dans les collections d'images. Cette construction doit bien évidemment s'appuyer sur les recommandations et prescriptions émises par les institutions de normalisation, OGC et ISO TC/211 du domaine de l'information géographique. Elles sont largement utilisées par les plateformes sur lesquelles nous souhaitons mettre en œuvre notre proposition et seront pour cela aisées à mettre en oeuvre.

Ces dernières années, de nombreux standards ont vu le jour pour structurer, échanger des jeux de données spatio-temporelles ou leur métadonnées au sein d'infrastructures de capteurs ou de données. Aussi, nous serions tentés d'exploiter ceux portant sur la définition des métadonnées notamment les standards ISO 19115 ou ISO 19115-2 relatifs à la description de l'information géographique et son extension pour les données matricielles. Mais leur portée applicative restreinte, c'est leur objectif, aux données spatiales, ne nous permet pas de fournir une vision suffisamment abstraite pour traiter les données spatio-temporelles dans leur ensemble. Par ailleurs, la vision associée à ces standards est centrée sur les besoins de structuration et d'interopérabilité inhérentes aux producteurs de données et non à ceux des utilisateurs.

3.3.2 Le méta modèle O & M comme modèle d'organisation

Le paradigme d'Observation proposé et formalisé par l'OGC au travers du standard *Observations and Measurement* (OGC, 07) a attiré toute notre attention. En effet, il apporte une vision centrée utilisateur. Bien qu'il soit moins universel que le modèle PMEST, le modèle *d'Observation & Measurement* présente un niveau d'abstraction suffisamment élevé pour pouvoir prendre en considération la majeur partie des descriptions des données spatio-temporelles acquises ou produites pour le suivi et la gestion de l'environnement. Le diagramme de classe UML en figure 3.2 présente le méta modèle du standard O & M.

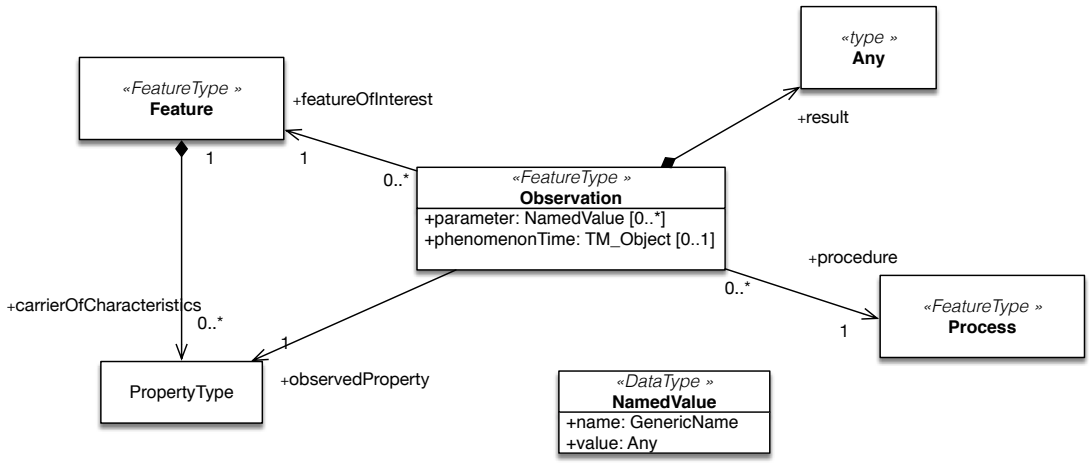


Figure 3.2: Méta modèle simplifié du standard *Observations and Measurements* (OGC, 07).

Présentation du méta modèle O & M

Le paradigme d'Observation tel que défini par l'OGC peut être énoncé comme suit : « une **observation** est une action dont le **résultat** est la valeur d'une **propriété** d'une **entité ou objet d'intérêt** à un **moment** donné, obtenu grâce à une **procédure**». Nous explicitons les principaux concepts autour de la notion d'Observation pour argumenter ensuite le choix des dimensions de notre classification à facettes.

(Fowler, 88) définit une observation comme une action associée, réalisée durant un instant ou une période de temps donnée, au cours de laquelle un nombre, un terme ou tout autre symbole est associé à un phénomène. (National Research Council, 95) définit quant lui un phénomène comme une propriété d'un objet identifiable qui correspond à l'entité d'intérêt sur laquelle porte l'observation. L'observation utilise une procédure pour produire une valeur estimée du phénomène observé. Il est à noter que le terme Observation est apparu que très récemment et a permis de faire une distinction entre cette notion et celle de mesure (Nieva, 01 ; Yoder et *al.*, 00). En effet, en métrologie, c'est généralement le terme mesure qui est utilisé (Sarle, 95 ; Vocabulaire International de Métrologie, 12). Il est réservé dans le cas où le résultat est quantitatif et dont le protocole s'appuie sur un instrument de mesure autre qu'un humain.

Le résultat ou *result* d'une Observation est une estimation de la valeur d'une propriété de l'objet d'intérêt. Les autres propriétés de l'Observation : *featureOfInterest*, *observedProperty*,

parameter et procedure fournissent le contexte de l'Observation pour assurer la découverte, l'interrogation puis l'évaluation et l'interprétation des résultats.

L'entité d'intérêt ou *featureOfInterest*, ou encore objet d'intérêt, est une entité ou *Feature* de type quelconque. La notion d'entité ou *Feature* est une notion centrale pour la modélisation de l'information géographique. Elle est définie par l'ISO 19101 (ISO, 02a) et ISO 1909 (ISO, 05) comme une représentation abstraite du monde réel. Elle peut être, entre autres, caractérisée par zéro ou plusieurs propriétés spatiales ou *Spatial attribute,* zéro ou plusieurs propriétés temporelles *Temporal attribute* ou bien encore zéro ou plusieurs propriétés thématiques ou *Thematic attribute*. Une entité est donc située dans l'espace et dans le temps comme un bâtiment, une rivière, une tempête, une forêt. Aussi, du point de vue de l'observation, une entité d'intérêt est une entité qui correspond à la cible de l'observation ou l'objet réel sur lequel l'observation est réalisée. *Thematic attribute* d'une entité ou *Feature* proposé par ISO 19109 correspond à la classe *PropertyType* O & M. Le diagramme de classe UML en figure 3.3 issu de l'ISO 19109 formalise cette notion de *Feature* en explicitant ses différents types d'attributs.

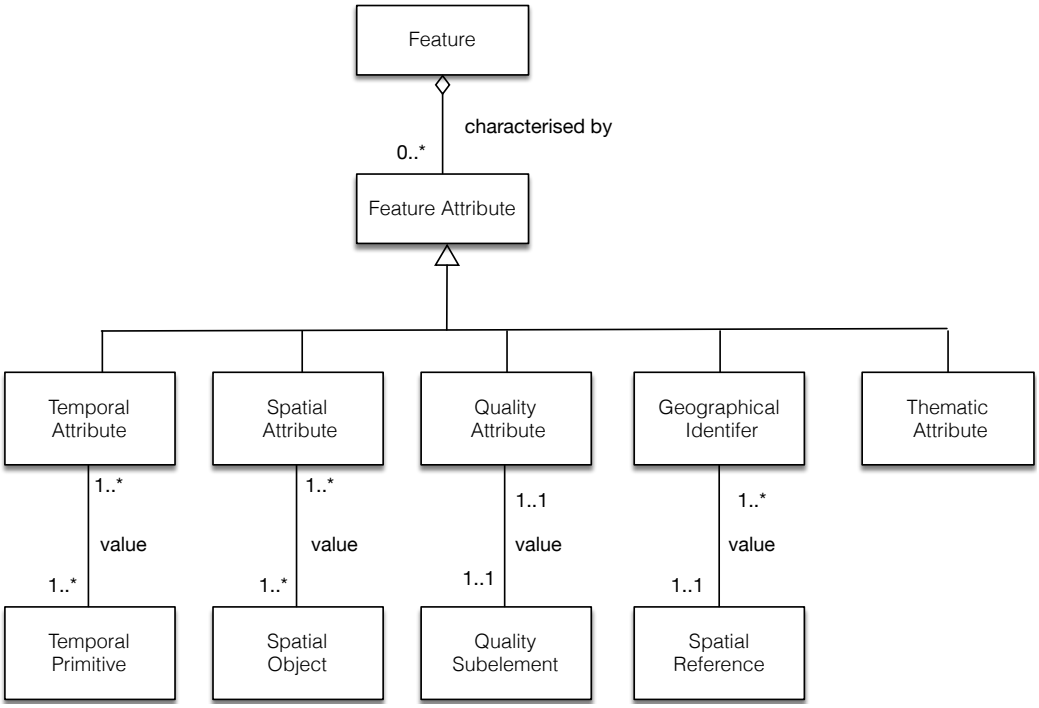


Figure 3.3: Formalisation de la notion de *Feature* - ISO 19109 (ISO, 05)

La propriété observée ou *observedProperty* identifie ou décrit le phénomène pour lequel l'observation effectuée fournit une valeur estimée. Cette propriété est obligatoirement une des propriétés de l'objet d'intérêt observé.

Un paramètre d'observation *parameter* est un paramètre associé à un acte d'observation. Il est généralement utilisé pour enregistrer les conditions environnementales dans lesquelles est effectuée l'observation. Enfin, le concept d'Observation est également caractérisé par une propriété qui définit son domaine temporel dont le type pourra être un instant ou une période. O & M distingue deux moments: *samplingTime* ou date d'échantillonnage qui correspond à l'instant ou la période où le résultat s'applique à l'objet d'intérêt. Le *resultTime ou* date d'acquisition est l'instant ou la période où l'observation a été effectuée. Souvent, ces deux temps sont identiques et dans ce cas, ce dernier n'est pas indispensable pour caractériser une observation.

La procédure ou *procedure*, ou encore protocole d'observation, est la description du traitement utilisé pour produire un résultat sur une propriété observée de l'objet d'intérêt. Elle est souvent assurée par un instrument ou un capteur mais peut aussi être une chaîne de traitements, un observateur humain, un algorithme ou un programme ou une simulation numérique. La figure 3.4 donne un exemple trivial d'instanciation du modèle O & M. Il s'agit de la mesure horaire de température de l'air fourni par la station météorologique située à Fréjorgues. L'objet d'intérêt est la station météorologique de Montpellier - Fréjorgues dont la position géographique est 43.58° N, 3.96° E. La propriété mesurée est la température de l'air. La procédure est le thermomètre à minimum et à maximum. Les conditions d'observation sont sous abri (station de type Stevenson) à 1.5 m du sol. Pour la journée du 4 mai 2017 à 16h00, la température observée est de 17° Celsius correspondant respectivement au *resultTime* et au résultat lui-même.

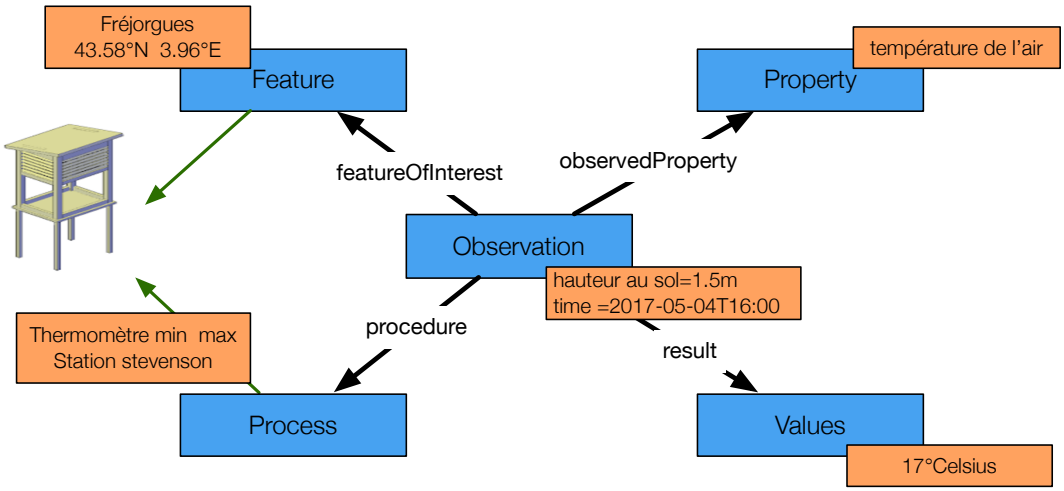


Figure 3.4: Schéma présentant un exemple d'instanciation du modèle O & M. Mesure de la température de l'air.

Les propriétés clés caractérisant le concept d'Observation ont l'intérêt de mettre en avant les éléments de contexte qui sont essentiels pour l'utilisateur dans ses activités de recherche d'information à savoir :

- l'objet de l'observation *featureOfInterest* pour laquelle nous distinguons ses propriétés spatiales S*patial attribute* (cf. figure 3.3) et ses propriétés *observedProperty ou thematic attribute*,
- les propriétés associées à l'événement de l'observation dont les conditions d'acquisition *parameter,*
- le moment de l'observation *samplingTime* ou *resultTime* et enfin,
- la procédure ou protocole d'observation, *procedure*, relatif aux méthodes et aux instruments utilisés.

Ces propriétés constituent les catégories fondamentales à fournir à un utilisateur pour découvrir et parcourir des collections de données spatio-temporelles. La propriété *result*, quant à elle, ne constitue pas une dimension de classification mais l'information souhaitée par l'utilisateur. Le schéma en figure 3.5 représente ces catégories et les rattachent à celles du modèle PMEST de Raganathan. Le sujet ou l'essence de l'observation ou *Personnality* est l'objet d'intérêt ou *FeatureOfInterest*. Les propriétés

observées ou *observedProperty* correspondent à la dimension *Matter*. Tout ce qui touche à la procédure est rattachée à la dimension *Energy*. *L*a dimension *Time* correspond au *samplingTime* ou *resultTime*. Enfin, la dimension *Space*, correspond à la spatialité de l'objet ou *spatial attribute*.

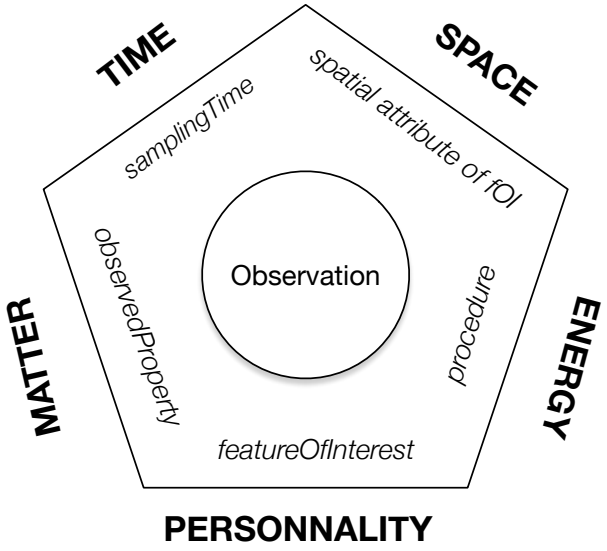


Figure 3.5: Les cinq catégories fondamentales pour classifier les données spatio-temporelles. Elles sont projetées sur le modèle PMEST

3.3.3 Application à l'Observation de la Terre

Le modèle O & M offre le cadre pour formaliser les activités d'Observation dont résultent les images satellitaires. En effet, elles sont le résultat de l'Observation d'une portion de la Terre dont la prise de vue a été effectuée à un instant donné grâce l'utilisation d'un instrument. Ces derniers sont embarqués sur un satellite qui décrit une orbite autour de la Terre. Le propriété *featureOfInterest* correspond à la portion de la terre photographiée, la propriété observée *observedProperty* est le rayonnement réfléchi par la surface de la Terre ou ses composantes. Les paramètres *parameter* d'observation ou conditions d'observation sont nombreux, comme par exemple l'angle de prise de vue, l'ennuagement, le rayonnement diffus de l'atmosphère, la hauteur du soleil… La procédure est liée à la plateforme et aux instruments de mesure qui y sont embarqués. Le résultat de l'observation *result* est une ou plusieurs matrices de valeurs géophysiques correspondant à une ou plusieurs composantes des ondes électromagnétiques émises par la surface de la terre.

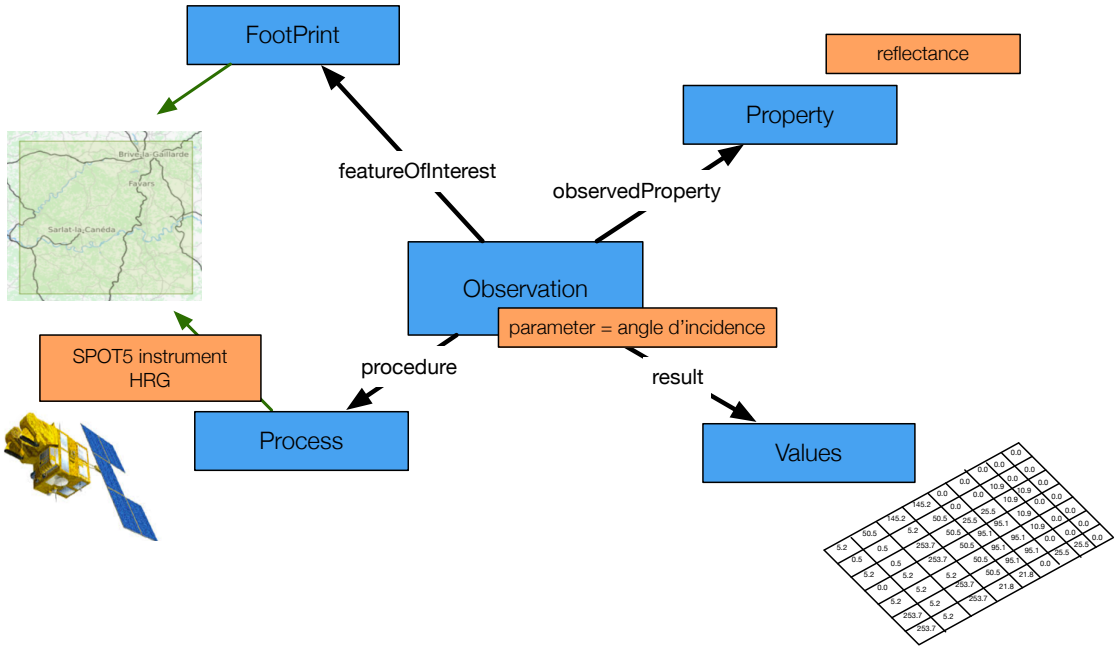


Figure 3.6: Schéma présentant un exemple d'instanciation du modèle O & M pour l'observation de la Terre. Cas de l'observation de la Terre par le satellite SPOT5 et l'instrument HRG

Comme le recommande le standard ISO 19106 (ISO, 04) relatif à la déclinaison d'un standard, les travaux de (Gasperi et *al*., 12), propose une spécialisation du modèle O & M (figure 3.7) en schéma de métadonnées pour le domaine de l'observation de la Terre. Cette proposition permet de raffiner les propriétés du modèle O & M pour tenir compte des spécificités de l'observation de la Terre. Les spécialisations proposées *OpticalSatelliteObs ::OM_Observation* et *OpticalSatelliteObs ::Process* des classes *Observation* et *Process* viennent enrichir les informations concernant la description du protocole de mesure en explicitant les propriétés associées aux conditions d'acquisition *acquisitionParameter*, aux instruments de mesure (*plateform*, *instrument*, *sensor*) et celles portant sur les propriétés de l'Observation en proposant une classe *EarthObservationMetadata*. Cette dernière décrit les informations sur les post-traitements qui suivent l'acquisition de la mesure. Ils sont très utiles pour connaître la qualité de l'image produite, notamment ses niveaux de qualité en géométrie et en radiométrie. Bien que le modèle O & M, comme celui spécialisé pour l'Observation de la Terre soit centré utilisateur, il est mis en œuvre par les agences spatiales, l'ESA dans le cas présent. Il est donc instancié à travers le point de vue des producteurs. Aussi, les différentes propriétés sont valuées à partir des nomenclatures de ces producteurs. L'exemple fournit dans la section 2.2.2 (Métadonnées pour les images satellitaires), les codifications relatives au niveau de traitements d'une image, ou encore les identifiants des appareils de mesure (plateforme, instrument) sont abscons pour des utilisateurs non familiers de la télédétection.

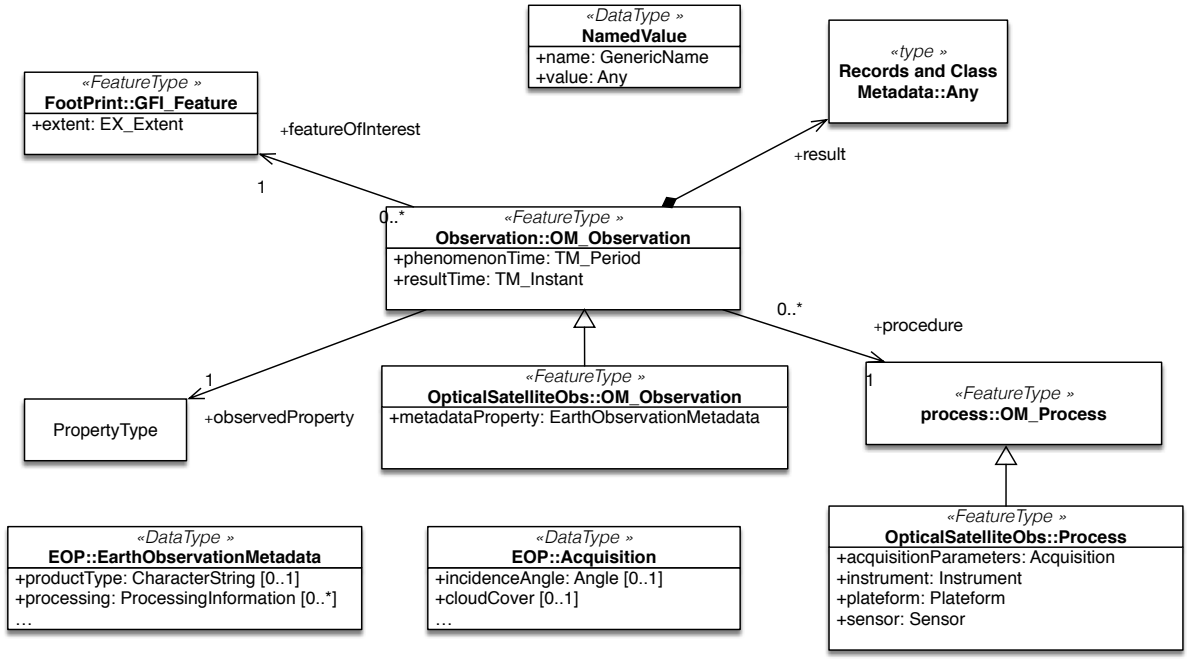


Figure 3.7: Instanciation partielle du méta modèle O & M pour l'observation de la Terre. Cas d'observation par capteur optique

Comme nous l'avons proposé pour le modèle général O & M, nous déclinons les cinq catégories qui nous semblent fondamentales pour classifier les images satellitaires à partir des spécialisations proposées par les travaux de (Gaspéri et *al*., 12). La figure 3.8 spécialise pour l'observation de la Terre les cinq catégories retenues. L'objet d'intérêt ou *FeatureOfInterest* est la Terre ou plus précisément la portion de la Terre observée lors d'une prise de vue. Les propriétés observées ou et leur déclinaison : réflectance, transmittance… correspondent à la dimension *observedProperty*. Les instruments de mesure (*plateform*, *instrument*, *sensor*) et les conditions (*acquisitionParameter*, *EarthObservationMetadata*) dans laquelle elle est effectuée sont rattachées à la dimension *procedure*. La dimension *Time* correspond au *phenomenonTime* également appelée date/heure d'acquisition. Enfin, la dimension *Space*, comme nous l'avons exprimé précédemment, est le *Footprint* ou emprise au sol de l'image. Elle explicite la propriété spatiale *spatial attribute* de l'objet observé car c'est également un des éléments de contexte indispensable à un utilisateur pour rechercher des images.

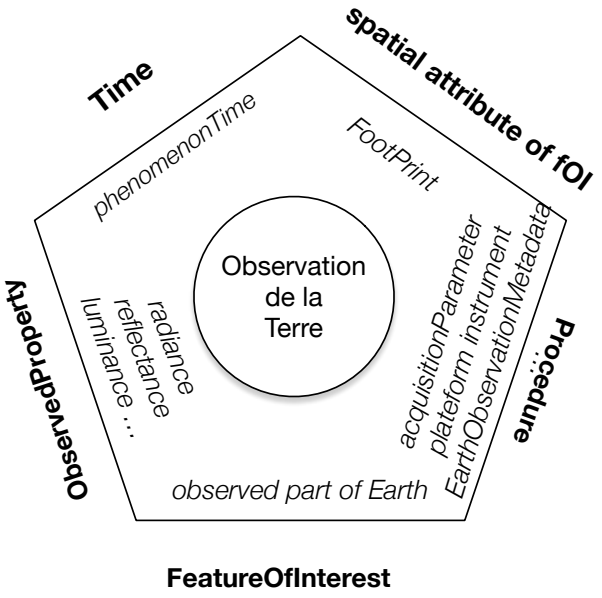


Figure 3.8 : Les catégories et facettes retenues pour la classification à facettes. Application à l'imagerie optique.

En nous appuyant à la fois sur les travaux de (Gaspéri et *al*., 12) et les standards ISO 19108 (ISO, 02b) et ISO 19115 (ISO, 03) qui fournissent respectivement des formalisations des propriétés temporelle, spatiale et des métadonnées, nous précisons catégorie par catégorie les concepts et sous concepts clés qui nous semblent nécessaires pour organiser une classification des images satellitaires optiques.

