Standards de métadonnées et définition de profils d'application Dublin Core

Isabelle Mougenot

HDR

Habilitation à Diriger des Recherches

École Doctorale I2S - Université Montpellier

Spécialité: Informatique

par

Isabelle Mougenot

Standards de métadonnées et définition de profils d'application Dublin Core

Soutenue le 18 décembre 2015 devant le jury composé de

Mme.	Anne Doucet - Université Paris 6	(Rapporteur)
Mme.	Régine Vignes Lebbe - Université Paris 6	(Rapporteur)
M.	Pierre Gançarski - Université Strasbourg	(Rapporteur)
Mme.	Thérèse Libourel - Université Montpellier	(Examinateur)
Mme.	Sylvie Ranwez - Ecole des Mines d'Alès	(Examinateur)
Mme.	Malika Smaïl Tabbone - Université Nancy 1	(Examinateur)

HDR préparée au sein de
l'UMR 228 Espace Dev
500, Rue JF. Breton
Université Montpellier
34 093 Montpellier

Table des matières

Première partie

CV Étendu

Sommaire

1 Curriculum Vitae

Identité

Nom : Mougenot
Prénom : Isabelle
Née le : 29 juin 1964 à Valence (26)
Nationalité : Française
Email : isabelle.mougenot@umontpellier.fr

Adresse professionnelle :
Isabelle Mougenot
Unité Espace-Dev
CC83 Bâtiment 21
Place Eugène Bataillon
34395 Montpellier
Téléphone : 04 67 91 72 63

Diplômes de fin de cursus

Doctorat de 3^{me} cycle Université Montpellier 2 : Soutenu le 11 décembre 1995 à Montpellier, mention *Très Honorable*.
Sujet : « LIGM-DB un système de gestion intégré des séquences nucléiques des récepteurs d'antigènes et de leurs annotations.»
Président Claude Boksenbaum, Professeur Université Montpellier 2.
Rapporteurs Christian Gauthier, Directeur de Recherche CNRS (Lyon).
Etienne Pichat, Professeur Université Lyon 1 (Grenoble).
François Rechenmann, Directeur de Recherche INRIA (Grenoble).
Examinateur Thérèse Libourel, Maître de conférence CNAM (Montpellier).
Directeurs Marie-Paule Lefranc, Professeur Université Montpellier 2.
Claude Boksenbaum, Professeur Université Montpellier 2.

1990 - 1991 : DEA **INGENIA** (Méthodes d'obtention et d'analyse des données pour l'ingénierie des protéines et le séquençage de génomes), Université Montpellier 1 et 2, Université Aix-Marseille, mention *Assez-Bien*.

Poste d'enseignant-chercheur

septembre 2001 - décembre 2010 : **maître de conférences UM2** - section 27 - faculté des sciences **pour la composante enseignement** - laboratoire du LIRMM (département informatique et équipe DOC (Données, Objets et Connaissances)) **pour la composante recherche** ;
janvier 2011 - En cours : **maître de conférences UM2** - section 27 - faculté des sciences **pour la composante enseignement** - unité ESPACE-DEV **pour la composante recherche** (équipe MICADO).

Autres expériences professionnelles 1996-2001

1996 - 2000 : **contrat ingénioral** - dans le cadre d'un projet Européen EUREKA ;
2000 - 2001 : **contrat ingénioral** - dans le cadre de projets Européens IST.

2 Travaux de recherche

2.1 Préambule et schéma de synthèse

Pour une meilleure compréhension du parcours de recherche sur un long terme (le temps s'écoulant de haut en bas), une figure de synthèse (figure 1) reprend les principaux jalons des activités scientifiques menées, et en particulier les thèses supervisées en co-encadrement et les projets de recherche investis. Les travaux sont menés pour la plupart à l'interdisciplinaire en collaboration avec différents experts de disciplines différentes (biologie moléculaire, écologie, sciences cliniques, géographie, télédétection). De fait un code couleur donne un apercu des travaux à l'interface entre l'informatique (jaune) et les sciences du vivant (bleu) d'une part, et l'informatique et les sciences de l'environnement (gris) d'autre part. Un nuage de mots récapitule les principaux descripteurs en informatique des recherches menées.