Dimension FeatureOfInterest

En télédétection, l'objet d'intérêt est déterminé par le point de vue adopté par celui qui utilise l'observation. Son point de vue guide l'interprétation de la valeur géophysique obtenue en vue d'identifier, délimiter spatialement un ou plusieurs objets d'intérêts et cela en fonction de leur signature spectrale. Le plus souvent, cette interprétation s'appuie sur une classification qui est propre à un champs disciplinaire (hydrographie, géologie) ou à une activité humaine (agriculture, forêt, urbanisation, infrastructures). Pour cela, l'observation de la Terre est une observation particulière pour laquelle il peut y avoir autant d'objet d'intérêts que de point de vues. Ce qui est très différent d'une observation, par exemple, de la hauteur d'eau d'un fleuve où quelque soit le point de vue, l'objet observé est toujours le même. La figure 3.9 schématise notre vision et associe à un point de vue une ou plusieurs classifications de référence des champs disciplinaires ou inter disciplinaires avec lequel la communauté d'utilisateurs souhaite utiliser les images. Néanmoins, il est actuellement délicat de mettre en correspondance, sans expertise humaine et lors de la phase de post-traitement de l'image, la valeur géophysique d'une unité élémentaire de l'image à une classe d'objets d'intérêt que l'on souhaite étudier. Seules les images ayant subi de tels traitements, supervisés par un expert, pourront ensuite faire l'objet d'une description des classes contenues dans l'image, ce qui est dans les faits encore rarement le cas. Divers travaux de recherche explorent les possibilités qu'offrent les techniques d'induction ou de déduction de la connaissance pour proposer une traitement automatique de l'interprétation des valeurs géophysiques (Arvor et *al*., 13) ou d'images classifiées (Bayoudh et *al*., 15 ; Toulet et *al*., 17). Mais ces travaux demandent à être optimisés pour pouvoir être mis en œuvre sur

des images haute résolution. Pour cela, cette dimension, bien que essentielle pour la classification des images, est aujourd'hui délicate à mettre en œuvre. Elle demeure néanmoins un enjeu majeur pour faciliter l'accès aux images.

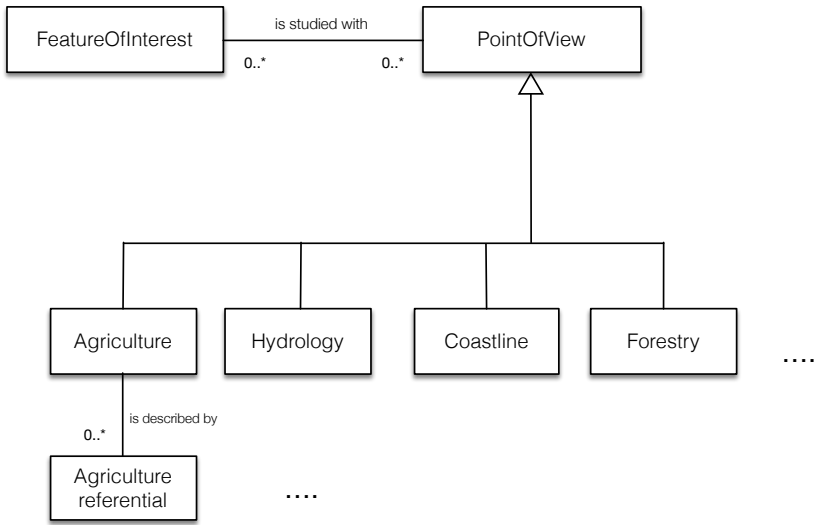


Figure 3.9 : Objet d'intérêt selon les différents points de vue associés à l'utilisation d'une image satellitaire. Ces points de vue ne sont pas exhaustifs mais sont donnés à titre d'exemple.

Dimension ObservedProperty

Bien qu'en télédétection optique, le résultat de l'observation soit une grandeur physique qui correspond au rayonnement réfléchi par la surface de la Terre, ce sont des rapports liés à cette grandeur qui sont utilisés pour analyser, interpréter la valeur du rayonnement comme la réflectance, l'absorption, la transmittance, la luminance ou encore l'albédo. De fait, pour la télédétection, cette dimension n'est discriminante car il est admis qu'implicitement une image correspond à une matrice de valeurs géophysiques sur laquelle la réflectance, par exemple, pourra être calculée par la suite.

Dimension Procedure

Avec les dimensions spatiale et temporelle, la dimension procédure apporte des informations contextualisant l'Observation. Une grande partie de ces informations sont essentielles pour juger de la qualité et l'adéquation de l'image avec les besoins d'analyse d'un phénomène. Pour cela, elle constitue une dimension essentielle pour naviguer dans les images. Nous distinguerons les informations portant sur les instruments de mesure et leur propriétés telles que proposées par (Gaspéri et *al*., 12) : les Classes *Platform*, *Instrument*, *Sensor* et les conditions d'acquisition décrites dans la classe *Acquisition* (figure 3.7). Par ailleurs, la classe *EarthObservationMetadata* permet également de fournir des informations sur la nature et la qualité du post-traitement qui a permis de transformer la mesure en image. Pour chacune de ses classes et de leurs propriétés, l'expertise et le contexte dans lequel la classification est organisée dirigeront le choix des concepts à développer au sein de cette dimension de recherche. Nous les détaillons dans la figure 3.10 en nous limitant aux propriétés qui nous paraissent essentielles aux besoins de recherche.

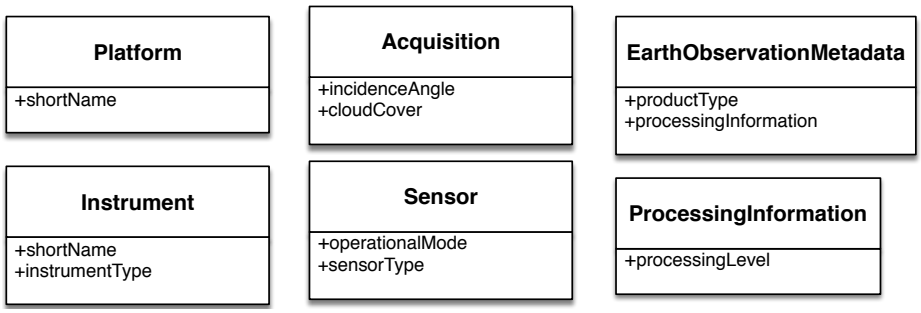


Figure 3.10 : Les principales classes permettant d'expliciter les catégories liées à la dimension *procedure*. D'après (Gaspéri et *al*., 12).

Dimension spatial attribute of FOI

Comme le propose l'ISO 19115, l'emprise au sol d'une image ou *Footprint* peut être déclinée sous différentes classes d'objets qui sont au nombre de trois : *BoundingPolygon*, *BoundingBox* ou *geographicIdentifier*. L'emprise au sol est généralement fournie sous la forme d'une géométrie de type polygone ou d'un jeu de coordonnées qui définissent un rectangle englobant. Afin d'utiliser des classifications ou référentiels géographiques pour naviguer dans cette dimension, nous préconisons d'associer la géométrie à un ou plusieurs identifiants géographiques inclus dans ou qui intersecte l'emprise spatiale de l'image. Ces identifiants pourront être issus d'un référentiel géographique qui fournit par exemple les unités administratives ou tout autre découpage du territoire. Cela s'apparente à ce que (Barde et *al*., 04) puis (Desconnets et *al*., 07) ont proposé pour exprimer la dimension spatiale de l'information. Aussi, de ces différentes représentations possibles, la représentation à partir d'un objet *geographicIdentifier* doit être privilégiée. Le diagramme en figure 3.11 formalise notre point de vue.

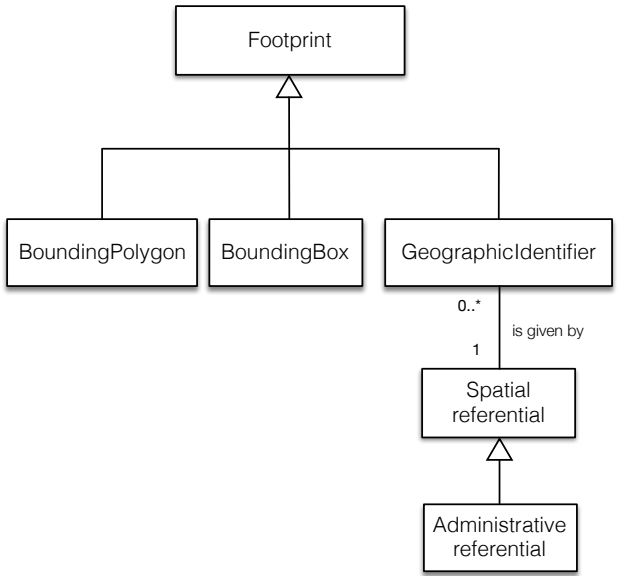


Figure 3.11: Relations entre le concept *FootPrint* et les différents types de représentation spatiale proposées par l'ISO 19115. Nous ajoutons l'association *is given by*. Elle indique qu'un identifiant géographique est issu d'un référentiel spatial qui peut être, par exemple, un référentiel administratif ou tout autre type de référentiel découpant un territoire.

Dimension Time

Comme le précise (Gaspéri et *al*., 12), l'objet temporel pertinent à associer à une prise de vue est *phenomenonTime*. Il correspond à la date et heure de la mesure aussi appelée en télédétection date d'acquisition de l'image. Elle est de type *TM_Object* c'est à dire peut être soit associé à un instant *TM_Instant* ou une période temporelle *TM_Period* (figure 3.12). Le temps d'acquisition d'une prise de vue est souvent de l'ordre de quelques secondes à plusieurs dizaines de secondes selon la longueur de la bande observée. Quoiqu'il en soit, l'ordre de grandeur du temps d'acquisition de la portion de la Terre observée est sans commune mesure avec les phénomènes observés à sa surface qui eux évoluent plutôt sur des périodes allant de la semaine à la saison voire à plusieurs années. Pour cela, il est suffisant de considérer la dimension temporelle comme un instant. Cette dimension est fortement discriminante car les problématiques environnementales étudiées ont une temporalité.

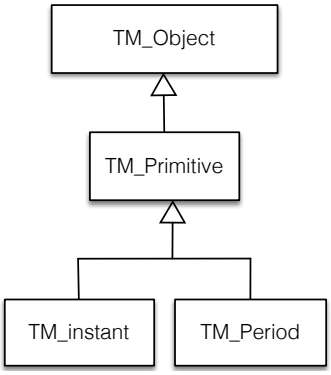


Figure 3.12 Primitives temporelles proposées par l'ISO 19108. Diagramme de classe simplifié

L'objectif est maintenant de croiser les cinq dimensions tirées de la vision Observation du standard O & M et les critères de recherche sur lequel un utilisateur construit une requête (cf. contexte). Le tableau 3.2 met en évidence les dimensions sur lesquelles doivent porter les efforts d'enrichissement et d'adaptation des terminologies, à savoir les dimensions *Space*, *Time* et *Procedure*. Bien que la dimension *FeatureOfInterest* soit essentielle pour un utilisateur car elle porte sur la nature de l'information, les techniques et les outils actuels ne permettent pas aujourd'hui d'associer, lors des opérations de post-traitements de l'image, la valeur géophysique d'un pixel à un point de vue disciplinaire ou à une problématique environnementale. Par conséquent, il n'est pas possible actuellement de disposer d'une telle information dans les métadonnées associées à une image. Enfin, la dimension *observedProperty,* comme nous l'avons exprimé précédemment est peu discriminante pour la recherche d'information sur des collections d'images.

Tableau 3.2: Mise en relation des grandes catégories retenues pour classifier les images et les critères de recherche.

	FeatureOfInterest	observedProperty	Space	Time	Procedure
Où ?					
Quand ?					
Quoi ?					
Comment ?					

Une spécialisation équivalente du modèle O & M et sa décomposition peuvent être aisément reproduites dans d'autres domaines environnementaux. En effet, cette approche a l'intérêt de s'appuyer sur un modèle très général et repose sur des primitives thématique, temporelle et spatiale qui sont formalisées par les standards de l'ISO et l'OGC. Ces standards sont par ailleurs largement implémentés dans les infrastructures d'observation et de diffusion des données spatio-temporelles en environnement.

3.3.4 Concrétisation de la classification à facettes

Notre motivation est à visée utilitariste et guider par le souhait de disposer d'un système d'organisation de connaissances (SOC) pour appuyer les activités d'indexation des images et leur interrogation par les utilisateurs. De l'approche analytico-synthétique qui est sous jacente à la conception d'une classification à facettes, nous retenons uniquement l'approche analytique qui nous a permis de faire émerger les grandes catégories de concepts et sous concepts que nous avons tiré du standard O & M. L'objectif est donc de proposer une déclinaison concrète de la classification à facettes, dans sa dimension analytique, à l'aide d'un SOC nous permettant d'organiser un ensemble de concepts à même de détailler les différentes dimensions de recherche sous la forme d'une poly hiérarchie, autour de chaque catégorie fondamentale. Chaque hiérarchie représente une catégorie fondamentale et décline les différents concepts associés à différents niveaux de spécialisation ou d'abstraction. Cette concrétisation peut être envisagée par la construction d'une ressource termino-ontologique (RTO) (Tissaoui et *al*., 13).

Une ressource termino-ontologique s'attache à décrire un domaine d'intérêt de manière organisée et consensuelle. Une RTO peut être une liste de valeurs, une taxonomie, un thésaurus ou une ontologie. Ce type de ressource organise la connaissance autour de deux dimensions :

- Une dimension ontologique. Elle porte sur la description des entités d'un domaine, leur niveau d'abstraction et leur interrelation avec les autres entités du domaine (voisinage, équivalence, subsomption) ;
- Une dimension terminologique qui permet d'organiser les termes qui désignent les concepts associés pour une langue donnée. Ces termes vont faciliter les opérations d'indexation et d'interrogation documentaires. Ce sont ces tâches auquel répondent les thésaurus.

Pour cela, la concrétisation de la classification à facettes par un thésaurus apparaît pertinente. Le thésaurus porte son attention sur la dimension terminologique tout en offrant une structuration des concepts autour de leurs relations sémantiques. L'agence française de normalisation (AFNOR, 81) définit un thésaurus comme « *une liste d'autorité organisée de descripteurs et de non descripteurs obéissant à des règles terminologiques propres et reliés entre eux par des relations sémantiques (hiérarchiques, associatives ou d'équivalence). Cette liste sert à traduire en un langage artificiel dépourvu d'ambiguïté des notions exprimées en langage naturel* ».

Le diagramme de classe en figure 3.13 proposé par (Zayrit, 10) traduit les principes directeurs, édictés par les organismes de normalisation, pour l'élaboration des thésaurus. La classe *Terme* est spécialisée par les sous classes *Descripteur* et *Non Descripteur*. Les trois relations sémantiques d'équivalence (*Employé*/*EmployéPour*), associative (*TA*) et hiérarchiques (*TS*/*TG*) sont précisées. L'extension de cette vision purement terminologique vers une vision ontologique est également proposée par (Zayrit, 10) (classes grisées de la figure 3.13). Elle permet d'introduire la classe *Concept* et la classe *Thésaurus*, cette dernière est un agrégat de concepts. L'association *Désigne* explicite la relation entre un Concept et son terme *Descripteur*. L'introduction de la classe *Thesaurus,* associé aux classes *Terme et Concept* permet d'envisager l'utilisation conjointe de plusieurs thésaurus comme cela est souvent nécessaire pour interconnecter des terminologies venant de disciplines diverses. Nous verrons plus loin dans notre exposé que nous en ferons usage afin de décrire les relations sémantiques entre termes issus de plusieurs thésaurus. Enfin, l'association *TermeTeteDe* va permettre d'accéder aux termes les plus génériques de chaque thésaurus. Ils correspondent aux catégories fondamentales qui constituent le point d'entrée de chaque dimension d'une classification à facettes.

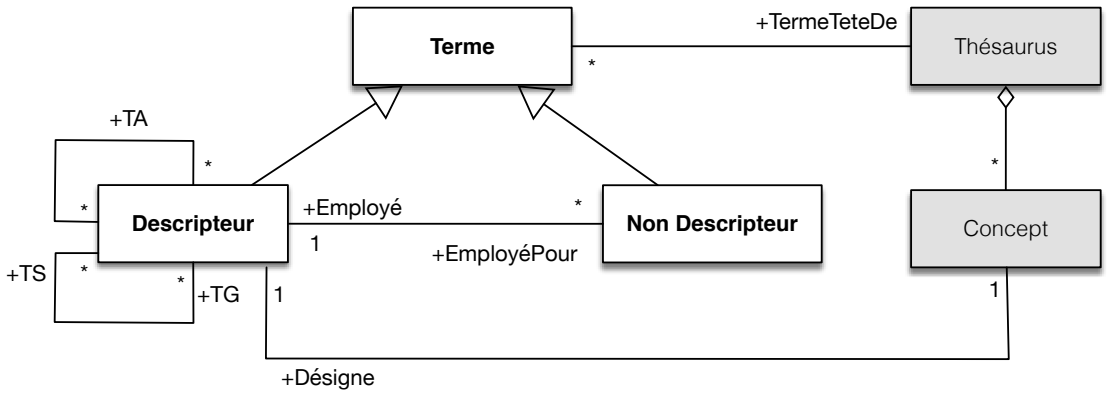


Figure 3.13 : Diagramme de classes formalisant la notion de Thesaurus telle qu'elle introduite par AFNOR, 81. Les classes grisées étendent cette notion pour introduire la dimension ontologique du Thésaurus (d'après Zayrit, 10).

Au fil des années, un ensemble de travaux de normalisation sont venus fournir les principes pour l'élaboration de thésaurus. Nous citerons celui qui fait office de référence en la matière, la norme internationale ISO 2788 (ISO, 86). Ces travaux visent également à faciliter l'interopérabilité des thésaurus. Cela prend tout son sens dans le contexte d'échange de données et d'information à travers le web comme celui qui nous anime. Les thésaurus comme la majorité des autres RTO s'adossent sur les langages de description du W3C. Nous évoquerons en particulier le format SKOS (*Simple Knowledge Organisation System*) (Isaac et *al*., 07) qui propose un cadre de représentation standard

pour la construction de RTO. Il a une portée suffisamment généraliste pour être aisément utilisable pour décrire les activités autour du domaine de l'environnement. SKOS reprend pour une grande part les travaux de normalisation entrepris précédemment autour de la conception, gestion et maintenance des thésaurus (Chichereau et *al*., 07).

Le vocabulaire SKOS est défini à partir des éléments de modélisation présents dans RDF, RDFS et OWL proposé par le W3C. Il est proposé sous différentes formes selon le besoin d'expressivité attendu lors de la construction RDFS ou OWL. Nous présentons en figure 3.14 les éléments cœur (SKOS Core) du vocabulaire sous la forme d'un diagramme de classe. Nous faisons le parallèle avec le modèle précédent (figure 3.13). Le vocabulaire SKOS est défini autour des classes *Concept*, *ConceptScheme* et *Collection*. *ConceptScheme* correspond à la classe *Thésaurus* du modèle précédent. *Concept* correspond à la fois aux classes *Descripteur* et *Concept*. La classe *Concept*, comme toutes les autres classes SKOS, est caractérisée par la propriété p*refLabel* (label préféré) qui est le label du descripteur, une collection de *altLabel* (labels alternatifs) que l'on peut faire correspondre aux termes non-descripteurs qui n'est autre que la traduction de la relation d'équivalence (ou synonymie). Les relations hiérarchiques *broader* (et son inverse *narrower*) et associatives (*related*) sont des associations réflexives sur la classe *Concept*. Elles correspondent respectivement aux associations *TG/TS* et *TA* du modèle en figure 3.13. Enfin, l'association *topOfConcept*, et son inverse *hasTopConcept*, modélisent les concepts les plus génériques d'un *ConceptScheme*. Elle correspond à l'association *TermeTeteDe* du diagramme précédent. Enfin, SKOS apporte un élément d'organisation nouveau avec la classe Collection. Il permet d'organiser les concepts au travers de catégories complémentaires.

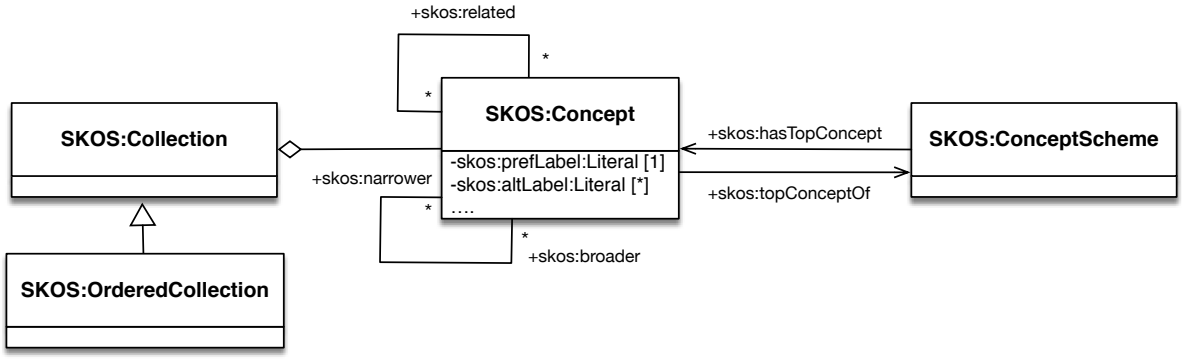


Figure 3.14 : Diagramme de classe du vocabulaire SKOS Core

SKOS est devenu, au même titre RDF, RDFS ou OWL sur lesquels il s'appuie, un standard du W3C. De nombreux outils ou librairies sont disponibles pour visualiser, gérer, manipuler le vocabulaire SKOS comme le SKOSeditor[46] plugin de Protégé, l'éditeur *standalone* ThManager[47], l'outil TemaTres[48] ou encore la plateforme PoolParty[49]. Dans notre domaine, les RTO construits sur le format SKOS sont nombreux. Nous citerons les plus emblématiques dans notre domaine AGROVOC[50], GEMET[51] , Getty Thesaurus[52] ou l'ontologie de nom de lieu GeoNames[53]. Il est également à noter que

[46] SKOS editor pour Protégé : https://code.google.com/archive/p/skoseditor/
[47] ThManager : http://thmanager.sourceforge.net/
[48] TemaTres : http://www.r020.com.ar/tematres/demo/index.php
[49] PoolParty : https://www.poolparty.biz/
[50] AGROVOC : thésaurus de l'agriculture maintenu par la FAO : http://aims.fao.org/
[51] GEMET : General Multilingual Environmental Thesaurus http://eionet.europa.eu/gemet.
[52] Getty Thesaurus: http://www.getty.edu/research/tools/vocabularies/lod/

nos travaux, effectués lors le projet Européen FP7 NatureSDIplus (Carlisle et *al*., 10) et le projet REFLECS du CNES, nous ont permis de concevoir et d'implémenter un composant dédié à l'accès et à la gestion de thésaurus SKOS au sein de l'outil MDweb (Boisson et *al*., 06 ; Desconnets et Libourel, 12).

Ainsi, la concrétisation proposée met à disposition la vision orientée utilisateur. Sa représentation standardisée au format SKOS va nous permettre de la manipuler au sein d'architecture ouverte et décentralisée sur le web. C'est en utilisant l'expression « référentiel terminologique utilisateur » ou « référentiel de valeurs utilisateur » que nous évoquons dans la suite de notre exposé l'utilisation de la classification à facettes proposée. En effet, nous envisageons son rôle comme référentiel de valeurs. Il s'agit d'exploiter les termes du thésaurus comme étant autant de valeurs potentielles des éléments de métadonnées qui viendront décrire les images et peuplés l'index de recherche. Ce rôle est déjà évoqué dans notre état de l'art (section 2.2.5.2 Verrous pour la mise en œuvre au sein des SRI, figure 2.10).

3.4 Schéma de métadonnées pour le partage des images satellitaires

Après avoir exposé les préoccupations des utilisateurs au travers de leur pratique de recherche d'information et les propositions d'organisation de connaissances pour promouvoir leur vision, nous abordons dans cette section les besoins des concepteurs de système de partage. Les métadonnées constituent le socle pour la recherche, l'identification, la localisation, la pérennisation des images au sein de tels systèmes. C'est à travers le prisme de la structuration des métadonnées pour la recherche d'information que nous revenons sur nos travaux entrepris de 2014 à 2017 autour des profils d'application de métadonnées. L'utilisation d'un schéma de métadonnées et des référentiels de valeurs associés pour assurer l'harmonisation des métadonnées sont considérés ici comme un préalable à toute opération d'indexation, d'enrichissement et d'adaptation des valeurs de métadonnées sur des collections d'images. Après avoir rappelé les motivations qui nous ont amené à proposer un profil d'application pour les données d'observation de la Terre, nous revenons sur la notion de profil d'application et les principes de construction. Nous présentons brièvement le profil EOAP conçu pour les besoins de la communauté environnementale. Le lecteur intéressé pourra approfondir ces aspects en lisant les publications suivantes : (Chahdi, 13 ; Desconnets et *al*,. 14 ; Mougenot et *al*., 15 ; Desconnets et *al*., 17). Les deux premières traitent de la définition du profil EOAP. La troisième reconsidère les recommandations autour du modèle structurel DSP (Description Set Profile) associé à la définition d'un profil d'application. La quatrième revient sur la méthodologie utilisée pour concevoir de tels schémas de métadonnées. (El Houssine, 14 ; Loukili, 14) ont exploité le profil d'application pour produire un environnement de gestion des profils d'application dans le premier cas et un environnement orienté utilisateur relatif à la découverte des ressources de renforcement de capacités dans le domaine de l'observation de la Terre, dans le deuxième cas.

3.4.1 Besoin d'un cadre d'interopérabilité

Le partage d'images multi plateformes au sein d'une infrastructure de données spatiales nécessite la mise en place d'un cadre d'interopérabilité pour remédier aux hétérogénéités structurelle et sémantique et ainsi assurer l'interrogation uniforme des métadonnées d'images par les services de découverte. Ce cadre s'appuie généralement sur l'adoption d'un standard de métadonnées propre à une communauté d'utilisateurs ou sa déclinaison. Il constitue le socle sur lequel repose l'implémentation

[53] GeoNames : http://www.geonames.org/

des services d'accès aux images. Au fil des diverses initiatives de partage, des pratiques ont émergé. Elles s'attachent à couvrir des exigences fonctionnelles qui leur sont propres et produisent ainsi des métadonnées hétérogènes qui rendent délicate la mise en œuvre d'outils communs de découverte. En effet, les divers cadres d'interopérabilité définis dans le domaine de l'observation de la Terre dont celui du GEOSS (Christian, 08; Nativi et Bigagli, 09) ou du pôle THEIA[54] (Leroy et *al*., 13) proposent des modèles de métadonnées qui sont à l'intersection des modèles des différents fournisseurs de données. Le modèle résultant correspond le plus souvent aux éléments coeur d'un standard. Il apporte les informations générales sur l'image pour couvrir les besoins de découverte et de localisation uniquement. Celles relevant des caractéristiques de l'image, des contextes d'acquisition ou de production sont rares, voire absentes. Il sera alors difficile de proposer des applications de recherche riches, d'envisager l'extension des fonctionnalités rendues par un tel système sans remettre en cause les modèles de description sous-jacents ou leur interopérabilité. Ce constat nous a amène à poser une réflexion autour du rôle des métadonnées dans ce contexte d'exploitation de grandes quantités de ressources hétérogènes et distribuées et autour des démarches menant à l'interconnexion de données d'observation de la Terre.

En nous appuyant sur la notion de profil d'application telle que proposée par l'initiative Dublin Core (Hillman et *al.*, 10), la démarche a pour objectif de fournir un modèle ouvert, extensible et exploitable pour faciliter le partage et la gestion des images distribuées au sein d'architectures décentralisées. Elle apporte des éléments de réponse en matière d'utilisation de systèmes interopérables sur le web. Il est destiné, à terme, à couvrir les besoins de découverte, de localisation, de consultation, de préservation et de traitements des données à des fins d'analyse. Elle est également envisagée dans une perspective d'ouverture des images sur le web.

3.4.2 Notion de profil d'application

Un reproche que l'on pourrait faire aux standards de métadonnées est qu'ils sont conçus indépendamment les uns les autres et de ce fait ne permettent pas de répondre à tous les besoins en matière d'usage de l'information. A cet effet, les profils d'application réutilisent les standards de métadonnées pour les amener à répondre soit à de nouveaux besoins, soit à des besoins plus ciblés. Le principe est d'emprunter différents éléments de métadonnées à différents standards et de les combiner, approche dite de *mix and match* (Heery et Patel, 2000) pour produire une nouvelle organisation d'éléments adaptée à une visée applicative cible. La figure 3.15 illustre ce principe.

[54] Pôle THEIA : Pôle Thématique Surface Continentale. https://www.theia-land.fr

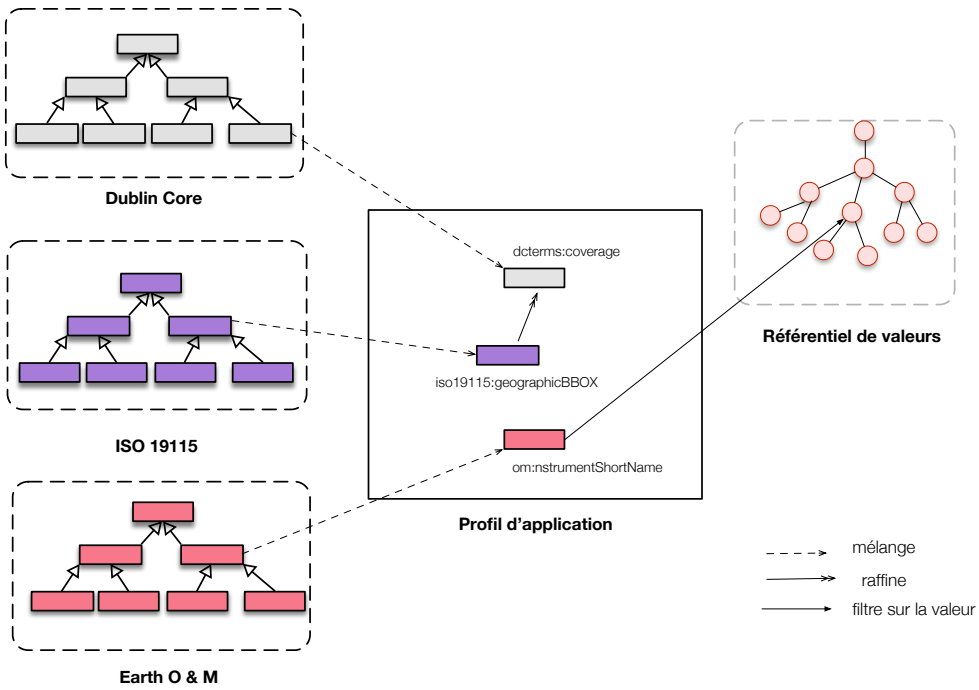


Figure 3.15: Illustration de l'approche mix and match (inspirée de Mougenot, 15).

La définition d'un profil d'application s'avère pertinent pour faciliter l'interopérabilité entre les différentes sources de données. Dans ce sens, il peut être vu comme un modèle de médiation. En effet, Il assure l'interopérabilité sémantique grâce à l'utilisation des éléments de métadonnées définis. Il assure également l'interopérabilité structurelle et syntaxique par le respect des règles syntaxique et structurelle relatives à sa matérialisation (Nilsson et *al.*, 2008).

3.4.3 Principes de construction

La construction d'un profil d'application est soumise à différentes règles. La toute première est de n'exploiter que des standards de métadonnées existants, ou à défaut de maintenir de manière ouverte et sur le long terme un nouveau standard de métadonnées qui vient couvrir les éléments de métadonnées nouvellement introduits. Les principes suivants concernent la publication des démarches d'enrichissement et d'exploitation des entités du modèle à partir des éléments de métadonnées provenant des standards. Nous nous adossons aux travaux menés, autour du Singapore Framework (Nilsson, 08) et du profil d'application nommé DCAP (Dublin Core Application Profile), par la communauté Dublin Core. Des préconisations d'ordre méthodologique (Nilsson et *al.*, 09) ainsi que la spécification de modèles conceptuels UML viennent faciliter les activités de construction d'un profil d'application.

Un premier modèle structurel nommé DCAM (Dublin Core Abstract Model) (Powell et *al.*, 07) explicite la notion de ressource et sa spécialisation en ressource décrite qui est alors une collection de couples propriété-valeur. La valeur s'envisage parfois aussi comme une ressource. La propriété comme la valeur peut être empruntée aux standards de métadonnées et aux référentiels de valeurs. Un second modèle structurel nommé DSP (Description Set Profile) (Nilsson et *al.*, 09) vient compléter le modèle DCAM pour fournir un cadre prescriptif à la construction du profil d'application. Un profil d'application est envisagé alors comme un ensemble de descriptions. Il est décrit au travers de la notion de *DescriptionSetTemplate*. Chaque description est appelée *DescriptionTemplate* et vient

enrichir de manière décentralisée une ressource d'intérêt en la documentant au travers d'éléments de métadonnées provenant de standards appropriés. Ces éléments, ainsi que les différentes contraintes syntaxiques et/ou sémantiques qui s'y appliquent, sont structurées au sein de déclarations appelées *StatementTemplate*. Les déclarations sont des *LiteralStatementTemplate* lorsque les éléments de métadonnées pointent sur des valeurs terminales (littéraux) ou des *NonLiteralStatementTemplate* lorsque les éléments de métadonnées pointent sur des ressources étiquetées provenant de référentiels de valeurs à l'exemple des thésaurus au format SKOS. Les contraintes sont explicitées au travers de la notion de *Constraint* qui se spécialise en *LiteralConstraint* et *NonLiteralConstraint*. Nous proposons un diagramme structurel simplifié du DSP.

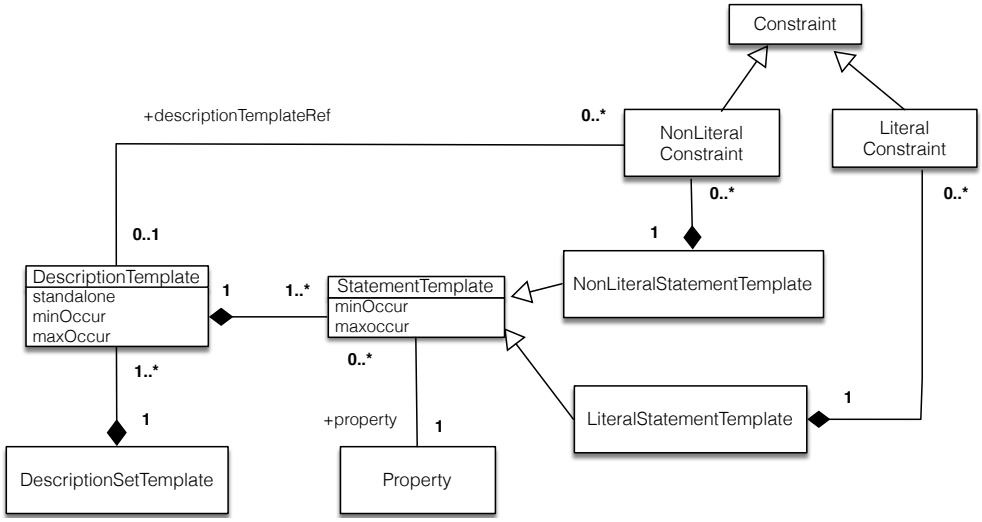


Figure 3.16: Diagramme de classes UML simplifié du modèle Description Set Profile (DSP)

3.4.4 Le profil EOAP

En s'appuyant sur les modèles structurels décrit précédemment (DCAM, DSP), la construction du profil EOAP reprend le cheminement préconisé par le Singapore Framework à savoir : la définition des exigences fonctionnelles du domaine applicatif visé ; la définition d'un modèle de domaine et enfin le modèle de description DSP. Nous revenons brièvement sur les éléments du profil EOAP.

Exigences fonctionnelles

Les exigences fonctionnelles attachées à la construction du profil EOAP (Desconnets et *al.*, 14) répondent au besoin de mettre à disposition un grand nombre d'images multi-capteurs vers la communauté des acteurs publics de l'environnement. A titre d'exemple, l'exigence « résoudre l'hétérogénéité des schémas de métadonnées et des valeurs de métadonnées pour assurer une interrogation uniforme » met en évidence le besoin de faire reposer la description des images, et la terminologie utilisée pour les annoter, sur un unique schéma afin que les opérations de recherche, de visualisation, d'évaluation de compatibilité ou de qualité soit uniforme quelque soit le capteur et rende plus efficace ces activités pour un acteur public.

Modèle de domaine

Un modèle de domaine est un modèle conceptuel. Il a pour objectif de définir les entités qui devront être décrites par le profil d'application conformément aux exigences fonctionnelles. Il s'agit bien à travers de ce modèle de définir le périmètre du profil d'application. L'entité coeur *Resource* fait référence à la notion de Resource telle que proposée par le DCAM (Dublin Core Abstract Model). C'est une entité abstraite qui représente l'ensemble des ressources qui sont à partager. Elle généralise les deux types de ressources spécifiques au domaine traité, à savoir les entités *Process* et *EarthImage*. La relation relation *IsPartOf* sur l'entité *Resource* permet de représenter, pour les ressources de type *EarthImage*, la relation d'agrégation entre une collection d'images, ensemble d'images ayant des propriétés communes, et une image. Sémantiquement équivalente à la Classe *dcmi:Agent*, l'entité *Agent* est également une entité abstraite qui est spécialisée en *Organisation* et *Sensor*. La première décrit les institutions prenant part à la création d'une ressource ou sa distribution au sein de la communauté à travers les relations *isCreatedBy*, *isDistributedBy*.

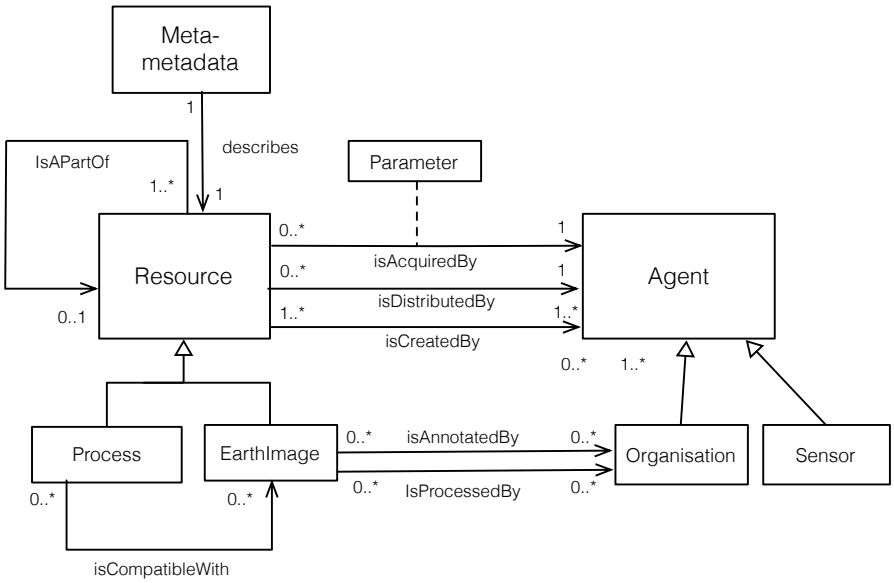


Figure 3.17: Modèle de domaine pour l'observation de la Terre (diagramme de classes UML)

La seconde, *Sensor*, représente les instruments de mesures assurant l'acquisition d'une image satellitaire. L'association *isAcquiredBy* permet de décrire les caractéristiques du capteur ayant servi à l'acquisition d'une image ou d'un ensemble d'images ainsi que les paramètres propres à l'acquisition d'une image (classe d'association *Parameter*). L'entité *EarthImage* représente les images d'observation de la Terre que l'on souhaite rendre accessibles. Elle apporte les caractéristiques intrinsèques de l'image notamment les informations portant sur les dimensions temporelle et spatiale, ses résolutions spatiale et spectrale ainsi que des informations pour assurer sa distribution telles que son format, les restrictions d'utilisation. Enfin, l'entité *Meta-metadata* apporte les informations nécessaires à la gestion des enregistrements de métadonnées (*e.g.* date de modification, langue de description, etc.). Les classes et les relations de notre modèle de domaine sont documentées à cette adresse : http://purl.org/eoap/. Nous avons délibérément omis de discuter les éléments du modèle relatifs à la classe *Process*. La démarche, les motivations qui nous ont conduit à poser ce modèle de domaine, ainsi que l'ensemble du modèle de domaine sont largement discuter dans (Desconnets et *al*., 17b).

Description Set Profile

Le modèle DSP (Description Set Profile) complète le modèle DCAM (Dublin Core Abstract Model) et fournit un cadre prescriptif pour construire un profil d'application (Nilsson et *al*., 2009). Ainsi, **le profil d'application est bien un modèle qui ne prescrit pas les données d'intérêt, dans notre cas les images, mais les éléments de métadonnées qui décrivent ces données**. Leur consultation est envisagée grâce aux instances du DSP, *i.e* les jeux de métadonnées. Chacune des entités identifiées dans le modèle de domaine sont déclinées en jeux de description. Nous illustrons ces relations dans la figure 3.18, ci-après.

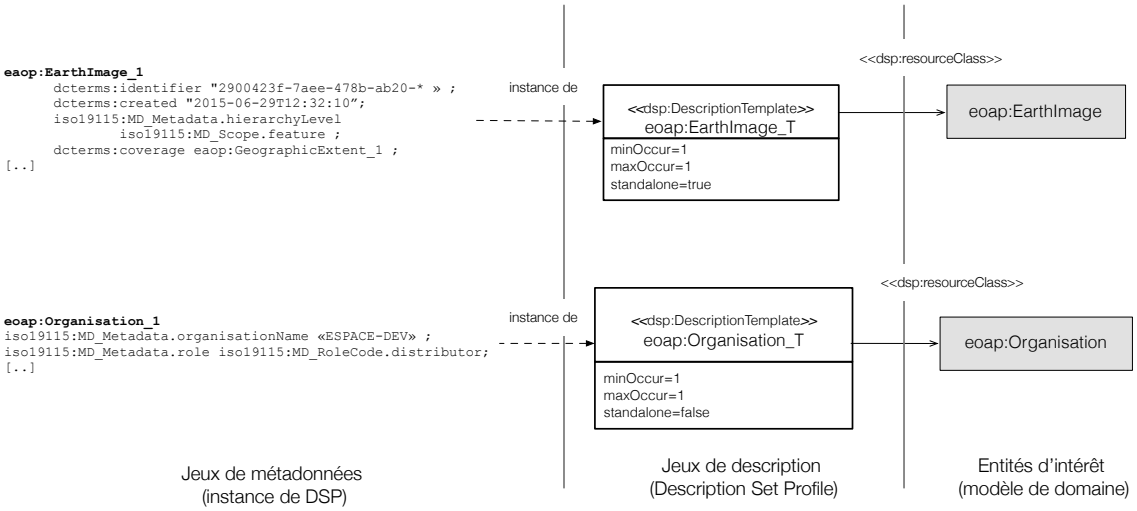


Figure 3.18 : Illustration des relations entretenues entre entités d'intérêt – jeu de description - jeu de métadonnées

Nous avons porté nos efforts de description sur les entités d'intérêt afin de couvrir les besoins de découverte et de localisation. Les propriétés qui permettent de décrire les entités d'intérêt réutilisent les éléments de description basique émanant du DCMI Metadata Terms[55]. Ils sont raffinés pour décrire des caractéristiques propres à une image satellitaire telle que l'empreinte spatiale d'une image, par le standard de métadonnées ISO 19115, ISO 19115-2 ou le profil O & M pour l'observation de la Terre. Nous donnons un extrait du jeu de description *EarthImage_T* portant sur le raffinement de l'élément *dcterms:coverage* par la classe *iso19115: EX GeographicBoundingBox* (Listing 3.1).

```
eoap:EarthImage_T
a       dsp:DescriptionTemplate ;
       dsp:maxOccur "1"^^xsd:nonNegativeInteger ;
       dsp:minOccur "1"^^xsd:nonNegativeInteger ;
       dsp:resourceClass eoap:EarthImage ;
       dsp:standalone "true"^^xsd:boolean ;
    [...]
       dsp:statementTemplate
[ a   dsp:NonLiteralStatementTemplate ;
dsp:maxOccur "infinity" ;
        dsp:minOccur "1"^^xsd:nonNegativeInteger ;
                    dsp:NonliteralConstraint
                         [ a       dsp:NonLiteralConstraint ;
```

[55] DCMI Metadata Terms : http://dublincore.org/documents/dcmi-terms/

```
                            dsp:DescriptionTemplate  <eoap:GeographicExtent_T>;
                            dsp:valueStringOccurence "disallowed"^^<dsp:Occurence>;
                            dsp:VocabularyEncodingSchemeOccurence
"disallowed"^^<dsp:Occurence>;                          ] ;
                      dsp:property  dcterms:coverage                              ]  .
eoap:GeographicalExtent_T
           a       dsp:DescriptionTemplate ;
           dsp:maxOccur "infinite" ;
           dsp:minOccur "1"^^xsd:nonNegativeInteger ;
           dsp:resourceClass iso19115:EX_GeographicalBoundingBox ;
           dsp:standalone "false"^^xsd:boolean .
    dsp:statementTemplate
    [ a  dsp:NonLiteralStatementTemplate ;
    [...]
    dsp:property iso19115:EX_GeographicBoundingBox.westBoundLongitude
    ] .
    [...]
```

Listing. 3.1 : Extrait du profil EOAP sérialisé en RDF (syntaxe N3). Pour plus de clarté, la description du système de référence spatial qui accompagne la description de l'emprise spatiale n'est pas fournie dans ce fragment.

Le listing 3.2 présente un extrait de l'instance du jeu de description *EarthImage_1* correspondant à l'instanciation du jeu de description présenté dans le listing 3.1.

```
    eaop:EarthImage_1
          dcterms:identifier "2900423f-7aee-478b-ab20-fe932d4adb" ;
          dcterms:created "2015-06-29T12:32:10";
          iso19115:MD_Metadata.hierarchyLevel
                  iso19115:MD_Scope.feature ;
          dcterms:coverage eaop:GeographicExtent_1 ;
    [..]
    dcat:downloadURL                                                  http://ids.equipex-
geosud.fr/constellation/rest/secured/download/data/SPOT6_2015_FRANCE-ORTHO_IGN-
MS/MD_S6X_2015062937945426CP.tar.gz
    [..]
    eoap:GeographicExtent_1
             iso19115:EX_GeographicBoundingBox.westBoundLongitude
                 "3.71";
          iso19115:EX_GeographicBoundingBox.eastBoundLongitude
                  "4.04";
             iso19115:EX_GeographicBoundingBox.northBoundLatitude
                 "43.72";
          iso19115:EX_GeographicBoundingBox.southBoundLatitude
                  "43.53";
```

Listing. 3.2 : Extrait d'une instance du profil EOAP (syntaxe N3)

3.5 Méthodes et outils pour élaborer un système de recherche orienté utilisateur

3.5.1 Indexation guidée par les thésaurus

Contrairement aux approches communément décrites (Santoro et *al*., 12 ; Gui et *al*., 13) pour traiter les ambiguïtés ou les « décalages » sémantiques dans les moteurs de recherche associés aux infrastructure de données spatiales, la démarche proposée aborde la désambigüisation des termes lors de l'indexation des données. Notre proposition tire partie d'une part des référentiels de valeurs orientés « utilisateur » ainsi que des référentiels de valeurs issus des spécifications des producteurs d'images et d'autre part, des référentiels spatiaux que les communautés environnementale et géographique mettent à disposition. Les deux premiers sont mis à contribution pour contrôler et réduire l'hétérogénéité sémantique au niveau des valeurs de métadonnées par l'alignement des nomenclatures des producteurs avec celles des utilisateurs. Les seconds sont exploités pour enrichir les métadonnées d'images par des hiérarchies administratives. Pour ce dernier point, notre motivation est

d'affranchir l'utilisateur des manipulations diverses pour formuler le critère Où ? (saisie des coordonnées, définition d'une zone à l'aide de la souris, etc.) en proposant une classification hiérarchique administrative sur laquelle il s'appuiera pour sélectionner le ou les noms de lieu sur lequel porte son besoin.

3.5.1.1 Entre harmonisation et publication des métadonnées

Positionnement dans le cycle de vie des métadonnées

La standardisation ou plus généralement l'harmonisation des métadonnées est un pré requis pour en assurer l'interopérabilité et la diffusion via des services standard de découverte. La figure 3.19 schématise une partie du cycle de vie d'une métadonnée au sein d'une infrastructure de données spatiales. L'indexation envisagée se situe en aval de l'harmonisation des métadonnées et en amont de leur publication dans les services de découverte. Cela permet d'une part, d'envisager les opérations d'enrichissement et d'adaptation sur des jeux de métadonnées harmonisés, et d'en diminuer ainsi la complexité, et d'autre part, de pouvoir publier, si besoin, les métadonnées enrichies.

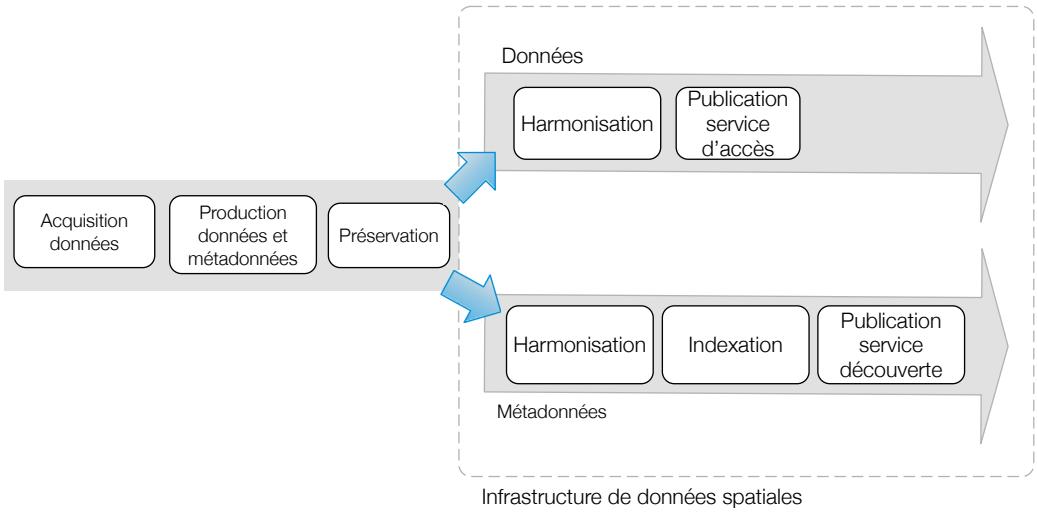


Figure 3.19: Place de l'indexation dans le cycle de vie des métadonnées au sein d'une infrastructure de données spatiales

L'indexation guidée par les thésaurus est donc un moyen de remédier aux hétérogénéités sémantiques, au niveau des valeurs des métadonnées en normalisant les valeurs prises dans les éléments de métadonnées. L'interopérabilité sémantique peut donc être améliorée dès lors que l'on souhaite agréger des métadonnées issues de divers services de découverte et disposer d'un moteur de recherche efficace.

Positionnement de notre approche au sein de la phase d'indexation

Le schéma en figure 3.20 positionne notre proposition d'indexation. Nous retiendrons que les opérations que nous souhaitons mener se situent entre la phase d'extraction des valeurs de métadonnées et leur analyse en vue de leur stockage sous la forme de vecteurs dans l'index.