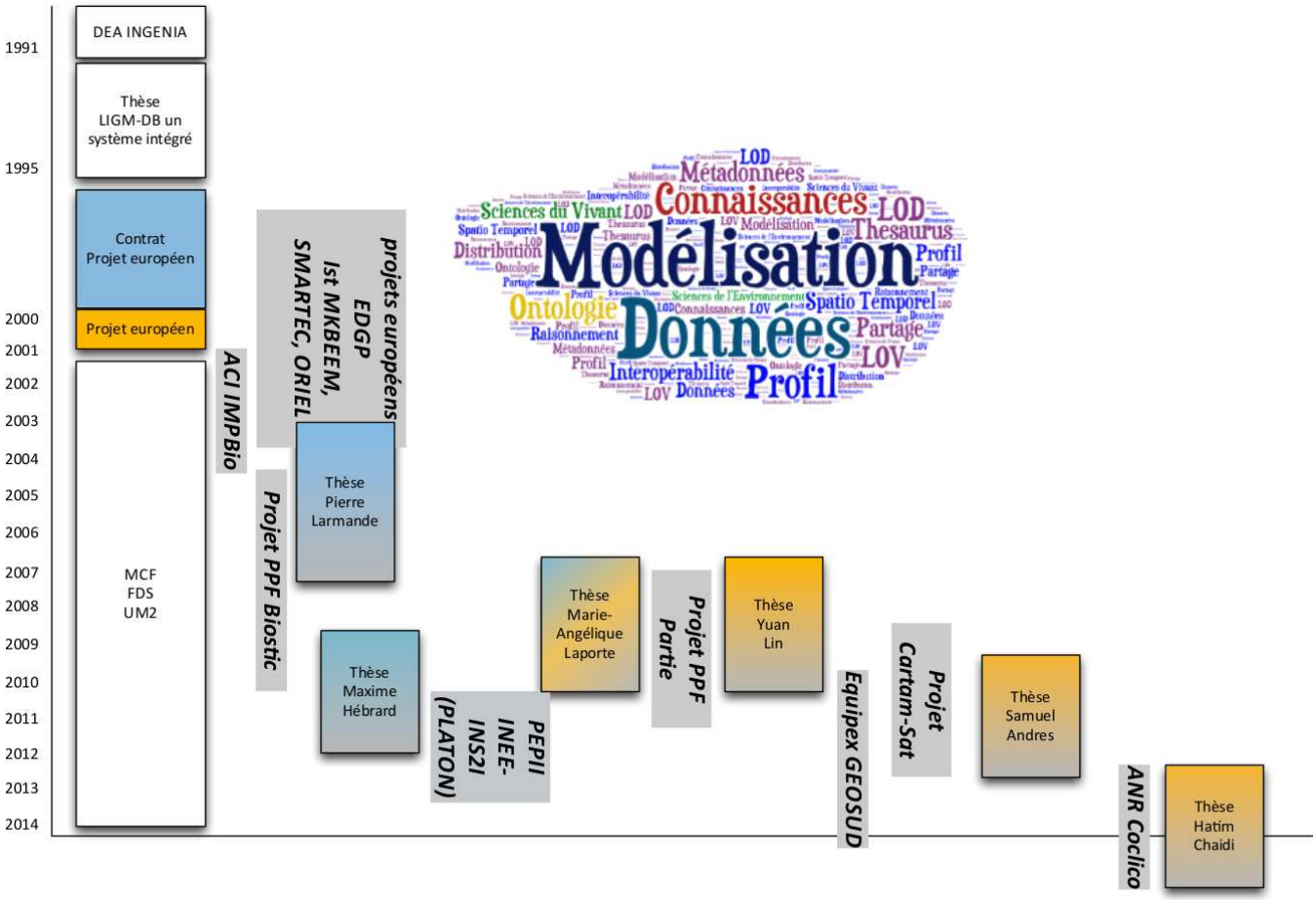

FIGURE 1 – Synthèse des principales activités

2.2 Conclusions sur les travaux passés

Les travaux menés sont objectivés par différents projets de recherche, qui pour certains, sont à l'interface entre l'informatique, les sciences du vivant et de l'environnement. La réflexion reste centrée sur la représentation, l'exploration et la mutualisation

de grandes masses de données. Outre la dimension "taille", la nature de ces données, à l'exemple des données biologiques ou géographiques, est particulièrement complexe.

2.2.1 Projets européens

- 1997-1998 - Activités de recherche bioinformatique au sein du projet de séquençage du chromosome X de la drosophile (European Drosophila Genome Project EDGP)[4] ;
- 2000-2002 - Construction interactive d'ontologies et de systèmes adaptatifs de gestion de transactions complexes - Projets Européens IST - Plateformes de commerce électroniques MKBEEM et SMARTEC[69] ;
- 2001-2004 - Représentation de connaissances dans le contexte de projets de grands séquençages - Projet Européen IST Oriel ; (interconnectivity between widely differing datasets that include genomic sequences, sequence-related information and multi-dimensional image collections. An EU-funded research project in information technology Online Research Information Environment for the Life Sciences Coordination EMBO)[84, 83, 34, 35].