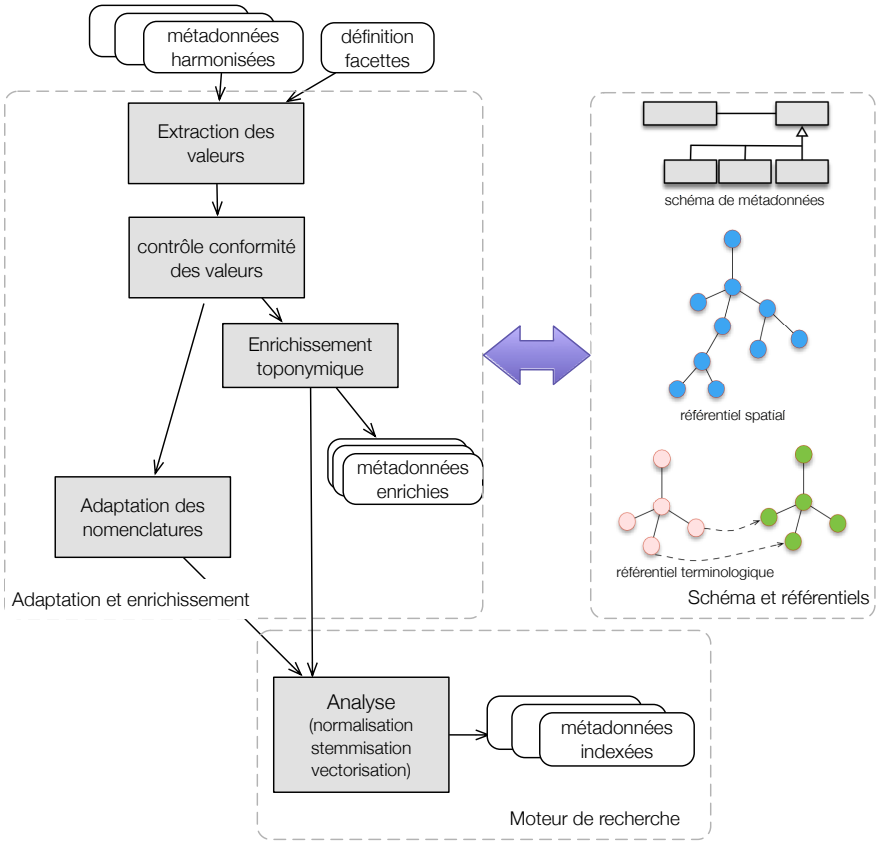


Figure 3.20: Les différentes étapes de l'indexation guidée par les thésaurus

Contrairement, à l'indexation de données textuelles non structurées, celle des images satellitaires n'a pas besoin de se préoccuper, lors de l'extraction des valeurs dans les fichiers de métadonnées, de leur association avec les concepts du domaine. En effet, le schéma de métadonnées sur lequel est réalisé l'harmonisation nous donne la sémantique de chaque élément. Le schéma de métadonnées peut être alors considéré comme une représentation formelle des principaux concepts sur lesquels on souhaite faire reposer l'indexation. De même, la phase d'analyse des chaînes de caractères qui a pour objectif la construction des vecteurs de termes ne revêt pas un caractère crucial. La structuration des métadonnées et la normalisation des valeurs par les référentiels *ad hoc* facilite l'étape de vectorisation. L'index nous permet avant tout de disposer d'une structure de données performante pour l'accès à la représentation des images. C'est également une structure de données adaptée à l'exécution des requêtes sur des facettes (Hostetter, 06 ; Kuć et Rogoziński, 13). Enfin, le calcul de similarité entre les vecteurs de termes et la requête utilisateur, préconisé pour assurer le classement des résultats, n'est pas pris en considération en première approche. En effet, le système de recherche à facettes envisagé ainsi que la structuration et l'adaptation des métadonnées assurent conjointement des mesures de précision et de rappel suffisamment hautes pour nous en affranchir.

Aussi, nous avons choisi de nous concentrer sur les étapes de contrôle, d'enrichissement et adaptation des métadonnées qui relèvent d'une approche « métier ». Nous nous appuierons sur un moteur de recherche existant pour assurer les opérations d'analyse, de requêtes et de restitution des résultats.

3.5.1.2 Contrôle et adaptation des nomenclatures discriminantes

3.5.1.2.1 Principes

Afin de réduire les hétérogénéités et les ambiguïtés sémantiques présentes dans les métadonnées, et en priorité celles qui sont associées aux dimensions de recherche, le principe est de contrôler les valeurs des caractéristiques d'une image sur un référentiel terminologique « producteur». Dans un deuxième temps et quand cela est pertinent, les nomenclatures « producteur» lexicalement validées sont alignées vers une nomenclature « utilisateur ». L'objectif est de remplacer la terminologie « producteur » par la terminologie « utilisateur » dans l'index du moteur de recherche. De fait, il s'agit de pouvoir disposer des classifications « utilisateur » pour construire les facettes de recherche. C'est par exemple, le besoin identifié pour les critères tels que la résolution spatiale ou le niveau de traitements que l'on souhaite mettre à disposition des utilisateurs peu familiers du domaine de la télédétection.

Comme nous l'avons exposé précédemment, nous mettons à profit les thésaurus comme référentiel de valeurs. L'objectif est d'exploiter les termes comme étant des valeurs potentielles des éléments de métadonnées. Il s'agit ensuite d'utiliser les relations sémantiques entre termes pour aligner les terminologies « producteur » et « utilisateur ». Nous choisissons d'organiser la démarche à partir de la représentation standard d'un thésaurus structuré à l'aide du vocabulaire SKOS. Dans une première approche, les propriétés associées à chaque concept *skos :prefLabel*, *skos :altLabel* ainsi que les relations hiérarchiques *skos :narrower* *skos :broader* et associatives *skos :related* ou *skos:relatedMatch* suffisent à traiter les besoins de contrôle et d'alignement envisagés.

3.5.1.2.2 Contrôle des valeurs des éléments de métadonnées

Le contrôle du domaine de valeurs d'un élément de métadonnées peut être délégué à un thésaurus. L'utilisation que nous proposons ici place le thésaurus comme une ressource terminologique qui nous permet d'automatiser ce contrôle. L'extrait du thésaurus SKOS du projet GEOSUD donné en figure 3.21 nous permet de dérouler le principe général.

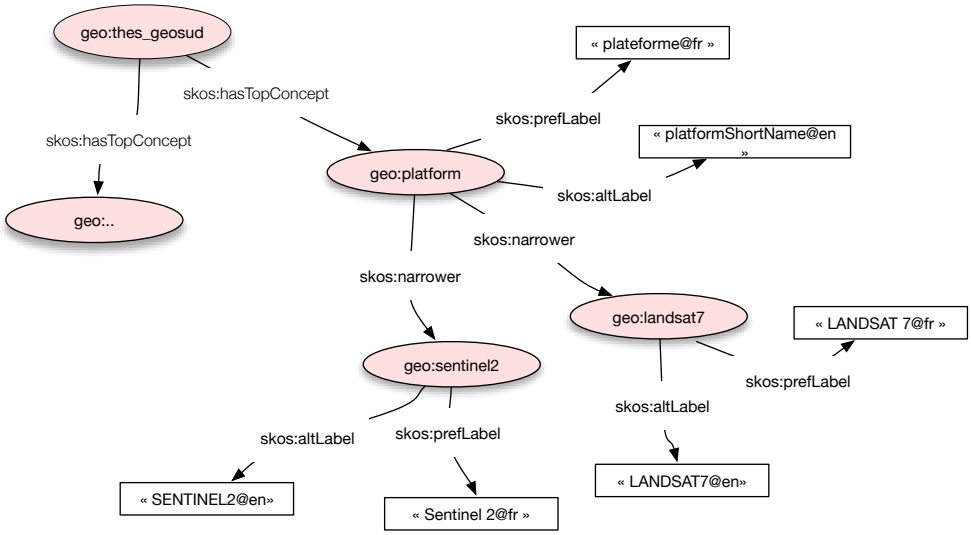


Figure 3.21 : Extrait du thésaurus SKOS GEOSUD.

Nous prenons l'exemple de l'élément de métadonnées *platformShortName* qui décrit le nom du satellite qui a servi à acquérir une image. Les éléments de métadonnées utilisés pour décrire une image constituent les *skos:TopConcept* du thésaurus utilisé. Les *skos:Concept* qui sont rattachés à chacun d'entre eux correspondent aux valeurs possibles de l'élément de métadonnées. Pour chacun des éléments de métadonnées, l'algorithme de contrôle parcourt les *skos:TopConcept* du thésaurus. Si une correspondance est trouvée, le parcours des *skos:Concept* fils est réalisé. Un test d'égalité entre la valeur de métadonnées est réalisée sur la propriété *skos:altLabel*. Une normalisation de la chaîne de caractères testée est réalisée au préalable. Si aucune correspondance n'est trouvée, un ensemble de sous-chaines, pour chaque chaîne de caractères, est créée et est comparé selon la méthode des N-gramm. Cela permet de tenir compte d'éventuelles inversions de caractères. Si une similitude est trouvée, la valeur de l'élément de métadonnées est corrigée. Suite à cette opération, certains éléments de métadonnées peuvent faire l'objet d'une adaptation, à la suite de cette opération.

3.5.1.2.3 Adaptation des nomenclatures discriminantes

Comme mentionné plus haut, l'objectif est d'associer à une propriété d'une image une sémantique plus proche du point de vue utilisateur. Techniquement, il s'agit d'aligner les termes issus de deux terminologies, celle du producteur d'images et celle qui représente le point de vue de l'utilisateur, pour que le moteur de recherche soit enrichi de la terminologie utilisateur. Dans la littérature, l'enrichissement d'annotations de contenus est généralement utilisé pour évoquer cette démarche. Nous tenons à la définir comme une adaptation car elle vise à bien adapter la sémantique fournie à l'utilisateur. L'alignement réalisé entre les deux terminologies utilise la propriété *skos:related* ou *skos:relatedMatch* si elle fait partie ou non du même thésaurus, *skos:ConceptScheme*. Elle traduit une relation d'équivalence entre concepts. L'alignement consiste en une mise en correspondance exacte entre 1 à *n skos:Concept* du référentiel « producteur » vers un et seul *skos:Concept* du référentiel « utilisateur ». La figure 3.22 schématise cet alignement.

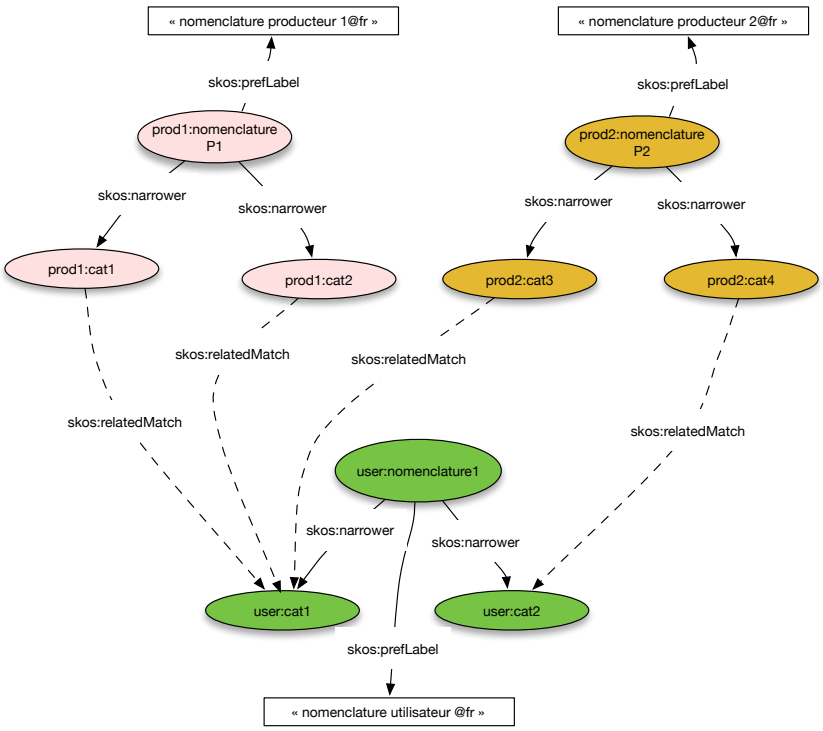


Figure 3.22: Alignement entre deux nomenclatures « producteur » et une nomenclature « utilisateur »

Cet alignement est construit en amont de la phase d'indexation par un expert. Nous verrons dans la section mise en œuvre que la gestion des alignements est facilitée par un outil dédié à la gestion des thésaurus SKOS et permet ainsi de les mettre à jour sans remettre en cause le traitement automatique de l'indexation.

Adaptation de la nomenclature niveau de traitements

Le cas des nomenclatures de produits, aussi appelé niveau de traitements, est emblématique de la diversité des nomenclatures utilisées pour désigner une même qualité de produit et par conséquence de la difficulté qu'un utilisateur peut rencontrer si il souhaite faire une recherche à partir de ce critère. Dans ce contexte, l'adaptation de la terminologie revêt tout son intérêt. La figure 3.23 donne un extrait des alignements réalisés autour de cette propriété pour mettre en correspondance les nomenclatures AIRBUS et USGS vers la terminologie utilisateur mise en place dans le cadre du projet GEOSUD.

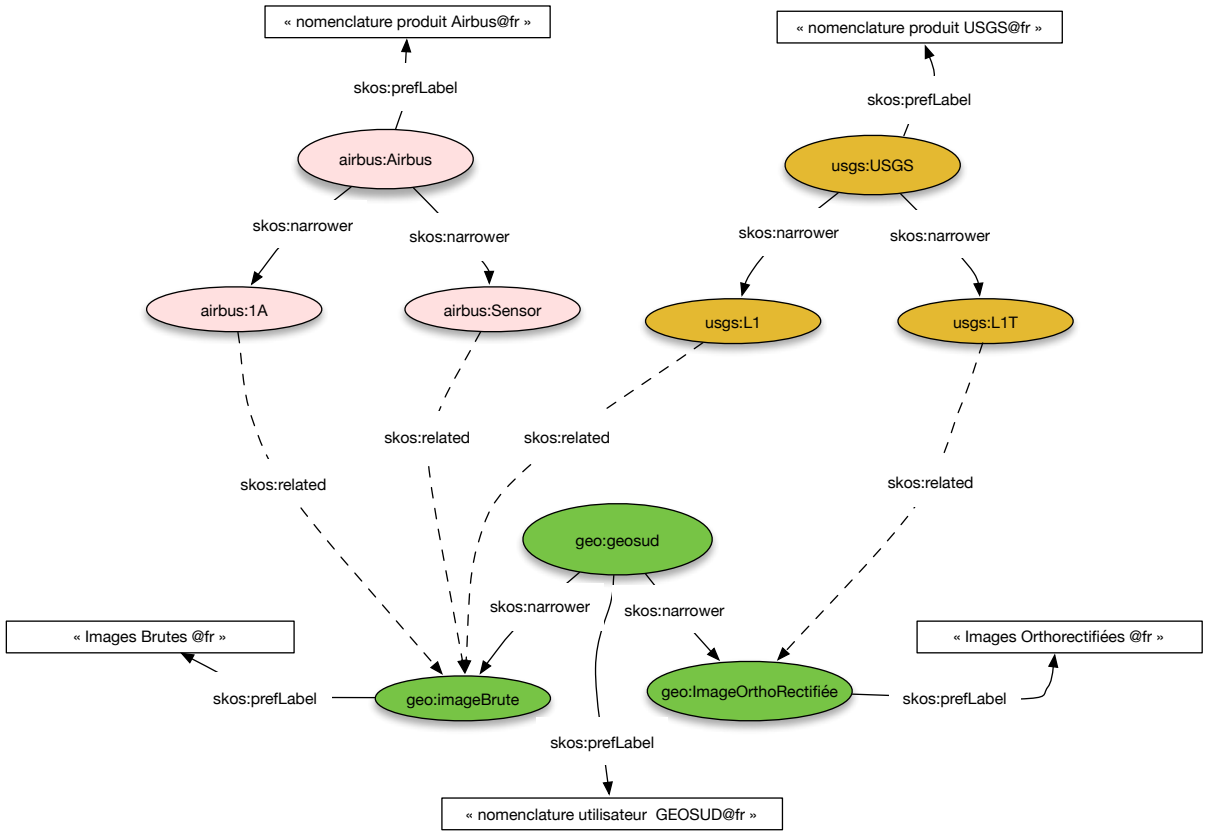


Figure 3.23: Extrait de l'alignement de la nomenclature des niveaux de traitements AIRBUS et USGS vers la nomenclature « utilisateur » retenue pour le projet GEOSUD

Adaptation pour la mise à disposition de classe de valeurs

Dans le cas d'une propriété d'image définie par une valeur numérique comme par exemple la résolution, l'angle d'incidence ou encore l'ennuagement lors de la prise de vue, il s'agit d'aligner les valeurs numériques vers des classes de valeurs communément adoptées par la communauté d'utilisateurs. C'est le cas par exemple de la résolution spatiale d'une image pour laquelle on souhaite que l'utilisateur l'appréhende à travers des catégories de résolution : basse résolution, moyen, haute et très haute résolution. Cette vision met en correspondance, de manière implicite, l'adéquation d'une

image au besoin applicatif. Le schéma en figure 3.24 donne un extrait de l'alignement proposé pour la propriété résolution spatiale d'une image optique vers les domaines de résolution.

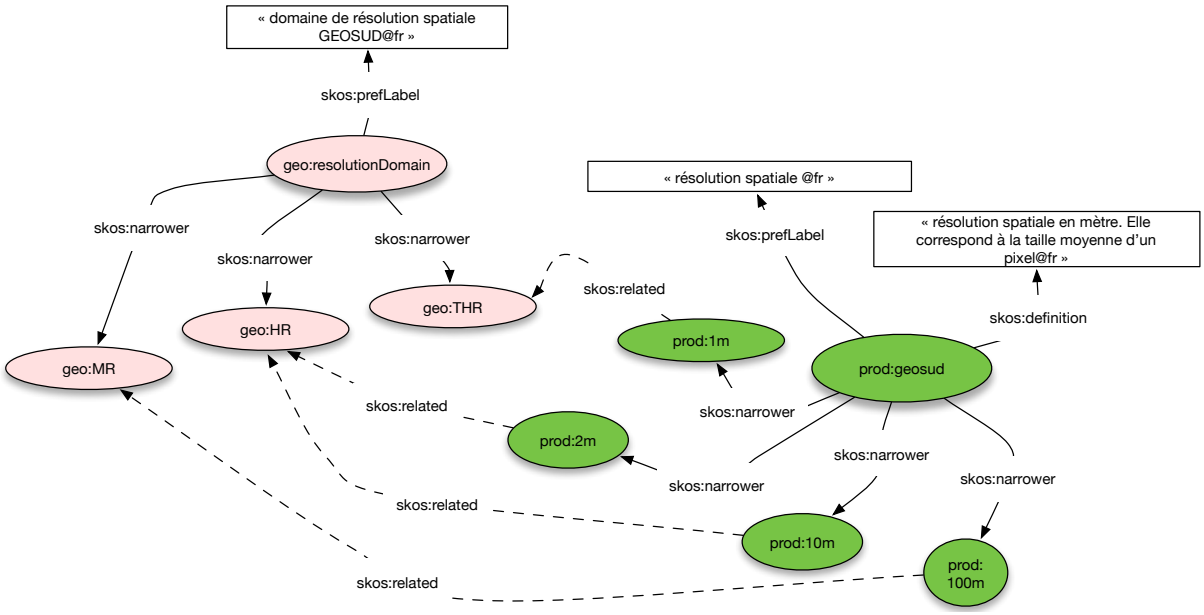


Figure 3.24: Extrait alignement entre les tailles de pixel et les catégories de résolution spatiale retenue pour le projet GEOSUD

3.5.1.2.4 Discussions

L'utilisation des thésaurus comme support à l'indexation relève bien du rôle attendu. La mise à profit des relations sémantiques apportées par le vocabulaire SKOS nous permet d'adapter les nomenclatures aux visions des utilisateurs finaux et de pouvoir être en capacité de le déployer dans d'autres contextes. L'apport et la flexibilité du vocabulaire SKOS nous permettent, en effet, de traiter nos problématiques d'adaptation à faible coût au regard de la complexité de mise en œuvre, de mise à jour et d'exploitation des thésaurus SKOS au sein d'un système opérationnel. Toutefois, l'expressivité relativement limitée et l'amalgame fait entre descripteur d'un concept (dimension terminologique) et concept (dimension ontologique) ne permettent pas d'envisager la définition et l'exploitation plus précises des relations sémantiques entre concepts qui seraient nécessaires pour mettre en relation les caractéristiques d'images avec les connaissances « métier » des acteurs de l'environnement.

3.5.2 Enrichissement des métadonnées par les référentiels spatiaux

3.5.2.1 Principes

Le principe est simple. Il s'agit de confronter l'emprise spatiale d'une image au découpage administratif d'un territoire et annoter l'image avec les unités administratives recoupant spatialement son emprise, et ceci à différents niveaux de hiérarchie. Par exemple, si l'on se situe sur le territoire français et que l'emprise de la donnée couvre une partie du département de l'Hérault, il sera pertinent que l'on annote une image SPOT6 (fauchée[56] d'environ 60 km) par les unités administratives du

[56] Fauchée : La largeur de la bande balayée, le long d'un parcours correspondant à la trace du satellite, définit la fauchée au sol.

niveau région Occitanie, département Hérault, ainsi que les différentes communes qui sont contenues dans l'emprise spatiale de l'image. Nous illustrons cet exemple en figure 3.25. Cette image, située à la pointe Bretonne pourra être enrichie des noms de la région Bretagne, du département du Finistère et des trentaines de communes incluses ou recoupant son emprise.

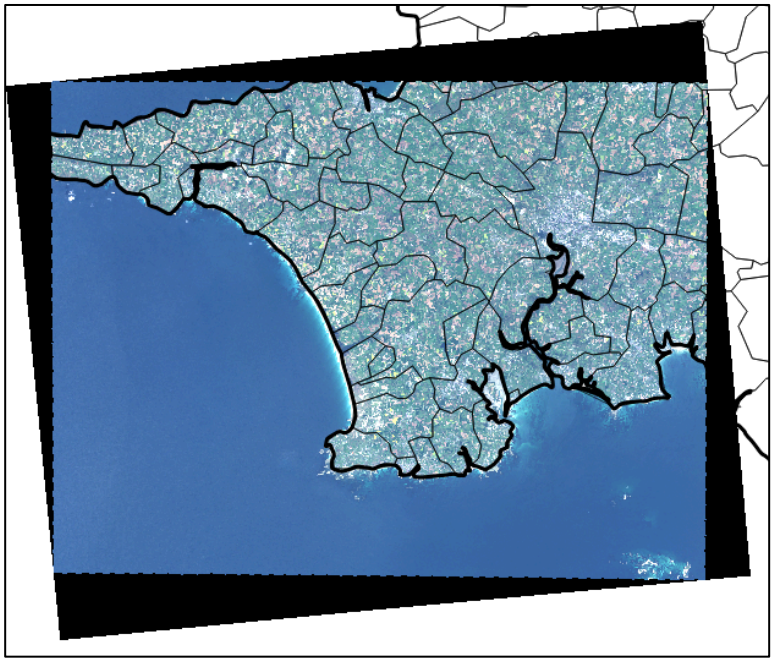


Figure 3.25: Emprise spatiale d'une image SPOT6. Les limites des communes et des départements qui se découpent sur l'image sont les unités administratives qui sont utilisées pour annoter les métadonnées de cette image. Les deux couches d'information sont projetées en RGF93.

Notre approche s'apparente à une fonction de géocodage inversée[57] qui prend en argument l'emprise spatiale d'une image et retourne une liste de libellés d'unités administratives qui sont contenues ou recoupent cette emprise. Les libellés des unités administratives ainsi récupérés viendront enrichir la description de l'image pour être ensuite indexés. Ils pourront ainsi être présentés à l'utilisateur sur la dimension de recherche *Espace*. L'opération d'affectation des unités administratives à une image est rendue possible si l'on peut extraire l'emprise spatiale de ses métadonnées et accéder à un référentiel présentant le découpage administratif du territoire sur lequel la collection d'images à annoter se situe. De nombreux référentiels spatiaux existent et sont rendus accessibles soient par leur téléchargement ou soient par l'interrogation via un service web.

3.5.2.2 Méthode d'affectation des unités administratives

Nous envisageons l'affectation d'une unité administrative à une image en considérant la pertinence que peut avoir cette annotation sur la recherche spatiale d'une image. En pratique, annoter une image à partir d'une unité qui la recouvrirait avec un très faible pourcentage produirait des résultats de recherche avec une faible précision. Par exemple, un utilisateur qui souhaite produire une cartographie sur le département de la Gironde et qui recevrait des images dont l'emprise spatiale est

[57] Géocodage inversé : Le géocodage inversé (ou reverse geocoding) consiste à effectuer l'opération inverse du géocodage, c'est-à-dire d'attribuer une adresse à des coordonnées géographiques

pour une grande part située sur d'autres départements considérera à juste titre que les résultats retournés sont d'une faible précision. Pour éviter cet écueil, il est nécessaire, suite à la détermination des unités administratives contenues, recoupant ou contenant une image, d'évaluer la pertinence d'affecter ou non cette unité à une image. Trois éléments sont utilisés pour la déterminer. Le premier est l'aire de l'intersection d'une unité administrative avec une image A_{int}. Le second est le rapport de cette intersection avec l'aire de l'unité administrative $Aire\text{-}int/Aire\text{-}unitAdm$. Enfin, le troisième est le rapport entre l'aire de l'image et celle de l'unité $Aire\text{-}int/Aire\text{-}img$. La méthode d'affectation repose sur le pré requis fondamental en géographie qui consiste à confronter les géométries des unités administratives et celle de l'emprise de l'image au sein d'un même référentiel spatial. En l'occurrence, les calculs d'affectation sont réalisés à partir des coordonnées géographiques, système géodésique global WGS 84. Bien que l'utilisation d'un système de projection planimétrique conforme apporterait une précision accrue dans le calcul des superficies, nous admettons l'imprécision dû à l'utilisation de coordonnées géographiques.

La figure 3.26 récapitule les principaux cas de figure qui peuvent être rencontrés :

- cas 1 : l'unité administrative est supérieure à l'aire de l'intersection,
- cas 2 : l'unité administrative est totalement incluse dans l'image,
- cas 3 : l'image est totalement contenue dans l'unité administrative.

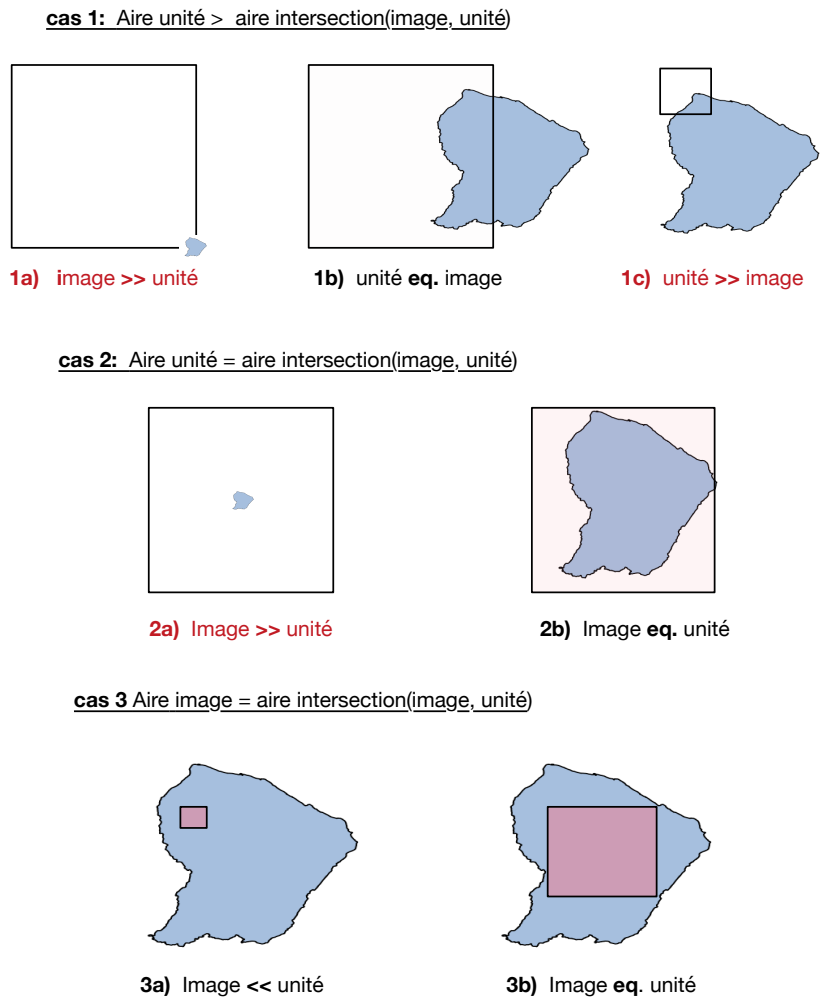


Figure 3.26: Différents cas considérés pour affecter une unité administrative à une image.

Pour les cas *1b*, *2b*, *3b*, seul le rapport *Aire-int/Aire-unitAdm* est proche ou supérieur à 1 et suffisant pour évaluer la pertinence. En effet, on peut considérer que l'unité administrative représente une partie significative de l'image. Pour traiter les cas *1a*, *2a*, *3a* où le rapport *Aire-int/Aire-unitAdm* est faible ou tend vers zéro alors le rapport *Aire-int/Aire-img*, c'est à dire quand une très faible partie de l'unité administrative est contenue ou recoupe l'image, est évalué.

La méthode proposée utilise successivement ces deux rapports *Aire-int/Aire-unitAdm* *et* *Aire-int/Aire-img* pour affecter une unité administrative à une image donnée. Pour cela, il est nécessaire de fixer une valeur seuil pour chacun d'entre eux. Elles sont déterminées et ajustées selon les collections de données indexées et l'échelle des découpages administratifs utilisés. Nous donnons dans la figure 3.12 le détail des opérations effectuées pour affecter puis enrichir la métadonnée de l'image.

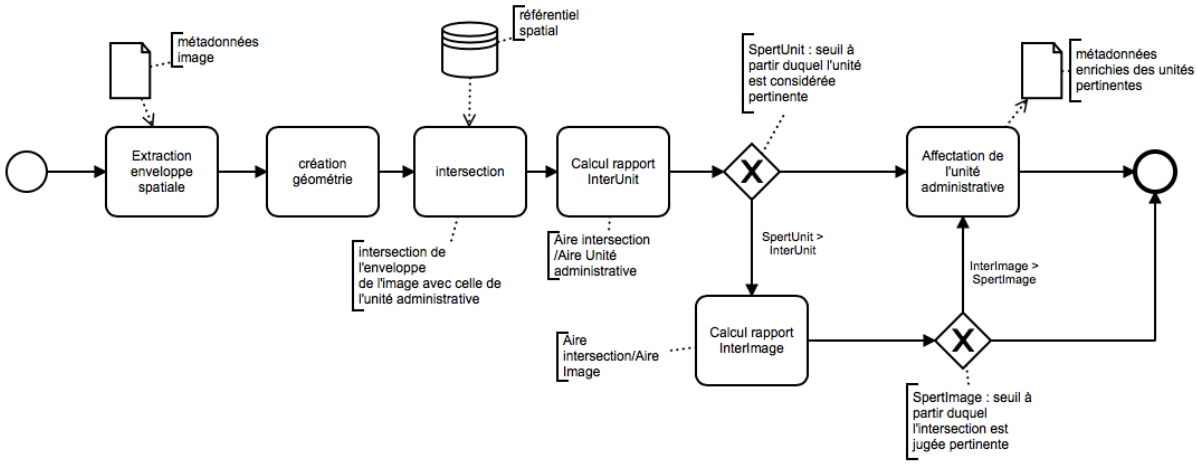


Figure 3.27 : Séquence des opérations effectuées pour affecter puis enrichir la métadonnée avec les unités administrative pertinentes (Diagramme BPMN).

Nous illustrons par deux exemples les résultats obtenus. Dans le premier, il s'agit d'enrichir les métadonnées d'un catalogue d'images satellitaires haute résolution en utilisant le découpage administratif de l'INSEE. Le deuxième exemple porte sur un catalogue référençant des ressources très hétérogènes sur lesquelles un référentiel proposant un découpage politique à l'échelle mondiale a été utilisé. L'analyse de ces résultats nous permet de discuter des limites de notre approche et d'envisager des pistes d'amélioration.

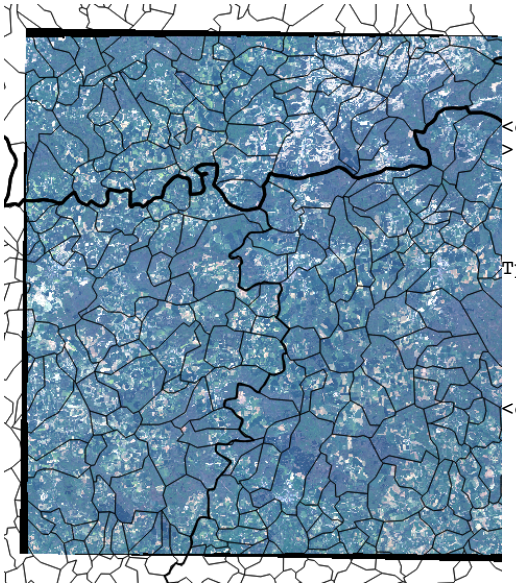
3.5.2.3 Annotation d'un catalogue d'images satellitaires haute résolution

Réalisé dans le contexte du projet GEOSUD, l'annotation a porté sur une collection d'images satellitaires relativement homogènes d'un point de vue de leurs caractéristiques. En effet, ces images satellitaires à haute résolution SPOT5, SPOT6 et Rapid Eye ont une emprise au sol d'une largeur d'environ entre 60 et 80 km, qui correspond à la fauchée du satellite. Leur longueur peut varier entre 60 et 600 km si l'on considère une scène ou une bande d'acquisition. Cela correspond à des emprises au sol allant de 3600 à 36000 km^2. Le référentiel utilisé pour annoter ces images est celui proposé par l'IGN : GEOFLA[58] Il décrit l'ensemble des unités administratives nationales (France métropolitaine et

[58] GEOFLA : http://professionnels.ign.fr/geofla

DROM): communes, cantons, arrondissements, départements, régions sous forme de fichiers de géométrie. Les seuils de pertinence, en relation avec la gamme de superficie d'images ont été fixés à 0.1 et 0.05. Ils assurent l'annotation sur les niveaux régions, départements et communes.

La figure 3.28 superpose l'emprise d'une image SPOT6 et les géométries du découpage administratif et donne un extrait des annotations ajoutées aux métadonnées de cette image.




```
…
<gmd:geographicIdentifier>
   <gmd:RS_Identifier>
     <gmd:code>
<gco:CharacterString>BOURGOGNE</gco:CharacterString
>
     </gmd:code>
   <gmd:codeSpace>
   <gco:CharacterString>
  http://…/cog.xsd#xpointer(//*[@name='CodeRegion
Type'])
  …
  <gmd:geographicIdentifier>
     <gmd:RS_Identifier>
       <gmd:code>
<gco:CharacterString>YONNE</gco:CharacterString>
     </gmd:code>
   <gmd:codeSpace>
   <gco:CharacterString>
  http://…/…(//*[@name='CodeDepartementType'])
  …
  <gmd:geographicIdentifier>
     <gmd:RS_Identifier>
       <gmd:code>
<gco:CharacterString>TANLAY</gco:CharacterString>
     </gmd:code>
   <gmd:codeSpace>
   <gco:CharacterString>
  http://…/…(//*[@name='CodeCommuneType'])
```

Figure 3.28 : Découpage administratif sur l'emprise d'une image SPOT6 d'avril 2017. Un extrait des métadonnées (XML ISO 19139) présente les annotations géographiques ajoutées. Les deux couches d'information sont projetées en RGF93.

3.5.2.4 Annotation d'un catalogue de ressources hétérogènes

L'annotation considérée, a été réalisée dans le contexte du projet européen EOPOWER[59], pour le compte de l'organisation intergouvernementale GEO[60]. Il porte sur une collection de ressources de renforcement de capacités très hétérogènes (document, données, organisation). De plus, ces ressources sont localisées sur 4 quatre des cinq continents (Desconnets et *al.*, 17a), et géolocalisées pour la plupart au travers d'un couple de coordonnées géographiques ou bien d'un rectangle englobant. De même, leur emprise spatiale peut porter aussi bien sur un territoire restreint que s'étendre à plusieurs pays, voire à un continent. Le référentiel utilisé pour l'annotation est la base géographique mondiale Geonames[61]. Elle décrit, entre autres, les unités administratives à l'échelle de la planète, sous la forme d'une hiérarchie agrégative. C'est cette hiérarchie qui est utilisée. Les seuils de pertinence ont été fixés à 0.2 et 0.1 pour permettre d'annoter les ressources sur les 2 premiers niveaux de la hiérarchie :

59 EOPOWER : Earth Observation for Economic Empowerment (projet européen FP7)
60 GEO : intergovernmental Group on Earth Observation, https://www.earthobservations.org/
61 Geonames : http://www.geonames.org/

continent, pays, cela en cohérence avec l'objectif du projet. Le référentiel Geonames est accédé à partir de son service web qui expose les fonctions nécessaires pour mettre en œuvre l'algorithme d'affectation. La figure 3.14 présente l'emprise de la succursale polonaise d'ASTRIUM Limited qui correspond à la Pologne. Le listing donne un extrait des annotations ajoutées aux métadonnées à savoir le continent Europe et la Pologne.


```
…
<gmd:geographicIdentifier>
   <gmd:RS_Identifier>
     <gmd:code>
<gco:CharacterString>Europe</gco:CharacterString>
     </gmd:code>
   <gmd:codeSpace>
   <gco:CharacterString>
http://www.geonames.org/
…
<gmd:geographicIdentifier>
   <gmd:RS_Identifier>
     <gmd:code>
<gco:CharacterString>Poland</gco:CharacterString>
     </gmd:code>
   <gmd:codeSpace>
   <gco:CharacterString>
http://www.geonames.org/
…
```

Figure 3.14 : découpage administratif sur la localisation d'une organisation privée ASTRIUM Ltd. Un extrait des métadonnées (XML ISO 19139) présente les descripteurs géographiques ajoutés.

3.5.2.5 Discussion

Constats et limites

Après plusieurs tests sur divers prototypes et la mise en œuvre opérationnelle dans les deux catalogues mentionnés plus haut, la méthode d'affectation est relativement robuste et génère peu d'incohérences dans les annotations de métadonnées. Par contre, sa robustesse repose essentiellement sur la détermination experte des valeurs de seuils qui dirigent l'enrichissement des unités administratives aux jeux de données. Les incohérences introduites, par exemple une ressource affectée à un pays ou à un continent dans lequel elle n'est pas située, apparaissent lorsque la géométrie de l'emprise de la ressource a été déterminée de manière imprécise, manuellement. Il est également à noter que l'enrichissement de jeux de données à large emprise (plusieurs dizaines de milliers de kilomètres) à partir de référentiel à grande échelle peut potentiellement générer la récupération de plusieurs centaines ou milliers d'unités. Ce qui peut fortement augmenter les temps d'indexation, la taille de l'index et être préjudiciable à terme aux temps de réponse à une requête utilisateur. Dans ce cas, la détermination de l'arrêt de l'annotation au bon niveau de hiérarchie devient essentielle. Enfin, la mise en œuvre d'une telle approche demande d'avoir à disposition ou d'accéder de manière performante et continue à un référentiel mis à jour. Une adéquation est à trouver entre accès via un service web et le téléchargement du référentiel.

Pistes d'amélioration

La première amélioration possible porte sur la détermination de la pertinence d'affecter ou non une unité administrative à un jeu de données. La méthode actuelle propose une détermination par seuil fixé arbitrairement. Une première alternative est d'affecter toutes les unités incluses ou qui intersectent la donnée et d'y associer leur degré de couverture de l'emprise du jeu de données considéré. Ce dernier pourrait être calculé à partir du rapport Aire d'intersection de l'unité/ surface de l'emprise du jeu de données. Cette valeur pourrait alors être utilisée pour améliorer le classement des résultats d'une recherche conduite par la dimension spatiale. L'utilisation de la théorie des ensembles flous pourrait également être mise à profit dans un même objectif.

Enfin, des référentiels tels que Geonames offrent, à travers les modèles ontologiques sur lesquels ils sont construits, des descriptions qui portent sur les relations hiérarchiques, d'inclusion ou de voisinage entre les objets. Il semble aisé d'envisager de compléter les annotations actuelles réduites aux libellées des objets par les relations de subsomption et de voisinage. Elles pourraient être mises à profit pour étendre une recherche de données à l'ensemble d'une unité administrative, à ses voisines ou encore au niveau hiérarchique supérieur.

Généralisation à d'autres référentiels spatiaux

La démarche peut être potentiellement étendue à d'autres référentiels. En effet, selon le contexte du projet, des référentiels mettant à disposition le découpage d'un territoire selon son occupation du sol, les aires protégées, les zones de gestion se rapportant à une activité économique peuvent être traités et assurer autant de nouvelles clés de recherche sur des collections de données spatio-temporelles. Aujourd'hui, un certain nombre de ces découpages sont facilement mobilisables et interrogeables via des services géographiques web standardisés, notamment OGC WFS (Web Feature Service). C'est le cas par exemple des classifications Corine LandCover[62] ou encore les aires protégées qui ont été mises à disposition par chacun des états membres dans le cadre de la directive INSPIRE.

3.5.3 Recherche à facettes

Les propositions associées à l'indexation guidée par les thésaurus nous permettent d'envisager la mise en œuvre d'une recherche à facettes sur les collections d'images satellitaires. En effet, elles nous ont amené à construire un index basé sur des terminologies se rapprochant du point de vue de l'utilisateur et offrant également un système de classification multi dimensionnel des images. Pour cela, l'indexation préconisée permet de lever deux des principaux verrous pour envisager un tel système de recherche, à savoir posséder des descriptions structurées, aussi appelées métadonnées facettées, et adapter le vocabulaire à l'utilisateur.

Dans cette section, nous présentons les choix sur lesquels repose la recherche à facettes pour des collections d'images satellitaires. Nous les justifions au regard de la connaissance que nous avons du comportement des utilisateurs cibles (cf. Contexte, section 3.2), des bases conceptuelles fournies par (Ingwersen et Wormell, 92, Ellis et Vasconcellos, 00, Broughton, 00), et les expériences relatées par divers auteurs : (Yee et *al*., 03 ; Denton, 03) qui ont mis en œuvre une telle approche ou l'ont analysé. Un de nos objectifs est de fournir les clés pour réutiliser et prolonger notre approche dans des contextes similaires. Pour cela, nous justifions le choix d'associer une navigation à facettes avec une

[62] Corine Land Cover WFS : CLC.developpement-durable.gouv.fr/geoserver/wfs?

recherche avancée, à base de formulaire. Nous donnons également les éléments pour comprendre nos choix tant en matière de facettes qu'en matière d'ergonomie. Finalement, un patron de conception des interfaces utilisateur est proposé.

Les questions ouvertes relatives à la performance des systèmes à facettes soulevées par (Tunkelang, 09) ne sont pas abordées. Nous considérons qu'elles ne constituent plus aujourd'hui un obstacle à la mise en œuvre d'un tel système. Les outils open source actuels assurent le passage à l'échelle vers de très grands ensembles de données.

3.5.3.1 Approche de recherche adoptée

Bien que la motivation ne soit pas identique, l'approche de recherche adoptée peut être définie comme une recherche à facettes si l'on se réfère à la définition de (Tunkelang, 09). Il la définit comme la combinaison d'une navigation à facettes complétée d'une recherche sur texte libre. Cette combinaison permet d'assurer une recherche à la fois sur des descriptions structurées et appuyées sur un système de classification (les facettes) et sur des descriptions non structurées comme des descriptions textuelles (résumé, titre, …). Nous adoptons une recherche à base de navigation à facettes couplée, selon le besoin, à une recherche booléenne que nous qualifions de « recherche avancée ». Cette combinaison nous permet de couvrir des scénarii de recherche non permis par les facettes : comme par exemple, la recherche d'images sur une emprise spatiale différente d'une unité administrative ou/et la recherche simultanée sur deux extensions temporelles (intra et inter annuelles) : « je veux toutes les images acquises au mois de juin depuis 2000 jusqu'à 2015. Cette dernière requête est particulièrement utile pour récupérer des images et mener des études diachroniques sur un territoire. Les principaux scénarii de recherche que nous souhaitons couvrir peuvent être résumés ainsi :

- Scénario 1 : recherche à base de navigation à facettes : la requête est formulée sur les dimensions présentes dans les facettes à savoir la zone d'étude (dimension Espace S), la date d'acquisition (dimension Temps T) et certains propriétés de l'image tels que la résolution, le niveau de traitements. Les utilisateurs avancés pourront également utiliser les facettes relatives aux conditions d'acquisition (dimension Procédure P) comme la plateforme, le capteur d'acquisition ou encore l'angle d'incidence,
- Scénario 2 : recherche spatiale couplée à une navigation à facettes : l'utilisation de la recherche avancée permet de définir une emprise spatiale quelconque. La navigation par facettes sur les dimensions STP permet ensuite d'affiner la recherche spatiale,
- Scénario 3 : recherche temporelle couplée à une navigation à facettes : l'utilisation de la recherche avancée permet de définir une emprise temporelle quelconque. La navigation par facettes sur les dimensions STP permet ensuite d'affiner la recherche temporelle,
- Scénario 4 : recherche spatio-temporelle couplée à une navigation à facettes : il correspond à la combinaison des scénarii 2 et 3.

3.5.3.2 Présentation et choix des facettes

L'expérience révèle que l'utilisation des facettes dans une application de recherche doit être réfléchie pour qu'elle remplisse pleinement et efficacement son rôle. Comme le souligne (Simon, 71) *"a wealth of information creates a poverty of attention and a need to allocate that attention efficiently"* (Une mine d'informations crée une pauvreté d'attention et une nécessité d'allouer efficacement cette attention). Plus prosaïquement et selon l'adage bien connu « trop d'informations, tue l'information », l'enjeu à considérer est : quelle facette et sous quelle forme doit-on la présenter à l'utilisateur ?

Deux facteurs sont à prendre en compter pour guider cette réflexion : le nombre de facettes et le nombre de valeurs mises à disposition. Le nombre de facettes reflète les différents chemins à partir desquels les données sont classifiées. Logiquement, notre réflexion s'appuie sur la classification à facettes préconisée plus haut en section 3.3. Elle identifie trois dimensions : Espace, Temps, et Procédure que l'on a pu mettre en correspondance avec les différents niveaux d'expertise d'un utilisateur (cf. tableau 3.2, section 3.3.3).

En premier lieu, et comme le souligne (Cubranic, 08 et Koren et *al*., 07), il est nécessaire de privilégier des facettes qui couvrent la plus grande majorité des documents avec si possible une distribution la plus uniforme. La présence répétée d'une valeur « valeur inconnue » en réduit considérablement l'intérêt. Il en est de même, si une valeur de facettes représente 99% de la collection de données. Pour cela, les dimensions Espace et Temps sont bien entendu des candidats à privilégier car elles sont construites sur des métadonnées élémentaires, même si l'on verra plus bas que la dimension Espace exprimée à travers une hiérarchie administrative nécessite une attention toute particulière.

Concernant le second point, il est effectivement important à ne pas fournir à l'utilisateur, de manière simultanée, de trop grandes listes de valeurs. Idéalement, une facette avec un grand nombre de valeurs doit être hiérarchique. En effet, la hiérarchie pourra être présentée progressivement à l'utilisateur. Après la sélection d'une valeur racine, uniquement les fils de cette valeur pourront être présentés par l'interface utilisateur. Pour ce qui concerne des facettes non hiérarchiques, plusieurs techniques sont envisageables (Tunkelang, 09), notamment afficher les valeurs ayant la plus haute fréquence au sein de la collection ou encore construire des plages de valeurs pour des facettes numériques. C'est de cette manière que nous avons traité la propriété telle que la résolution spatiale pour la dimension Procédure. Enfin, même si les moteurs de recherche proposant la gestion des facettes sont aujourd'hui performants, les intersections à exécuter sur les différentes dimensions lors d'une requête sont coûteuses en temps de calcul. Aussi, on privilégiera, pour maintenir une réelle simultanéité dans l'interaction filtrage/retours des résultats, un nombre limité de facettes. Face à ces recommandations et pour illustrer nos choix qui restent néanmoins liés à l'expertise du concepteur, trois exemples de facettes sont détaillés. Ils sont mis en relation avec les efforts d'enrichissement et d'adaptation effectués lors de l'indexation.

Facettes hiérarchiques

La facette portant sur la dimension Espace est une illustration intéressante de l'usage d'une facette hiérarchique. Comme nous l'avons proposé en section 3.4.3, cette facette est construite sur une hiérarchie administrative de plusieurs niveaux et comptant par exemple en France plus de 40 000 unités administratives sur une profondeur d'au moins 3 niveaux : Région, Département, Commune. La recherche sur cette facette se fera par sélection de valeurs niveau après niveau. Le choix de la région Occitanie nous permettra de présenter les départements Aveyron, Hérault et ainsi de suite dans le cas où des données y sont rattachées. Les niveaux inférieurs portent sur des unités administratives de faible superficie et doivent être traités spécifiquement car ils peuvent présenter potentiellement une grande quantité de valeurs. Par exemple, un département tel que la Gironde compte 540 communes. Deux options sont envisageables : la construction de plage de valeurs permettant de regrouper par ordre alphabétique. Une deuxième option est d'associer à cette facette un champ texte libre dans lequel pourra être saisi le nom de la commune. Nous illustrons ces deux choix dans la figure 3.30 ci-dessous.

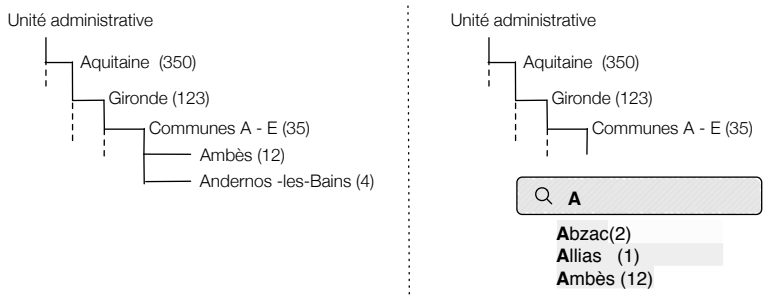


Figure 3.30: Exemple d'une facette hiérarchique, unité administrative de la dimension Espace.

Un autre exemple concerne la facette portant sur les propriétés Plateforme et Instrument de la dimension Procédure. Ces deux propriétés ne peuvent être dissociées, un instrument ne correspond qu'à une plateforme. Une facette hiérarchique peut être construite sur deux niveaux : le premier propose les valeurs de plateformes, le choix d'une plateforme permet à l'utilisateur de faire apparaître le deuxième niveau, les instruments associés sur lesquels il pourra raffiner sa recherche. Issu de deux éléments de métadonnées respectifs, les associations entre plateforme et instrument sont fournies par le référentiel « producteur » au travers de relation hiérarchique *skos:narower*/*skos:broader* (figure 3.31).

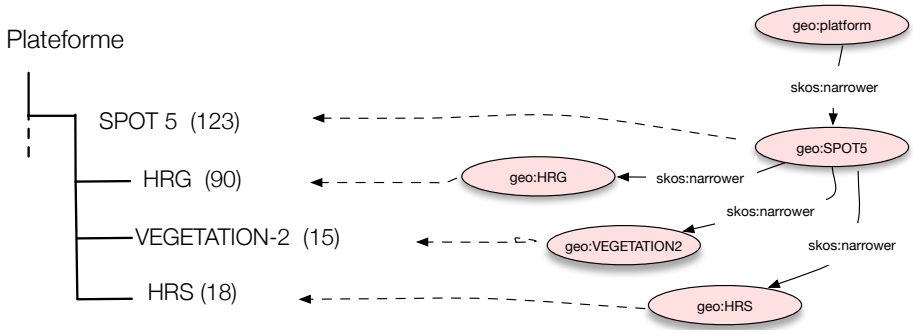


Figure 3.31: Exemple de facette hiérarchique sur les propriétés Plateforme et Instrument. L'exemple concerne le satellite SPOT5 qui réalise des prises de vue à l'aide de trois instruments : HRG, VEGETATION-2 et HRS. Nous mettons en regard un extrait de la hiérarchie correspondante telle qu'elle peut être définie dans un thésaurus SKOS.

Facettes de type « plage de valeurs»

Les facettes « plage de valeurs » (*Range*) permettent de regrouper les valeurs numériques en catégories. Ce type de présentation se prête particulièrement aux propriétés portant sur les paramètres d'observation ou conditions d'acquisition (dimension Procédure) qui sont décrit par des types numériques : entier, réel, réel double. Dans ce cas, la construction des plages de valeurs peut être guidée soit par la distribution statistique des valeurs de la propriété au sein de la collection d'images ou peut faire l'objet d'une représentation au sein du référentiel « utilisateur ». La figure 3.32 montre la facette « angle d'incidence ». La dimension Temps utilise une présentation par plage de valeurs en regroupant la date d'acquisition par année.

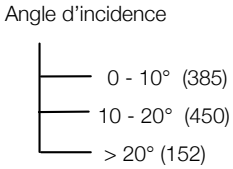


Figure 3.32: Exemple de facette «plage de valeurs » : Angle d'incidence de l'instrument lors de la prise de vue. Les plages de valeurs sont celles proposées pour la recherche d'images optique.

Facettes de type « énumération »

Les propriétés décrites par une liste de valeurs réduite peuvent être représentées par une facette de type énumération (*Field*). Son usage demande de contrôler a priori la validité des chaînes de caractères au regard du domaine de valeurs défini pour chaque élément de métadonnées. Néanmoins, et comme nous l'avons détaillé dans la section adaptation de nomenclatures discriminantes, les énumérations présentées peuvent être le résultat d'une adaptation de valeurs numériques ou de code d'identification en une catégorie qui a du sens pour l'utilisateur. C'est le cas des propriétés résolution spatiale ou niveau de traitement pour lesquelles l'alignement entre les concepts « producteur » et les concepts « utilisateurs » assurent cette adaptation. La figure 3.33 donne une illustration d'une facette construite sur une énumération portant sur le niveau de traitements des images.

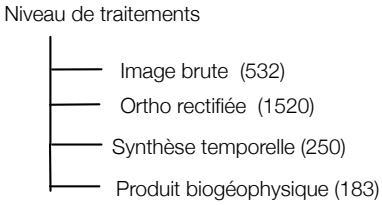


Figure 3.33: Exemple de facettes « énumération » : Niveau de traitements. Extrait de la nomenclature proposée par le pôle de données Surfaces Continentales THEIA.

Nous récapitulons dans le tableau 3.3, les différentes facettes proposées pour la recherche dans les collections d'images satellitaires optiques. Cette liste n'est pas exhaustive. Elle se base sur les travaux entrepris dans le cadre du projet GEOSUD (Kazmierski et *al.,* 14).

Tableau 3.3: Récapitulatif des facettes élémentaires préconisées pour une recherche dans des collections d'images satellitaires optiques

Dimension de recherche	Facette	Type de facettes	Commentaires
Espace	Unité administrative	Hiérarchique	La hiérarchie est construite à partir des relations décrites dans le référentiel utilisé
Temps	Date d'acquisition	Plage de valeurs	La plage choisie est l'année
Procédure	Résolution spatiale Domaine spectral Niveau de traitements	Enumération Enumération Enumération	Les énumérations sont le résultat d'une adaptation d'un code ou d'une valeur numérique vers une nomenclature utilisateur
	Angle d'incidence Mode opérationnel Ennuagement Plateforme/instrument	Plage de valeurs Enumération Plage de valeurs Hiérarchique	La hiérarchie Plateforme/Instrument est construite à partir du référentiel terminologique « producteur »

3.5.3.3 Lignes directrice pour l'ergonomie de l'IHM de recherche

De manière concomitante à la présentation des facettes, l'ergonomie et la dynamique de l'interface utilisateur font partie de nos préoccupations. Plusieurs points sont soulevés par Tunkelang, 09 pour aborder le sujet :

1- Quand et où doivent êtres présentées les facettes à l'utilisateur ?
2- De quelle manière organiser les facettes et leurs valeurs ?
3- Doit on utiliser débuter la recherche en s'appuyant sur une expression libre ?

Sur ce dernier point, rappelons que la forte expérience accumulée dans la gestion des métadonnées montre que les métadonnées portant sur les données spatio-temporelles sont relativement pauvres, voire dénuées de description textuelle libre (résumé, titre). Ceci est d'autant plus vrai lorsque les métadonnées sont produites de manière automatique comme celles des images satellitaires. Pour cela, la recherche sur texte libre a très peu d'intérêt. L'approche recommandée est davantage d'aborder la recherche par navigation sur les facettes. Elle pourra être complétée par une recherche avancée lorsque cela s'avère nécessaire.

Concernant le premier point, le parti pris est de mettre en relation les quatre dimensions de recherche dégagée par la classification à facettes et de privilégier, à la fois dans l'ordre de présentation et la visibilité immédiate, les facettes des dimensions Espace et Temps.

Les facettes des dimensions Procédure pourront être présentées dans un deuxième temps comme une option de raffinement de la recherche. Trois directions peuvent être envisagées :

1. Une première qui consiste à une approche assez répandue qui est de masquer les facettes de la dimension Procédure que l'on nommera « experte ». C'est l'utilisateur qui choisira de les afficher si il juge nécessaire le raffinement de sa recherche,
2. Une deuxième option « plus intelligente » sera d'afficher les facettes au regard du nombre de résultats retournés. Nous donnons alors le rôle à ses facettes de raffiner les résultats si ces derniers sont trop nombreux ou trop hétérogènes, c'est à dire contenant une grande diversité de propriétés,

3. Une dernière possibilité est de faire apparaître une ou plusieurs nouvelles facettes selon la valeur sélectionnée par l'utilisateur sur une facette active, approche que l'on appellera « contextuelle ».

Enfin, concernant la présentation des facettes dans l'IHM, Tunkelang, 09 préconise de les positionner horizontalement et au dessus des résultats pour des utilisateurs peu experts (figure 3.34). Ce positionnement facilite la mise en relation entre les raffinements effectués sur les facettes et les résultats obtenus. Un positionnement vertical juxtaposé aux résultats sera privilégié pour des utilisateurs plus familiers de tels environnements (figure 3.35).

Patrons de conception pour la recherche à facettes dans les images satellitaires

De nombreux sites fournissent des patrons de conception (*design pattern*) qui pourront appuyer la construction d'une IHM de recherche à facettes. Le site de Peter Morville[63] propose un certain nombre de patron orientés navigation à facettes. Atkinson, 05 fournit également des recommandations pour guider la conception d'une IHM. Ces exemples ont l'inconvénient de ne considérer que la recherche de collections de produits manufacturés ou de documents. Pour y remédier, nous complétons ces exemples avec deux patrons (figure 3.34 et 3.35) qui reprennent les deux alternatives ergonomiques proposées dans le précédent paragraphe et replace le formulaire de recherche avancée.

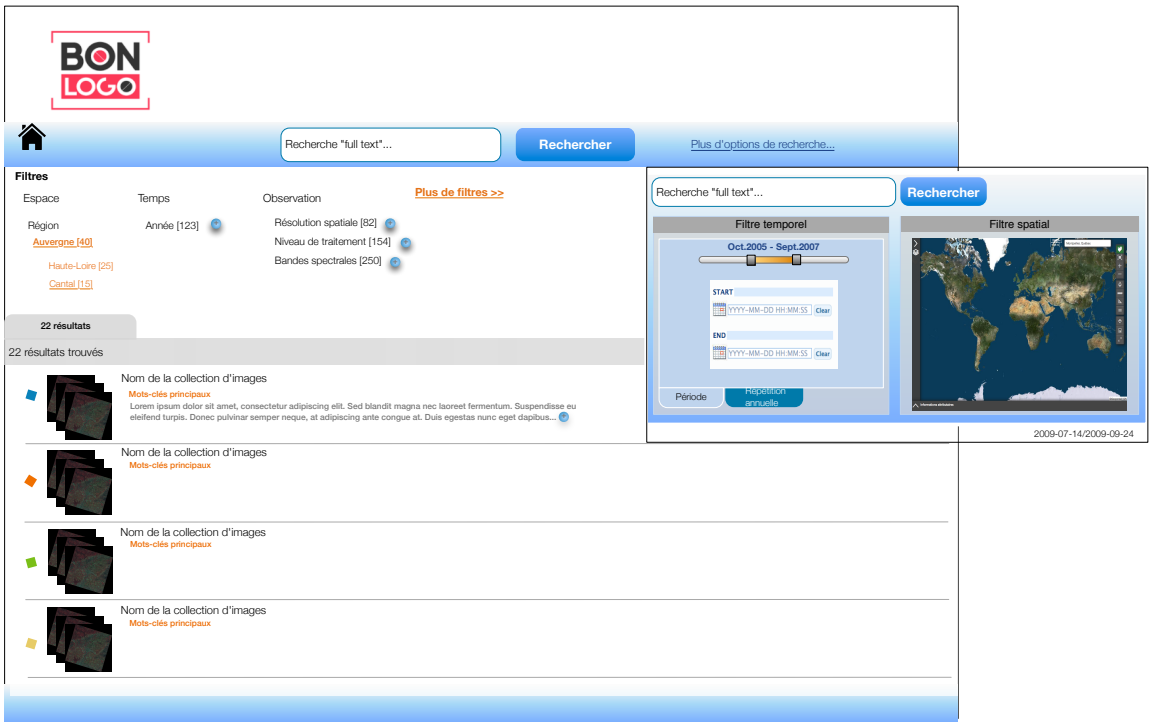


Figure 3.34: Patron de conception pour une recherche à facettes sur des collections d'images, à destination d'utilisateurs peu familiers des environnements de recherche. (d'après une maquette conçue par Geomatys en collaboration avec GEOSUD)

[63] site de Peter Morville : https://www.flickr.com/photos/morville/collections/72157603789246885/

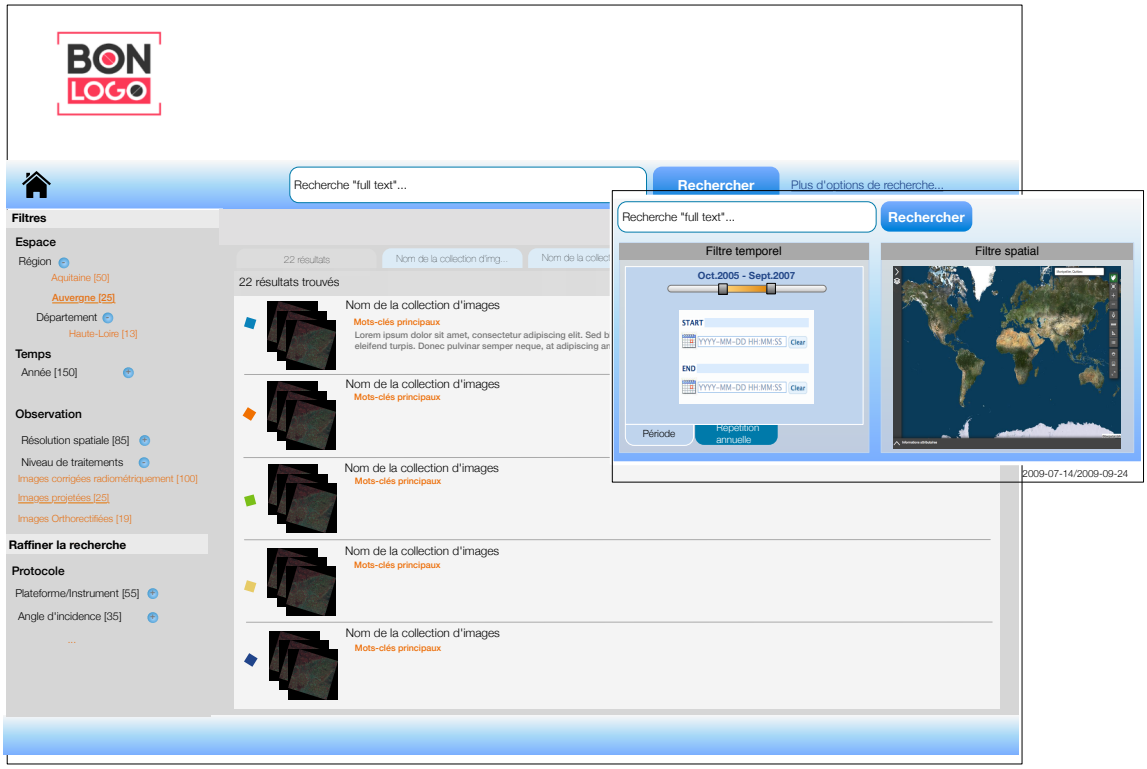


Figure 3.35: Patron de conception pour une recherche à facettes sur des collections d'images, à destination d'utilisateurs experts des environnements de recherche (d'après une maquette conçue par Geomatys en collaboration avec GEOSUD)

3.5.3.4 Discussion

Plus qu'une conceptualisation de ce que doit être une recherche à facettes, des pistes de conception sont proposées. Elles tiennent compte de l'expertise du domaine. Elle est notamment apportée par la classification à facettes proposée en section 3.3 et les comportements connus des utilisateurs auxquels s'adressent les applications. L'utilisation d'une recherche avancée pour couvrir davantage de scénarii de recherche est nécessaire mais réintroduit les limites d'une recherche paramétrique. La reformulation de requêtes lors de l'usage de la recherche avancée, notamment sur les dimensions spatiale et temporelle pourrait y remédier. La description des relations topologiques entre unités spatiales telles qu'elles peuvent décrites par les référentiels spatiaux constitue une piste à explorer. La reformulation de requêtes sur l'extension temporelle est à envisager par l'apport des outils de visualisation analytique (Keim et *al.*, 08 ; Maciejewski et *al.*, 10).

Enfin, quelques pistes restent à approfondir pour mieux guider l'utilisateur dans l'exploration des collections de données via la navigation par facettes. Nous pensons en particulier à reconsidérer la classement statique des valeurs de facettes en prenant en compte selon les cas : la représentativité de la valeur d'un point de vue statistique ou encore son indice de pertinence en fonction de la requête envoyée.

4. Exemple de mise en œuvre

Depuis 2012, tout ou partie des propositions déclinées dans les sections précédentes ont fait l'objet de mises en œuvre que ce soit dans l'implémentation de prototypes ou d'outils qui font aujourd'hui l'objet d'une utilisation opérationnelle. Chronologiquement, nous citerons l'outil de

catalogage et de recherche sur les missions spatiales du CNES : REFLECS dans lequel l'indexation sémantique guidée par des thésaurus dédiés a été mis en œuvre. Les projets européens FP7 GEONETCAB[64] puis FP7 EOPOWER[65] ont permis de prototyper puis déployer l'ensemble de la démarche préconisée dans ce mémoire. L'implémentation résultante, GEOCAB portal[66], des travaux menés sur ces deux projets est depuis 2015 mis à disposition de la communauté d'Observation de la Terre au travers de l'organisation inter gouvernementale GEO[67]. Il assure le catalogage et la recherche des ressources de renforcement de capacités (formation, tutoriel, programme d'accès aux images, service web de données) qui sont disséminées sur 4 continents (Desconnets et *al*., 17a). Enfin, l'infrastructure de données spatiales GEOSUD[68] ainsi que son extension nationale, la fédération d'infrastructures de données du pôle THEIA[69] ont mis en œuvre notre démarche sur de grands ensembles d'images satellitaires ou de données dérivées. Dans cette section, mon propos est de présenter quelques éléments de la mise en œuvre de notre démarche dans un contexte opérationnel : l'infrastructure de données spatiales GEOSUD.

4.1 Partenariat

La démarche entreprise autour des systèmes de recherche sur les données spatio-temporelles en environnement et leurs diverses implémentations ont été menées en partenariat avec la société Geomatys. Durant ce partenariat de recherche (2009-2013) puis le projet GEOSUD (2013-2016), les volets autour du catalogage et de la recherche d'information ont été développés. Ils ont été concrétisés par le développement de l'outil MDweb (Desconnets et Libourel, 12). Dans un deuxième temps, les principaux composants de MDweb ont fait l'objet d'un transfert vers l'outil open source Constellation-sdi[70]. Divers évènements de la sphère académique ou de celle des logiciels libres m'a permis, avec Geomatys, de les communiquer (Desconnets et *al*., 08 ; Desconnets et *al*., 09 ; Desconnets et *al*., 10). Plus récemment et dans le cadre du pôle national données des surfaces continentales (THEIA), des travaux de même nature ont été menés en partenariat avec le CNES, l'IGN et l'IRSTEA. Ils portaient sur la conception et l'implémentation d'une fédération de catalogues d'images satellitaires et son moteur de recherche.

4.2 Infrastructure de données spatiales GEOSUD

4.2.1 Contexte

Le projet GEOSUD est né du constat de sous utilisation de l'imagerie satellitaire par les acteurs du secteur public travaillant sur la gestion des milieux naturels. Aussi, un des ses principaux objectifs est de faciliter et garantir l'accès à des images satellitaires haute et très haute résolution dans des conditions facilitées. Une infrastructure de données spatiales (IDS) a été développée pour les mettre à disposition. En complément de la constitution de couvertures nationales issues de capteurs optiques

[64] GEONetCab : GEO Network for Capacity Building, FP7 projet

[65] EOPOWER : Earth Observation for Economic Empowerment

[66] GEOCAB portal : www.geocab.org

[67] GEO : Group of Earth Observation

[68] GEOSUD : GEOspatial for SUstainable Development

[69] THEIA : Pôle de données Surfaces Continentales, https://www.theia-land.fr/

[70] Constellation-sdi : http://constellation-sdi.org/

tels que Rapid Eye ou SPOT5 durant les années 2009 à 2015, le projet GEOSUD s'est doté d'une antenne de réception assurant l'acquisition d'images. Une unité de traitement de la télémesure acquise par l'antenne de réception permet d'assurer leur production et leur intégration de manière continue dans l'IDS. Elle doit également donner accès aux images très haute résolution PLEIADES détenues par l'IGN. Le flux annuel d'images produites et diffusées est de l'ordre de 3000 à 5000 soit des volumes annuels de plusieurs Teraoctets. La communauté des utilisateurs GEOSUD compte aujourd'hui plus de 500 entités adhérentes pour environ 1000 utilisateurs.

A plus d'un titre, le contexte du projet GEOSUD est propice à l'implémentation d'un système de recherche à facettes. En effet, plusieurs éléments rendent possibles et nécessaires sa mise en œuvre : le processus de production d'une image prévoit toujours la création de métadonnées ; les métadonnées produites sont hétérogènes et doivent être standardisées pour être exposées vers les services interopérables de découverte ; le flux quasi continu d'images produites par une antenne de réception demande de prévoir un composant qui puisse assurer de manière continue et automatique le traitement des images et des métadonnées en vue de produire les flux de données standard (CSW, WMTS[71], WCS). Enfin, les utilisateurs de l'IDS GEOSUD ne sont pue ou pas familiers des données issues de l'observation de la Terre.

4.2.2 Architecture

L'infrastructure de données GEOSUD repose sur le déploiement d'une architecture à base de services web géographiques standardisés. Il s'agit de garantir l'interopérabilité des données et des services avec d'autres infrastructures à l'échelle nationale (pôle THEIA), européenne (INSPIRE) et mondiale (GEOSS). Ce sont les implémentations des standards OGC CSW, OpenSearch pour les services de découverte et WMTS, WCS pour ce qui concerne les services d'accès aux données qui sont au cœur de l'interopérabilité entre ces systèmes. L'architecture de l'IDS GEOSUD offre aux utilisateurs un portail qui en est le point d'entrée. Il assure les fonctionnalités de recherche, de visualisation et de téléchargement des images.

La gestion et la diffusion standardisée des images reposent sur l'outil libre open source Constellation-SDI. Il assure la gestion des différents services géographiques OGC. Il propose également un environnement d'exécution de chaîne de traitements. L'implémentation de l'indexation guidée par les thésaurus tire partie de ces fonctionnalités. Elle s'appuie également sur l'implémentation d'une chaine de traitements qui assure les opérations d'enrichissement et d'adaptation terminologiques sur les métadonnées et la production des flux de données standardisés. Les référentiels terminologiques SKOS nécessaires à l'indexation sont exposés par un service web (RESTful) qui fournit les interfaces programmatiques (API) en lecture et écriture. Elles permettent ainsi d'en faire un composant autonome. Enfin, le moteur de recherche ElasticSearch vient compléter ces composants pour apporter les fonctions de stockage, d'analyse et d'indexation sur lequel sont exécutés les requêtes de l'application de découverte des images.

[71] WMTS : OGC Web Mapping Tiling Service

4.2.3 Chaîne de traitements pour la production automatique des flux standardisés

Dans le contexte de production continue d'images satellitaires, la présente chaîne de traitements permet d'assurer de manière automatique, cohérente le pilotage des opérations de transformation des données et des métadonnées en vue d'une part de les exposer via les flux standardisés CSW pour les métadonnées et WMTS, WMS et WCS pour les données et mettre à jour l'index ElasticSearch sur lequel l'application de découverte du portail vient poser ces requêtes.

La figure 3.36 décrit les principales étapes de cette chaîne. Les opérations portant sur les métadonnées assure leur standardisation via à un modèle pivot. La phase d'indexation est réalisée sur les métadonnées standardisées. Les étapes d'enrichissement et d'adaptation des métadonnées s'appuient sur un service web RESTful qui fournit l'accès aux thésaurus SKOS gérés par l'IDS GEOSUD et le référentiel administratif GEOFLA. Cette chaîne de traitements est exécutée sur un ensemble d'images regroupées au sein d'une collection[72]. Son exécution peut être ainsi paramétrée en relation avec les caractéristiques de la collection d'images. Le résultat de la séquence permet la mise à jour de l'index ElasticSearch et le service de découverte CSW. Enfin, elle met à jour un index spatial. Ce dernier stocke les géométries des emprises spatiales des images sous forme d'un index de type Rtree (Heurteaux, 16). Il est mis à contribution pour répondre aux requêtes spatiales construites soient sur les coordonnées d'un rectangle « bounding box » saisies ou sur la base d'une géométrie télétransmise par l'utilisateur. De manière concomitante, les flux de données WMTS, WCS, les vignettes de prévisualisation et les archives servant au téléchargement de l'image sont produites par la séquence portant sur les fichiers d'images.

[72] Collection d'images : Une« collection d'images décrit un ensemble d'images, acquises au cours de la même mission, selon les paramètres propres à cette dernière (même caractéristiques optiques, même instrument, même plateforme) homogène en terme de niveau de traitement. Exemple : la collection d'images SPOT6 multi spectral ortho rectifiée acquise en 2017 sur le territoire métropolitain.

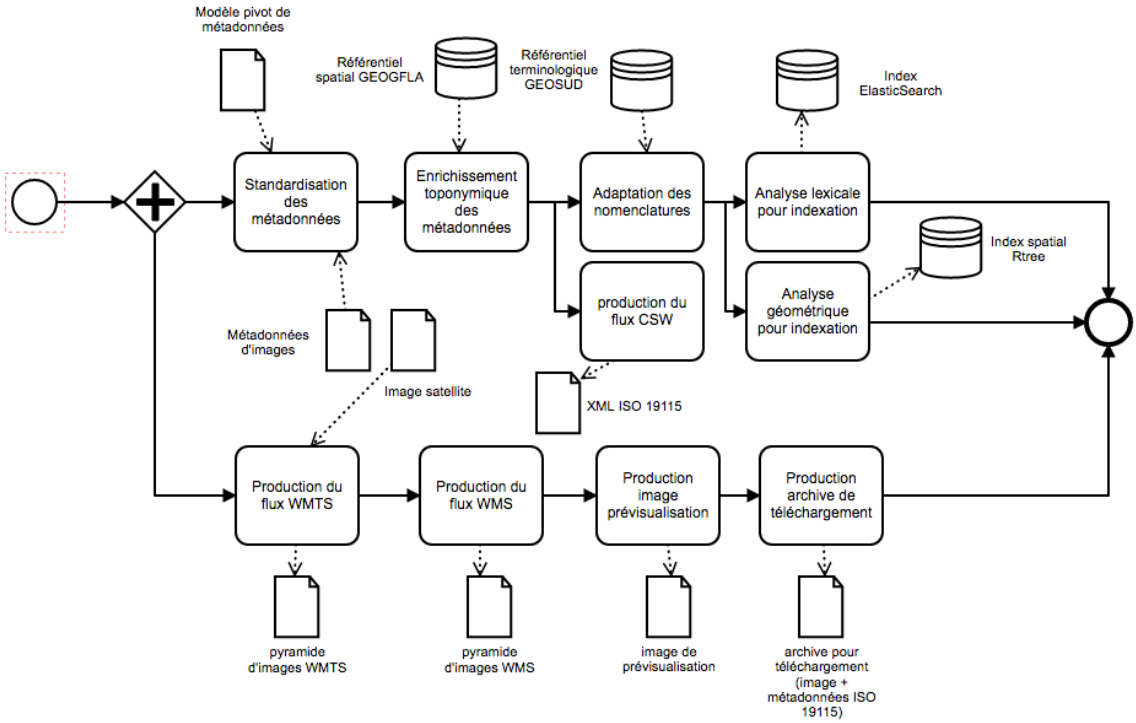


Figure 3.36 : Chaîne de traitements GEOSUD pour l'indexation guidée par les référentiels et la production de flux standardisés pour la diffusion des métadonnées et des images satellitaires (formalisme BPMN).

4.2.4 Architecture pour l'indexation

L'exécution automatique des opérations d'enrichissement et d'adaptation des métadonnées guidées par les référentiels est rendue possible par un ensemble de composants qui viennent étendre le processus d'indexation proposé par ElasticSearch. Pour chaque opération envisagée sur les métadonnées, un module Java est développé et est associé à l'élément de métadonnées sur lequel l'adaptation ou l'enrichissement doit être réalisé. Deux grands types de modules ont été développés (figure 3.37):

- *AdminBoundary handler* qui assure l'enrichissement des métadonnées à partir d'un référentiel spatial, ici le référentiel français GEOFLA,
- *SKOS related*, *SKOS hierarchy* et *SKOS property* qui permettent la manipulation des relations sémantiques et des propriétés SKOS pour traiter l'adaptation des terminologies.

Ces correspondances et la configuration de l'indexation sont décrites par un fichier de configuration indexProfiles.xml.

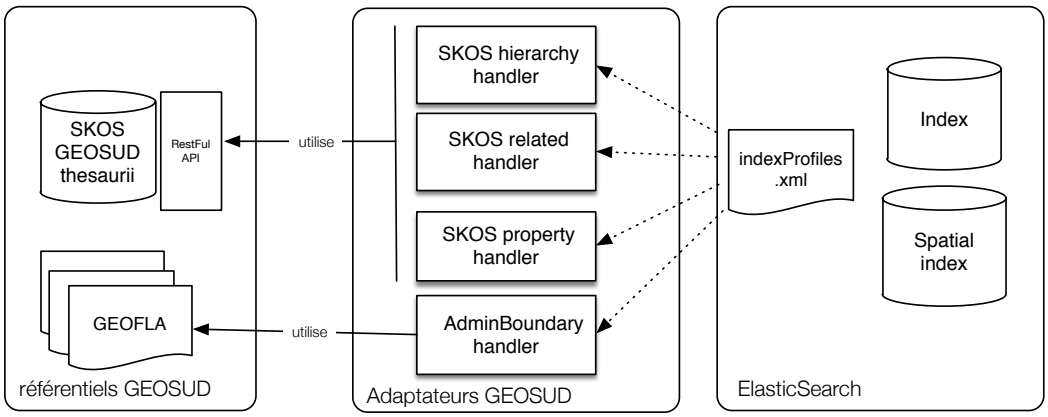


Figure 3.37: Schématisation des modules Java permettant d'étendre les fonctionnalités d'ElasticSearch pour assurer l'enrichissement et l'adaptation des métadonnées d'images.

4.2.4 Application de recherche sur les images satellitaires

L'application de recherche sur le catalogue d'images GEOSUD propose une découverte du catalogue en deux étapes :

Etape 1 (figure 3.38) : Une page d'accueil propose à l'utilisateur soit de basculer directement vers une navigation à facettes sur l'ensemble du catalogue, sans filtre préalable, soit de rechercher sur une des collections d'images proposées, ce qui revient à concentrer la recherche uniquement sur les images de cette collection.

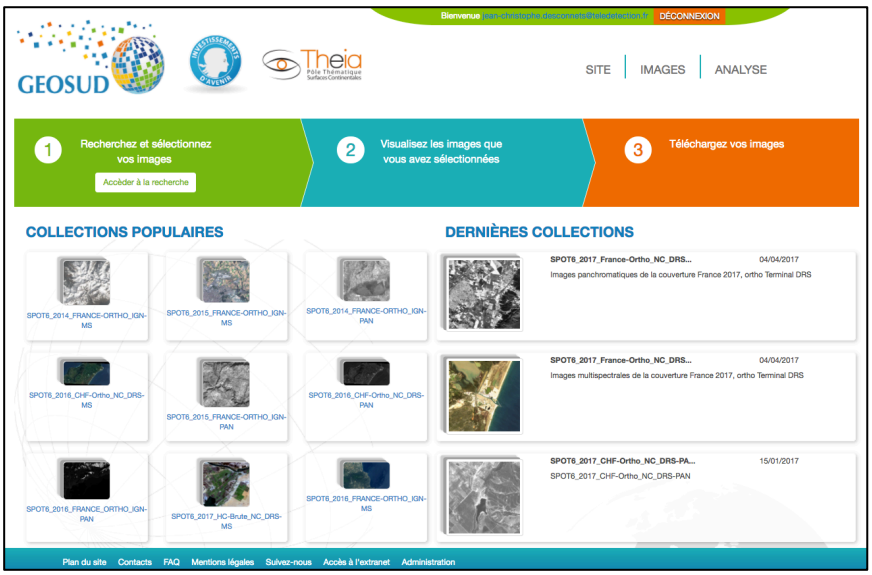


Figure 3.38: Page d'accueil de l'application de recherche de l'IDS GEOSUD

Etape 2 : quelque soit l'option choisie par l'utilisateur, la page où s'effectue la recherche et l'affichage des résultats sont présentés (figure 3.39). Les différentes facettes de recherche sont positionnées dans le volet vertical gauche en regard des résultats dans le volet droit. Les emprises spatiales des images retournées par la recherche sont affichées dans un encart cartographique (coin haut gauche). L'utilisation des facettes « Zone géographique », « Date » et « niveau de traitements »

permet de manière instantanée d'explorer puis de raffiner une recherche tout en visualisant les résultats correspondants. L'utilisateur pourra, dans un deuxième temps, choisir de visualiser (volet bleu turquoise) les images pour évaluer visuellement leur pertinence vis à vis du besoin puis les télécharger (volet orange). L'exemple de recherche fourni dans la figure 3.39 montre l'utilisation de la classification hiérarchique administrative. La facette hiérarchique permet de filtrer progressivement la recherche sur l'extension spatiale, de l'échelle de la région Aquitaine à l'échelle de la commune Lanton en passant par le département de la Gironde.

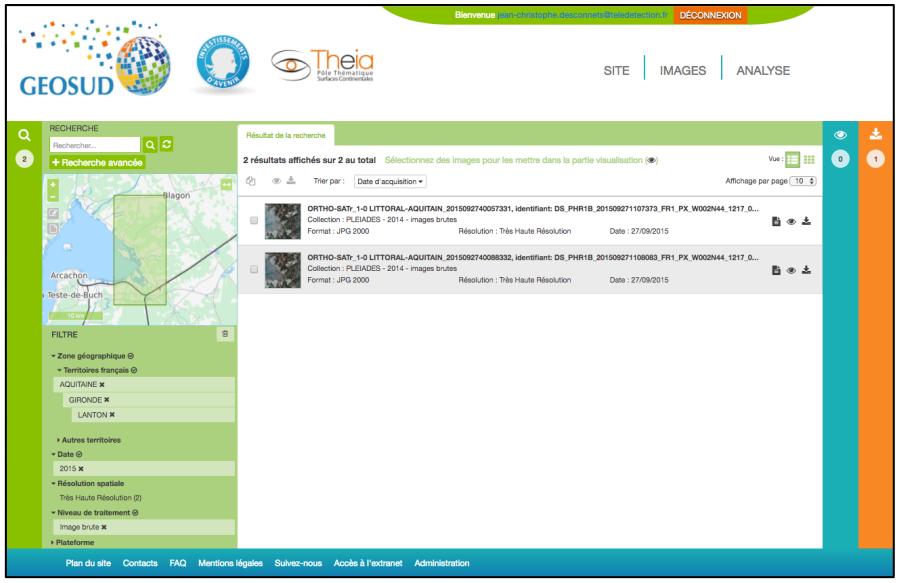


Figure 3.39 : Recherche d'images à partir des facettes « zone géographique », « Date et « niveau de traitements »

4.2.5 Quelques chiffres sur l'utilisation de l'application de recherche

Le déploiement de l'ensemble de la pile de composants ElasticSearch ELK (ElasticSearch, Logstach[73] et Kibana[74]) permet de récupérer et d'analyser les requêtes envoyées au moteur de recherche de notre plateforme ainsi que le type d'utilisateur qui les exécutent. L'analyse des fichiers de logs, sur la période janvier à avril 2017, montre que plus de 99% des requêtes exécutées s'appuient sur les facettes proposées. Plus particulièrement sont privilégiées les facettes portant sur la dimension spatiale et temporelle. Elles correspondent respectivement à 50% et 25% des requêtes (figure 3.40). Dans un moindre part, les facettes « niveau de traitements », « plateforme » et « résolution » sont utilisées, respectivement à 13, 6 et 5% des requêtes analysées. Enfin, l'analyse des profils utilisateurs nous permet de constater que les recherches effectuées sur notre plateforme le sont à 73% par des utilisateurs travaillant pour un établissement de recherche et d'enseignement (EPST). Les utilisateurs issus des services de l'état et les collectivités territoriales constituent le reste du contingent (figure 3.41). L'analyse plus fine de la structure d'appartenance des utilisateurs montre qu'une bonne majorité est issue de services ou de laboratoires dont la télédétection n'est pas le domaine d'expertise.

[73] Logstach est un ETL open source faisant partie de la suite ElasticSearch: www.elastic.co/fr/products/logstash

[74] Kibana est un outil de consultation et d'analyse de données open source faisant partie de la suite ElasticSearch : www.elastic.co/fr/products/kibana

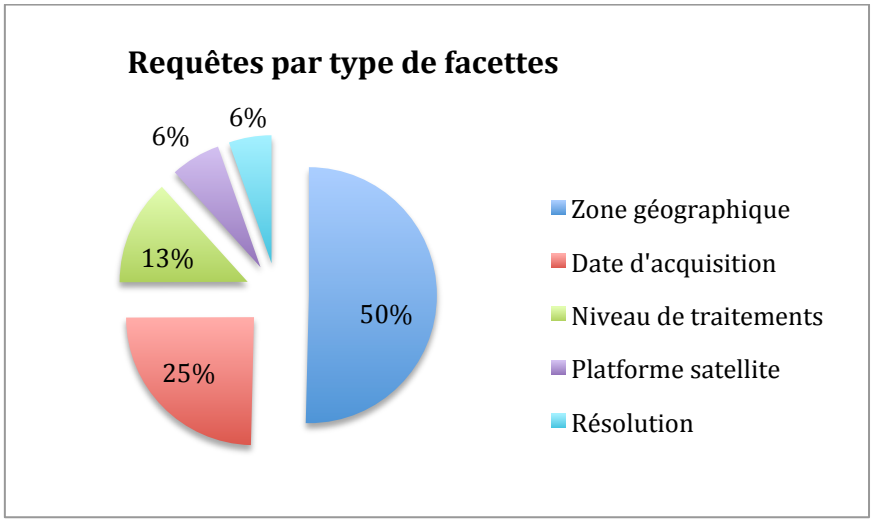


Figure 3.40: Répartition des requêtes selon le type de facettes. Période : janvier à avril 2017

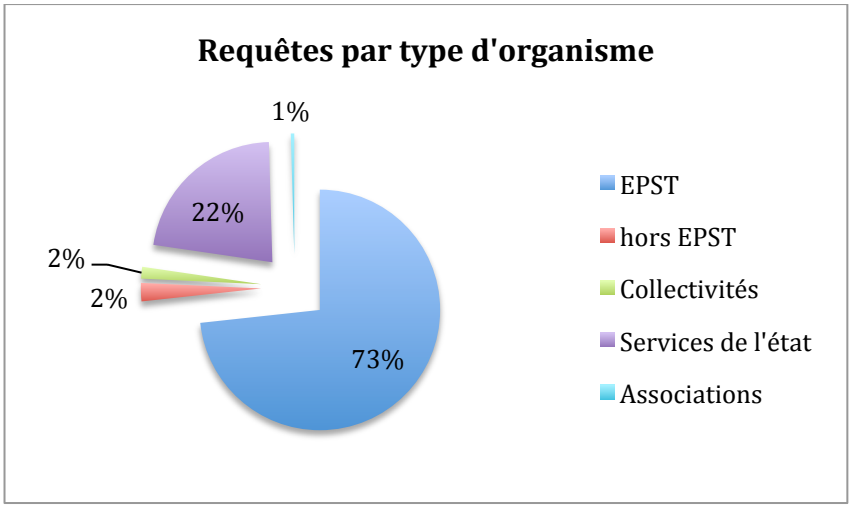


Figure 3.41: répartition des requêtes selon le type de facettes et le type d'organisme d'origine de l'utilisateur. Période : janvier à avril 2017

5. Conclusion

Synthèse

Les propositions déclinées dans ce mémoire sont le résultat des travaux menés autour du partage de données spatio-temporelles durant plusieurs années. Elles n'en sont pas la synthèse mais leur assemblage. Elles constituent à ce titre des propositions originales qui pourront être valorisées, dans un proche avenir, sous forme de publications scientifiques. En effet, elles mobilisent les trois domaines autour desquels nous avons investi: la médiation de données, la recherche d'information, et les représentations issues du web sémantique. Avec les connaissances et l'expertise accumulées, il s'agissait de mieux répondre aux attentes et aux enjeux de découverte et d'accès aux données spatio-temporelles tels qu'ils sont posés aujourd'hui face aux masses de données toujours plus importantes et de nature variées qui sont produites.

Aussi, nos réflexions nous ont amené à considérer et à positionner de manière centrale le besoin de l'utilisateur dans ses activités de recherche d'information. Ces activités sont le préalable à l'analyse puis à la production de nouvelles connaissances. Orienté vers l'utilisateur, l'objectif des propositions est de rendre les processus de découverte et de sélection des données, plus particulièrement ceux des images satellitaires, sémantiquement et ergonomiquement « accessibles » aux utilisateurs potentiels. Cet objectif est doublé d'une volonté de proposer une méthodologie qui puisse être mise en œuvre dans d'autres contextes opérationnels de découverte de grands catalogues d'images satellitaires multi capteurs.

Pour cela, les propositions développées dans ce mémoire s'attachent à décliner et à formaliser les différents éléments pour concevoir un système de recherche à facettes. Elles s'attachent également à adapter et généraliser la recherche à facettes dans le contexte particulier des données spatio-temporelles. L'adaptation d'un tel système de recherche nous a amené à articuler les propositions autour de 4 principaux éléments : une classification à facettes adaptée aux images satellitaires qui permette de formaliser les différentes dimensions avec lesquelles il est pertinent de rechercher ces données ; la définition d'un profil d'application qui constitue le socle pour interconnecter et harmoniser les métadonnées d'images issues de différentes plateformes ; une indexation des images guidée par les thésaurus pour réduire les ambiguïtés sémantiques, enrichir et adapter la terminologie à celle de l'utilisateur et enfin les lignes directrices pour concevoir les IHM relatives à une application mettant en œuvre une recherche à facettes.

Réutilisation de l'approche

La mise en œuvre de tout ou partie de ces différents éléments dans plusieurs projets a montré leur faisabilité, leur pertinence et leur robustesse. L'exemple de l'infrastructure de données spatiales GEOSUD (Kazmierski e *al*., 14 ; Maurel et *al*., 16) nous a permis de le démontrer. L'implémentation du portail d'accès aux ressources de renforcement de capacités GEOCAB en est une autre déclinaison (Desconnets et *al*., 17a). Les outils et les architectures utilisés permettent d'envisager une réutilisation de nos propositions vers d'autres grands ensembles de données spatio-temporelles. En premier lieu, le modèle *Observation & Measurements* utilisé et décliné pour les images optiques peut être vu comme une ontologie cadre ou un méta modèle, selon le point de vue adopté, sur lequel pourra être dérivé la classification de jeux de données issus des sciences environnementales. L'enjeu à court terme est de pouvoir concevoir des systèmes de recherches qui puissent offrir la découverte à de très grands ensembles de données à la fois issues de l'observation de la Terre, des observations *in situ* ou de leur produits dérivés.

La mise en place de l'infrastructure de recherche virtuelle nationale « Pôle de données Système Terre » dans lequel nous sommes impliqués est un terrain sur lequel nous allons pouvoir évaluer l'intérêt et la généricité de l'approche. Dans ce contexte, l'enjeu est de pouvoir partager des données issues d'observation dans des domaines aussi variés que la géologie, l'hydrologie, géologie, la télédétection, la géophysique… Un composant de fédération des données issu des différents pôles reposera sur la vision fédératrice que propose O & M. Nos travaux ont également été repris pour concevoir la fédération de catalogues d'images de l'initiative Dinamis[75]. Cette initiative française a pour objectif de mutualiser l'approvisionnement, l'achat, la gestion des demandes et la distribution d'un panel d'images satellitaires HR et THR auprès des acteurs publics. Enfin, les travaux entrepris pour implémenter une plateforme web GEOSUD d'analyse des images satellitaires HR et THR nous a

[75] *Dinamis : Dispositif Institutionnel Approvisionnement Mutualisé Images Spatiales*

permis de transposer notre approche sur la découverte au sein de catalogues de traitements. Dans ce contexte, la description des traitements repose sur l'utilisation de la spécification de l'OGC WPS[76]. Le schéma de données permettant de structurer la description d'un traitement (opération *DescribeProcess*) est enrichi d'annotations issues de classifications ou de liste de valeurs (catégorie de traitements, maturité des implémentations,..). Les dimensions de recherche proposées aux utilisateurs s'appuient sur ces annotations.

Limites

A plusieurs reprises, nous avons souligné les limites de nos propositions. Elles sont principalement d'ordre informationnel et conceptuel. Si notre approche tente de rapprocher la description des images vers la sphère des utilisateurs en proposant des adaptations et des enrichissements de valeurs de métadonnées, ces améliorations portent uniquement sur des caractéristiques d'images ou le contexte de son acquisition. La mise à disposition d'annotations, et donc d'une dimension de recherche, sur le contenu des images n'est pas traité car il est délicat d'assurer aujourd'hui, de manière automatisée et opérationnelle, de tel traitement sur des catalogues d'images HR et THR. Cela constitue un sérieux frein, et un enjeu majeur, pour qu'un utilisateur puisse aborder, de manière plus directe et en utilisant ses connaissances « métier », la sélection d'images en relation avec ses besoins applicatifs.

D'un point de vue conceptuel, notre approche souffre du manque d'expressivité des représentations choisies pour formaliser et structurer les connaissances sur lesquelles nous nous appuyons pour concrétiser les dimensions de recherche (les facettes), assurer les adaptations et les enrichissements des valeurs de métadonnées. En effet, dans une première approche, nous avons choisi d'utiliser la dimension terminologique des RTO (Resource Termino-Ontologique). Si ce choix relève d'un certain pragmatisme quant à la facilité de mise en place, de gestion et de maintenance, l'expressivité est relativement limitée. D'autre part, l'amalgame qui fait entre les dimensions ontologique et terminologique ne permet pas d'envisager son extension à une exploitation de sa dimension ontologique dans laquelle les besoins applicatifs de la communauté environnementale seraient mis en relation avec les caractéristiques des images. Par ailleurs, nous avons souligné tout l'intérêt que revêt l'utilisation de la dimension ontologique des référentiels spatiaux ou terminologiques, et notamment la représentation des relations topologiques entre entités spatiales, pour apporter de nouvelles fonctionnalités d'exploration des catalogues d'images.

Perspectives

Ces dernières années, la mise en place d'infrastructures de données spatiales ont permis de répandre les pratiques de standardisation des métadonnées et la mise en place de service découverte. Même si ces pratiques demandent encore à être améliorées et ouvertes à de plus larges besoins fonctionnels, la standardisation a grandement participer à favoriser le partage et l'accès aux données d'observation de la Terre au delà de la sphère des télédétecteurs.

Aujourd'hui, face à la quantité et à la diversité toujours plus grandes des données mise en ligne pour répondre aux attentes en matière de suivi et de gestion des territoires, les réponses apportées à l'exploration et à la sélection des images pertinentes doivent aller au delà de ces pratiques. La standardisation, l'adaptation ou l'enrichissement des valeurs de métadonnées que nous avons proposées est un premier pas. Mais il n'est pas suffisant. En premier lieu, les systèmes de recherche

[76] *WPS : Web Processing Service*

d'information doivent mieux intégrer la connaissance « métier » des utilisateurs afin de fournir des capacités d'exploration et de sélection des images orientées « application» et non pas orientée « données » comme c'est encore le cas aujourd'hui. D'autre part, nous pouvons attendre des applications de recherche qu'elles offrent davantage de moyens de représentation visuelle des données spatio-temporelles, cela pour améliorer l'exploration de grands ensembles d'images en utilisant les capacités cognitives de l'utilisateur. Les perspectives que nous pouvons dégager aujourd'hui pour orienter nos travaux dans ce sens empruntent trois directions.

La première vise à poser une réflexion ontologique sur les référentiels de valeurs « producteur », « utilisateur » et spatiaux utilisés afin de mieux exploiter la connaissance. Il s'agit de pouvoir contextualiser les facettes de recherche au cours d'un processus de filtrage. L'objectif serait d'utiliser les capacités de raisonnement associé à une représentation ontologique d'une classification à facettes pour proposer de nouvelles dimensions de recherche qui seront adaptées aux résultats des filtrages précédents. Pourront notamment être mis à contribution les travaux de (Compton et *al*., 12) ou (Islam et *al*., 04) qui ont représenté les standards *O & M* et ISO 19115 au format OWL. En ce qui concerne, la dimension spatiale de la recherche, les relations topologiques associées aux hiérarchies agrégatives des territoires doivent pouvoir être exploitées pour apporter un nouveau moyen de naviguer dans la dimension spatiale des images notamment au travers des relations de voisinage, d'inclusion ou d'appartenance entre unités administratives. Des ontologies comme celle de Geonames les décrivent.

La deuxième vise à combler l'absence d'information sur le contenu des images dans les métadonnées proposées par les producteurs d'images. Les dernières avancées en matière d'extraction de connaissances dans les images satellitaires (Chahdi et *al*., 16) nous paraissent exploitables pour venir compléter notre approche d'enrichissement des métadonnées. Les pré- classifications des pixels d'images qui en résultent, alignées à des classifications de référence d'une communauté de pratiques, pourront venir enrichir les métadonnées d'images et fournir ainsi les informations de contenu manquantes. Une nouvelle dimension de recherche, portant sur les objets d'intérêt thématiques, serait alors disponible.

De manière complémentaire à l'enrichissement des métadonnées par les pré-classifications de pixels, l'évaluation de la compatibilité d'une image à une application donnée serait également pertinente. Par compatibilité, nous entendons que les caractéristiques de l'image soient syntaxiquement et sémantiquement conformes aux paramètres d'entrée d'une chaîne de traitements et que les objets sur lequel porte le traitement soient présents. On parlera de compatibilité thématique. Traiter la compatibilité des images à une application revient à croiser les caractéristiques d'images à ceux des traitements qui sont mis à disposition. Les travaux de (Lin et *al*., 10 ; Lin, 11) nous ont fourni les pistes pour aborder au moins les compatibilités syntaxiques et sémantiques entre données et traitements. Proposer l'évaluation de la compatibilité thématique d'une image à un traitement demandera alors de formaliser la connaissance « métier » domaine d'application par domaine d'application et pouvoir le mettre en relation avec le contenu des images.

6. Bibliographie

Adkisson, H.P. (2005). Web design practices: use of faceted classification. www. webdesignpractices.com/navigation/facets.html, accédé en avril 2017.

AFNOR, (1981). Z 47-100: 1981. Règles d'établissement des thésaurus monolingues. Paris, AFNOR.

Aitchison, J., Gomershall, A. et Ireland, R. (1969). Thesaurofacet: A Thesaurus and Faceted Classification for Engineering and Related Subjects, English Electric, Whetstone.

Ahlberg, C. et Shneiderman, B. (1994). Visual information seeking: tight coupling of dynamic query filters with starfield displays. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating Interdependence (Boston, MA, USA, April 24–28, 1994). B. Adelson, S. Dumais, and J. Olson (Eds.), CHI '94. pp. 313–317. NewYork: ACM.*

Arvor, D., Durieux, L., Andrés, S., et Laporte, M. A. (2013). Advances in geographic object-based image analysis with ontologies: A review of main contributions and limitations from a remote sensing perspective. *ISPRS Journal of Photogrammetry and Remote Sensing, 82, 125-137.*

Austin, D. (1984), PRECIS: A Manual of Concept Analysis and Subject Indexing, 2nd ed.. The British Library Bibliographic Services Division, London.

Baghdadi N., Zribi M. (2017). Observation des Surfaces continentales par télédétection optique. Techniques et méthodes. Collection Système Terre – Environnement. Série Télédétection pour l'observation des surfaces continentales. ISTE Edition. Vol. 1. 366p.

Barde, J., Libourel, T., et Maurel, P. (2004). Ontologies et métadonnées pour le partage d'information géographique. *Revue internationale de géomatique, 14(2), 199-216.*

Barragáns-Martínez, A. B., Rey-López, M., Costa-Montenegro, E., Mikic-Fonte, F. A., Burguillo, J. C., et Peleteiro, A. (2010). *Exploitation of social tagging for outperforming traditional recommender system techniques. IEEE Internet Computing, 14(6), 23–30.*

Bayoudh, M., Roux, E., Richard, R. et Nock, R. (2015). Structural knowledge learning from maps for supervised land cover/use classification: Application to the monitoring of land cover/use maps in French Guiana. *Computers and Geosciences, Elsevier, 76, pp.31-40.*

Baziz M., Aussenac-Gilles N. et Boughanem M. (2003, Décembre). Désambiguisation et Expansion de Requêtes dans un SRI, Etude de l'apport des liens sémantiques. *Dans : Revue des Sciences et Technologies de l'Information (RSTI) série ISI, Hermes, 11, rue Lavoisier, F-75008 Paris, V. 8, N. 4/2003, p. 113-136.*

Baziz M., Boughanem M., Pasi G. et Prade H. (2007). An information retrieval driven by ontology from query to document expansion. *In : Proceedings of the Large Scale Semantic Access to Content (Text, Image, Video, and Sound) RIAO'2007, Pittsburgh,30-01 June, USA, CID : Paris, France, p. 301–313.*

Boisson, P., Clerc, S., Desconnets, J.C., et Libourel, T. (2006, Octobre). Using a semantic approach for a Cataloguing Service. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 1712-1722). Springer Berlin Heidelberg.

Beghtol C., (2008). From the universe of knowledge to the universe of concepts : The structural revolution. *In classification for information retrieval, Axiomathes, vol. 18, pp. 131-144.*

Belkin, N.J., Ingwersen, P., Pejtersen, A.M. (1992). *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Copenhagen, Denmark, June 21-24, 1992 ACM*

Boubekeur, F. (2008). Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets. Thèse de Doctorat, Université Paul Sabatier-Toulouse III.

Broughton, V. (2000). A new classification for religion. *International Cataloging and Bibliographic Control, No. 4, pp. 2-4.*

Broughton, V. (2006, Janvier). The need for a faceted classification as the basis of all methods of information retrieval. *In Aslib proceedings (Vol. 58, No. 1/2, pp. 49-72). Emerald Group Publishing Limited.*

Buckley C., Salton G. et Allan J. (1994). The Effect of adding information in a relevance Feedback environment, in the Proceedings of the ACM SIGIR Conference On Research and Development in Information Retrieval (SIGIR), pp 292-300.

Card, S. K., Mackinlay, J. D., et Shneiderman, B. (1999). Readings in information visualization: using vision to think. Morgan Kaufmann.

Carlisle M., Green D.R., De Martino M., Albertino R ., Desconnets J.C**.**, Waver R. et Cabello M. (Janvier 2010): INSPIRE and Nature-SDIplus: further progress towards a spatial data infrastructure (SDI) for nature conservation in the EU. *Proceedings of the seventeenth annual IALE, UK Conference.*

Chahdi, H. (2013). Conception d'un modèle abstrait de métadonnées pour l'interconnexion de jeux de données géoréférencées. Mémoire de Master Informatique, spécialité DECOLL. Université de Montpellier 2.

Chahdi, H., Grozavu, N., Mougenot, I., Berti-Equille, L., et Bennani, Y. (2016, Janvier). Génération de contraintes pour le clustering à partir d'une ontologie-Application à la classification d'images satellites. *In Extraction et Gestion des Connaissances (EGC).*

Chiaramella, Y. et Mulhem, P.. (2007). La recherche d'information. De la documentation automatique à la recherche d'information en contexte. *Recherche d'information, Document numérique 2007/1 (Vol. 10), p. 11-38.* http://www.cairn.info/revue-document-numerique-2007-1-page-11.htm, accédé le 12 mai 2017

Chichereau, D., Contat, O., Dégez, D. Deniau, A., Lénart, M. Masse, C. et Dominique Ménillet, D. (2007) Les normes de conception, gestion et maintenance de thésaurus. *Documentaliste-Sciences de l'Information, 44(1) :66–74.*

Christian, E.J. (2008). GEOSS Architecture Principles and the GEOSS Clearinghouse. *Systems Journal, IEEE , vol.2, n°.3, pp.333,337.*

Cleverdon C. W., Mills J., Keen E. M. (1966). Factors Determining the Performance of Indexing Systems, *ASLIB Cranfield Research Project.*

Compton, M., Barnaghi, P., Bermudez, L., GarcíA-Castro, R., Corcho, O., Cox, S., et Huang, V. (2012). The SSN ontology of the W3C semantic sensor network incubator group. *Web semantics: science, services and agents on the World Wide Web*, *17*, 25-32.

Cossu, R., Pacini, F., Brito, F., Fusco, L., Li Santi, L., et Parrini, A.. (2010). GENESI-DEC: a federative e-infrastructure for Earth Science data discovery, access, and on-demand processing. *In 24th International Conference on Informatics for Environmental Protection*.

Cubranic, D. (2008). Polestar: assisted navigation for exploring multi-dimensional information spaces. *Workshop on Human-Computer Interaction in Information Retrieval (HCIR'08)*.

Deerwester, S., Dumais, S., Furnas, G. W., Landauer, T. K., et Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science 41(6)*

Denton, W. (2003, Octobre). Putting facets on the web: an annotated bibliography. www.miskatonic.org/library/facet-biblio.html, accédé Avril 2017.

Desconnets, J.C., Libourel T., Maurel, P., Miralles, A., et Passouant, M. (2001, Septembre). Proposition de structuration des métadonnées en géosciences: Spécificité de la communauté scientifique. *In Journées Cassini'2001: Géomatique et espace rural (pp. 69-82)*.

Desconnets, J.C., Moyroud, N., et Libourel, T. (2003, Juin). Méthodologie de mise en place d'observatoires virtuels via les métadonnées. *In INFORSID (pp. 253-267)*.

Desconnets, J.C., Libourel T., Clerc, S., et Granouillac, B. (2007, Mai). Cataloguing for distribution of environmental resources. *In AGILE'07: 10th International Conference on Geographic Information Science (p. 15). Aalborg University*.

Desconnets J.C., Heurteaux V. (2008, Septembre): MDweb 2.0 - A Java/JEE Metadata Catalog. Free Open Source Software For Geography (FOSS4G). Cape Town, South Africa.

Desconnets J.C., Libourel T., Heurteaux V. (2009, Juillet) : Partage et mutualisation en environnement : des concepts à l'usage. L'expérience MDweb. OGRS 2009, International Opensource Geospatial Research Symposium, Nantes. 8-10 juillet 2009.

Desconnets J.C., Libourel T., Desruisseaux M. (2010, Mai) : Métamodélisation pour les services de catalogage et de localisation de ressources environnementales. Atelier SIDE, XXVIIIèeme Conférence INFORSID. 25-28 Mai, 2010. Marseille.

Desconnets, J.C. et Libourel, T. (2012) MDweb : Catalogage et localisation de ressources environnementales. *Dans Développements logiciels en géomatique: innovations et mutualisation (eds : B. Bucher and F. Le Ber). Traité IGAT. Hermès-Lavoisier. Paris, France*.

Desconnets, J. C., Chahdi, H., et Mougenot, I. (2014, Novembre). Application profile for Earth Observation images. In Research Conference on Metadata and Semantics Research (pp. 68-82). Springer International Publishing.

Desconnets, J.C., et Kazmierski, M. (2015). Mutualiser des données spatiales et des traitements en environnement. *Ingénierie des Systèmes d'Information, 20(3), 89-115*.

Desconnets J.C., Giuliani G., Guigoz Y., Lacroix P., Mlisa A., Noort M., Ray N., Searby N.D. (2017a): GEOCAB Portal: a gateway for discovering and accessing capacity building resources *in Earth Observation*. *International Journal of Applied Earth Observation and GeoInformation*. *Volume 54, February 2017, Pages 95–104*.

Desconnets J.C., Mougenot I. et Chahdi H. (2017b) : A methodology for effective Metadata Design in Earth observation. In *Developing Metadata Application Profiles* (pp. 65-97). IGI Global.

Desfriches Doria, O. (2012). Contribution de la classification à facettes pour l'organisation des connaissances dans les organisations. *Études de communication. lLangages, information, médiations, (39), 173-200*.

Dupuy, S., Barbe, E., et Balestrat, M., (2012). An object-based image analysis method for monitoring land conversion by artificial sprawl use of RapidEye and IRS data. *Remote Sensing 4.2, 2012, 404-423*.

El Hassouni, N. (2014). Environnement web pour l'exploitation des métadonnées RDF. Application aux données issues de l'observation de la terre. Mémoire de Master Informatique. Université de Montpellier.

Ellis, D. et Vasconcelos, A. (2000). The relevance of facet analysis for world wide web subject organization and searching. *Journal of Internet Cataloging, Vol. 2 Nos 3/4, pp. 97-114*.

ESA, (2015): Sentinel-2 Products Specification Document. https://sentinel.esa.int/documents/247904/349490/S2_MSI_Product_Specification.pdf, accédé le11 avril 2017.

European Commission Joint Research Centre (2013). INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119. inspire.ec.europa.eu/documents/Metadata/MD_IR_and_ISO_20131029.pdf Accédé le 12 avril 2017.

Faure, J.F. et Caminade, J.P. (2015, Décembre). SEAS Haïti. Phase 1, Lot 2 – Etude de dimensionnent. Etape 3 : Panorama des missions existantes. Rapport d'étape. UMR ESPACE-DEV, Université d'état d'Haïti.

FGDC, 1998: Content standard for digital geospatial metadata. FGDC-STD-001-1998.

Fowler, M. (1988). Analysis Patterns: reusable object models. Addison Wesley Longman, Menlo Park, CA.

Furnas, G., Landauer, T., Gomez, L., et Dumais, S. (1987). The Vocabulary Problem in Human-System Communication. *Communications of the ACM 30(11): pp. 964–971. doi:10.1145/32206.32212*

Gaspéri, J., Houbie F., Woolf A. et Smolders, S. (2012). Earth observation metadata profile of Observations & Measurements. OGC Document Number 10-157r3. Accés avril 2017 de https://portal.opengeospatial.org/files/?artifact_id=47040

Gayte, O., Libourel, T., Cheylan, J.P. et Lardon, S. (1997). Conception des systèmes d'information sur l'environnement. Géomatique. Paris, France. Edition Hermès, 153 p.

Georis-Creuseveau J. (2014). Les Infrastructures de Données Géographiques (IDG): développement d'une méthodologie pour l'étude des usages. Le cas des acteurs côtiers et de la GIZC en France. Mémoire de doctorat. En Géographie. Université de Bretagne Occidentale. 85p.

Gui, Z., Yang, C., Xia, J., Liu, K., Xu, C., Li, J., et Lostritto, P. (2013). A performance, semantic and service quality-enhanced distributed search engine for improving geospatial resource discovery. *International Journal of Geographical Information Science, 27(6), 1109-1132.*

Guttman, A. (1984). R-trees: a dynamic index structure for spatial searching. *ACM Vol. 14, No. 2, pp. 47-57.*

Hajalalaina, A.R., Grizonnet M., Delaître E., Rakotondraompiana, S. et Dominique Hervé. (2013). Discrimination des zones humides en forêt malgache, proposition d'une méthodologie multirésolution et multisource utilisant ORFEO toolbox. *Revue Française de Photogrammétrie et de Télédétection, 2013, pp. 37-48.*

Harman, D. (1992): Relevance Feedback Revisited. *In the Proceedings of the ACM SIGIR Conference On Research and Development in Information Retrieval (SIGIR), pp1-10.*

Haslhofer B. et Klas W. (2010): A survey of techniques for achieving metadata interoperability, *ACM Comput. Surv., vol. 42, n°2.*

Heery, R. et Patel, M. (2000). Application profiles: mixing and matching metadata schemas. Ariadne, 25

Heurteaux V. (2016). Création d'un plugin Elasticsearch chez Geomatys. https://www.elastic.co/fr/blog/creation-d-un-plugin-elasticsearch-chez-geomatys, accédé le 22 Avril 2017.

Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., et Yee, K. P. (2002). Finding the flow in web site search. *Communications of the ACM, 45(9), 42-49.*

Hearst, M. A. (2006). Clustering versus faceted categories for information exploration. *Communications of the ACM, 49(4), 59-61.*

Hillman, D.I., Phipps, J. et Coyle, K. (2010). Introduction to Application Profiles.

Hostetter, C. (2006). Faceted searching with apache SolR. *ApacheCon US.*

Hudon M. et Mustafa El Hadi W., (2010). Organisation des connaissances et des ressources documentaires. De l'organisation hiérarchique centralisée à l'organisation sociale distribuée. *Les Cahiers du Numérique, vol. 6, n° 3, pp. 9-38.*

Ingwersen, P. et Wormell, I. (1992). Ranganathan in the perspective of advanced information retrieval.*Libri, Vol. 42, pp. 184-201.*

Ingwersen, P. (1999). Cognitive Information Retrieval. *Annual Review of Information Science & Technology, vol. 34, p. 3-52.*

Isaac, A., Phipps, J., et Rubin, D. (2007, Mai). SKOS use cases and requirements. W3C working draft, W3C.

Islam, A. S., Bermudez, L., Fellah, S., Beran, B. et Piasecki, M. (2004). Implementation of the Geographic Information-Metadata (ISO 19115: 2003) Norm using the Web Ontology Language (OWL). *Transactions in GIS*.

ISO (1986). Principes directeurs pour l'établissement et le développement de thésaurus monolingues. 2788:1986. International Organization for Standardization, Geneva, Switzerland.

ISO (2002a). Geographic Information General Reference Model, ISO19101:2002, International Organization for Standardization, Geneva, Switzerland.

ISO (2002b). Geographic Information temporal schema, ISO 19108:2002, International Organization for Standardization, Geneva, Switzerland.

ISO (2003). Geographic Information Metadata, ISO 19115:2003, International Organization for Standardization, Geneva, Switzerland.

ISO (2004). Geographic Information Profiles, ISO 19106:2004, International Organization for Standardization, Geneva, Switzerland.

ISO (2005). Geographic Information Rules for application schema, ISO 19109:2005, International Organization for Standardization, Geneva, Switzerland.

ISO (2009). ISO19115-2 :2009. Geographic information — Metadata — Part 2: Extensions for imagery and gridded data. International Organization for Standardization, Geneva, Switzerland.

Jacko, J. A., et Sears, A. (2003). Handbook of Research on Ubiquitous Computing Technology for Real Time Enterprises.

Kazmierski, M., Desconnets J.C, Guerrero B. et Briand D. (2014, Juin). GEOSUD SDI, Accessing Earth Observation data collections with semantic-based services. In *Proceedings of the 17th AGILE Conference on Geographic Information Science, Connecting a Digital Europe through Location and Place,* Castellon, Spain.

Keim, D., Mansmann, F., Schneidewind, J., Thomas, J., et Ziegler, H. (2008). Visual analytics: Scope and challenges. *Visual data mining*, 76-90

Koren, J., Leung, A., Zhang, Y., Maltzahn, C., Ames, S., et Miller, E. (2007). Searching and navigating petabyte-scale file systems based on facets. *Proceedings of the 2nd international Workshop on Petascale Data Storage (PDSW '07): pp. 21–25. doi:10.1145/1374596.1374603*

Kuć, R., et Rogoziński, M. (2013). *Mastering ElasticSearch*. Packt Publishing Ltd.

Kumar, K. (1974). Theory of classification. 14th rev. New Delhi: Vikas Publishing House Pvt Ltd.

LaBarre, K. (2004), "Adventures in faceted classification: a brave new world or a world of confusion?", *in McIlwaine, I.C. (Ed.), Knowledge Organization and the Global Information*

Society, Proceedings of the 8th International Conference of the International Society for Knowledge Organization, University College London, 13-16 July, 2004, Advances in Knowledge Organization, Vol. 9, Ergon, Wu¨rzburg, pp. 79-84.

La Barre K., (2010), A Semantic (faceted) Web *? Les cahiers du numérique, vol. 6, n° 3, pp. 103-131.*

Lamb, J. (2001). Sharing best methods and know-how for improving generation and use of metadata. *New Techniques and Technologies for Statistics and Exchange of Technology and Know-how, 175-194.*

Leroy M., Kosuth P., Hagolle O., Cherchali S., Maurel P. et Desconnets J.C. (2013, Septembre). Theia Land data center. *ESA Living Planet Symposium. Edimburgh, UK. 9-13 Septembre 2013.*

Livre Blanc (1998). L'information géographique française dans la société de l'information. Rapport CNIG/AFIGEO.

Lin, Y., Pierkot, C., Mougenot, I., Desconnets, J.C., et Libourel, T. (2010, Juin). A framework to assist environmental information processing. In International Conference on Enterprise Information Systems (pp. 76-89). Springer Berlin Heidelberg.

Lin Y. (2011, Décembre): Méthodologie et composants pour la mise en oeuvre de workflows scientifiques. Thèse de Doctorat. Université de Montpellier II. Ecole doctorale I2S.

Loukili, A. (2014). Développement d'une application web de recherche à facettes sur les métadonnées au format RDF. Application au domaine de l'Observation de la Terre . Mémoire de Master Informatique. Université de Montpellier.

Lutz, M., et Klien, E. (2006). Ontology ‐ based retrieval of geographic information. *International Journal of Geographical Information Science, 20(3), 233-260.*

Maniez, J. (1999). Des classifications aux thésaurus: du bon usage des facettes. *Documentaliste, 36(4-5), 249-260*

Manning, C., Raghavan, P., et Schu¨tze, H. (2008). Introduction to Information Retrieval. New York: Cambridge University Press.

Maciejewski, R., Rudolph, S., Hafen, R., Abusalah, A., Yakout, M., Ouzzani, M. et Ebert, D. S. (2010). A visual analytics approach to understanding spatiotemporal hotspots. *IEEE Transactions on Visualization and Computer Graphics*, *16*(2), 205-220

Maurel P., Faure J.F.,· Cantou J.P., Desconnets J.C., Teisseire M., Mougenot I., Martignac C. et Bappel E. (2015, Septembre) : The GEOSUD remote sensing data and services infrastructure. *ISPRS Conference - RSDI (Remote Sensing Data Infrastructure) Workshop. 7-14 septembre 2015. La Grande Motte, France.*

Mills, J. et Broughton, V. (1977), Bliss Bibliographic Classification, 2nd ed.. Butterworth, London.

Mougenot, I., Desconnets, J.C., et Chahdi, H. (2015, Septembre). A DCAP to promote easy-to-use data for multiresolution and multitemporal satellite imagery analysis. *In Proceedings of the*

2015 International Conference on Dublin Core and Metadata Applications (pp. 10-19). Dublin Core Metadata Initiative.

Mougenot, I. (Décembre 2015). Standards de métadonnées et définition profils de d'application Dublin Core. Mémoire pour l'habilitation à diriger des recherches. Ecole Doctorale I2S, Université de Montpellier.

Moldovan D. et Mihalcea, R. (2000). Using WordNet and lexical operators to improve Internet searches. *IEEE Internet Computing, 4(1) :34- 43.*

Nativi S. et Bigagli L. (2009) Discovery, Mediation, and Access Services for Earth Observation Data Selected Topics *in Applied Earth Observations and Remote Sensing, IEEE Journal of , vol. 2, n°4, p.233-240*

National Research Council (1995). Expanding the Vision of Sensor Materials. Committee on New Sensor Technologies: Materials and Applications. National Academy Press. http://books.nap.edu/books/0309051754/html/index.html. Accédé le 9 Mai 2017

Nieva, T. (2001) : Remote data acquisition of embedded systems using internet technologies: a role-based generic system specification. Thesis, Ecole Polytech. Fed. Lausanne 2001. https://infoscience.epfl.ch/record/32864/files/EPFL_TH2388.pdf , accédé le 9 Mai 2017.

Neal, M. (1997). Parametric search: Evolving information retrieval for the Web. *CADIS Inc, 1501(2), 5.*

Nilsson, M. (2008). The Singapore framework for Dublin Core application profiles. http://dublincore.org/documents/singapore-framework/, accede en mai 2017

Nilsson M., Baker T. et Johnston P., (2008). The Singapore Framework for Dublin Core Application Profiles. Accédé le 12 Mai 2017 de http://dublincore.org/documents/singapore-framework/

Nilsson, M., Miles, A. J., Johnston, P., et Enoksson, F. (2009). Formalizing Dublin Core Application Profiles–Description Set Profiles and Graph Constraints. *In Metadata and Semantics* (pp. 101-111). Springer US.

OGC (2007). Observations and Measurements. Part 1 – Observation Schema. OGC 07-022r1. (Accédé le 18 avril 2017 : http://portal.openspatial.org/files/?artifact_id=224566)

Ponte, J. M., et Croft, W. B. (1998). A language modeling approach to information retrieval. research and development in information retrieval. *In Proc. of the International ACM-SIGIR Conference (1998), Proc. Of the International ACM-SIGIR Conference, pp.275–281.*

Pouget, F. (2005). Notions de base de télédétection. Université de La Rochelle, France.

Powell, A., Nilsson, M., Naeve, A., Johnston, P. et Baker, T. (2007, Juin). DCMI Abstract Model. DCMI Recommendation,

Ranganathan, S.R. (1933). Colon clasification. Madras Library Association, Madras.

Ranganathan, S.R. (1952). Elements of classification. 3rd ed. Bombay: Asia Publishing House.

Ranganathan, S.R. (1967). Prolegomena to library classification. *Asia Publishing House (New York)*.

Rivest, S., Bédard, Y., Proulx, M. J., Nadeau, M., Hubert, F., et Pastor, J. (2005). SOLAP technology: Merging business intelligence with geospatial technology for interactive spatio-temporal exploration and analysis of data. *ISPRS journal of photogrammetry and remote sensing, 60(1), 17-33*.

Robertson, S. E. et Sparck Jones, K. (1976). Relevance weighting of search terms. Journal of the American Society for Information Science, 27, 129–146.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In The SMART Retrieval System, in Experiments in Automatic Document Processing G. Salton, editor, Prentice-Hall, Englewood Cliffs, NJ,pp. 313–323.

Salton, G., Wong, A., et Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM 18(11): pp. 613–620*. doi:10.1145/361219.361220

Salton, G., Fox, E.A. et Wu, H., (1983). Extended Boolean information retrieval system. *Communications of ACM 26(11), pp. 1022-1036*.

Santoro, M., Mazzetti, P., Nativi, S., Fugazza, C., Granell Canut, C., et Díaz Sánchez, L. (2012). Methodologies for augmented discovery of geospatial resources. *Geographic Information Systems: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications, 305*

Sarle, W.S. (1995) Measurement theory: frequently asked questions. *Originally published in the Disseminations of the International Statistical Applications Institute, 4th edition, 1995, Wichita: ACG Press, pp. 61-66. Revised 1996, 1997*.

Sayah H. (2011) : Composant Sémantique pour l'amélioration de la recherche de données environnementales. Mémoire de master recherche en informatique. Université de Montpellier. Juin, 2011. 63 p.

Simon, H. (1971). Designing organizations for an information-rich world. *In Computers, Communication, and the Public Interest. Baltimore, MD: The Johns Hopkins University Press*.

Smeulders A., Worring M., Santini S., Gupta A. et Ramesh, J. (2000). Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. Pattern Anal. Mach. Intell. p. 1349-1380*.

Shneiderman, B. (1994, Novembre). Dynamic queries for visual information seeking. *IEEE Software 11(6): pp. 70–77. doi:10.1109/52.329404*

Spärck Jones K. et Willett P. (1997). Readings in Information Retrieval. *Morgan and Kaufmann publishers, inc. San Francisco*.

Spink A. et Saracevic T. (1997) Interactive information retrieval: Sources and effectiveness of search terms during mediated online searching. *Journal of the American Society for Information Science, 48, 8, 1997, p. 741-761*.

Spiteri L., (1998). A simplified model for facet analysis. *Canadian Journal of Information and Library.Science, vol. 23, pp. 1-30.*

Shvaiko, P., Ivanyukovich, A., Vaccari, L., Maltese, V., et Farazi, F. (2010). A semantic geo-catalogue implementation for a regional SDI.

Stvilia, B., Twidale, M. B., Smith, L. C., et Gasser, L. (2008). Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology, 59, 983–1001. doi:10.1002/asi.20813*

Tissaoui, A., Aussenac-Gilles, N., Laublet, P. et Hernandez, N. (2013). Evonto : un outil d'évolution de ressource termino-ontologique pour l'annotation sémantique. *Technique et Science Informatiques, 32(7-8) :817–840, 2013.*

Toulet, A., Roux, E., Laques, A.E., Delaître, E., Demagistri, L. et Mougenot, I. (2017, Janvier). Extraction automatique de paysages en imagerie satellitaire et enrichissement sémantique. *EGC: Extraction et Gestion des Connaissances, Grenoble, France. 17ème conférence Extraction et Gestion des Connaissances.*

Tunkelang, D. (2009). Faceted search. Synthesis lectures on information concepts, retrieval, and services, 1(1), 1-80.

USGS (2015). Landsat 8 OLI/TIRS Collection 1 Data Dictionary. https://lta.cr.usgs.gov/Landsat_8_C1.html, accédé le 11 avril 2017

Vaccari, L., Craglia, M., Fugazza, C., Nativi, S., et Santoro, M. (2012). Integrative research: the EuroGEOSS experience. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 5(6), 1603-1611.*

Vilches-Blazquez, L. M., et Corcho, O. (2009). A heuristic approach to generate good quality linked data about hydrography. *In Proceedings of the Database and Expert Systems Applications, international Workshop, (pp. 99-103). IEEE Press.*

Vocabulaire International de Métrologie (2012). Termes basiques et généraux de métrologie. Bureau International des Poids et des Mesures. http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf, accédé le 9 Mai 2017.

Voorhees E. (1994). Query expansion using lexical-semantic relations . *In : Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, Dublin, 1994, Ireland, Springer-Verlag New York, Inc., p. 61-69*

Weibel, S., Kunze, J., Lagoze, C. et Wolf, M. (1998). Dublin core metadata for resource discovery (No. RFC 2413).

W3C (2004): SKOS Simple Knowledge Organization System. W3 Consortium, 2004. https://www.w3.org/2004/02/skos/, accédé le 14 Avril 2017

Yee, K. P., Swearingen, K., Li, K., et Hearst, M. (2003, Avril). Faceted metadata for image search and browsing. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 401-408). ACM.

Yoder, J. W., Balaguer, F. et Johnson, R. (2000). From analysis to design of the observation pattern. http://www.joeyoder.com/Research/metadata/Observation/ObservationModel.pdf, accédé le 9 Mai 2017.

Zayrit., K. (2010) Modèles de données adaptés à la construction partagée d'un thésaurus dédié aux traits fonctionnels. Mémoire de Stage de Master DECOL, UM2, Montpellier, France, 2010.