2.2.2 Activités d'animation

- 2003-2006 - Portage du PPF UM2 (Plan Pluri-Formations) d'animation *BioStic*, partagé avec Jean-Pierre Nougier ;
- 2000-2003 - Responsabilité de l'ACI IMPBio (Action Concertée Incitative) *OntoBio* (Ontologies pour la Biologie) ;
- 2003-2006 - Implication dans le comité scientifique de l'ACI IMPBIO *ISIBio* Intégration des systèmes d'information en biologie ;
- 2003-2006 - Implication dans le PPF UM2 (Plan Pluri-Formations) *Gibi* Infrastructures à base de métadonnées pour la biologie ;
- 2007-2010 - Portage du PPF UM2 (Plan Pluri-Formations) d'animation *BioStic* ;
- 2007-2010 - Implication dans le PPF UM2 (Plan Pluri-Formations) d'animation *Partie* (Partage d'informations environnementales spatialisées) ;
- 2011-2013 - Portage du PEPII (PEPS Inter-Instituts) INEE - INS2I *Construction d'une ontologie centrée sur les traits des plantes pour l'écologie et l'agronomie (PLATON)*, partagé avec Eric Garnier.

2.3 Synthèse des travaux en cours

Mes travaux de recherche portent sur la modélisation de données et de connaissances, principalement dans les sciences du vivant et de l'environnement. Différentes questions de recherche sont explorées, qui mobilisent d'une part les principes d'interopérabilité de sources de données distribuées, et d'autre part les mécanismes qui vont permettre de poser des interprétations sur des grandes masses de données mises en commun. La distribution tout comme l'interopérabilité des sources de données sont abordées au travers des notions de métadonnée, de profil d'application et de solutions de gestion de données nommées NoSQL qui offent des mécanismes de persistance de l'information plus ouverts et moins contraints que le relationnel. Le partage et la mise en œuvre de mécanismes d'interprétation sur de grandes masses de données sont regardés sous l'angle

du web de données avec la conception et la construction de composants ontologiques et de thésaurus. Les langages RDF, RDFS, OWL et SPARQL constituent aujourd'hui le socle organisationnel de notre savoir collectif et permettent de mettre à disposition de tous, ce qu'il est convenu d'appeler les LOD (sources de données ouvertes) et les LOV (vocabulaires de données ouverts). Les communautés de pratique autour des domaines d'application considérés sont placées au centre du dispositif de recherche et les travaux menés s'attachent à restituer à ces communautés des approches méthodologiques, des modèles et des composants logiciels à même de leur faciliter l'accès aux données et la mise en commun de leur expertise. Différents axes scientifiques ont été investis sur les dernières années. Je reviens ci-dessous sur les contributions scientifiques les plus significatives.

2.3.1 Activités expertes autour de la construction de vocabulaires contrôlés

Organisation collective de la terminologie propre à un domaine d'intérêt [59, 63] Les sciences du vivant et de l'environnement s'appuient traditionnellement sur de nombreux dispositifs qui sont à l'origine de la production de larges volumes de données d'observation comme d'expérimentation. Ces données sont organisées au sein de multiples sources de données locales qui vont en assurer la gestion sans véritable réflexion sur leur mutualisation possible. Le besoin d'accéder, d'échanger, de partager et d'intégrer l'ensemble des données distribuées au sein de ces sources de données, est cependant essentiel pour la pérennisation de la donnée sur le long terme dans un premier temps et pour rendre optimale la valorisation de la donnée qualifiée dans un second temps.
La montée en puissance du web de données a apporté de nouvelles solutions aux verrous posés par l'hétérogénéité syntaxique, structurelle ou sémantique des données. Chaque source de données peut en effet être repensée, à tout moment, en terme de collection de triplets (déclarations formées d'un sujet, d'une propriété et d'un objet) qui sont comme autant de ressources sémantiques, échangeables et partageables. Ces ressources peuvent, en outre être complétées par différentes structurations spécifiées dans des langages de représentation tels que RDFS ou OWL, et/ou enrichies par d'autres ressources qui leur sont complémentaires. Cette flexibilité des modèles possède différents avantages. Il est ainsi possible de faire évoluer en continu la structure et le contenu d'une collection, sans avoir à modifier ni la structure, ni le contenu d'origine. Ces évolutions peuvent s'opérer par différents apports de collections tierces, et de manière décentralisée, puisque rien n'empêche un même triplet d'être construit sur la base des trois éléments, sujet, prédicat ou objet, qui proviennent de différentes collections. En conséquence, la mutualisation et l'intégration des collections de ressources n'affecte en rien l'autonomie de chacune des collections impliquées. Il s'agit donc de donner aux scientifiques les moyens de construire ces collections de ressources, de manière à les partager à moindre effort. L'attendu est non seulement de favoriser le partage de l'expertise mais aussi de faciliter la qualification et la valorisation des données Pour ce faire, il est convenu de s'accorder sur la définition des termes désignant les concepts d'intérêt pour les réseaux scientifiques d'expertise. Un cadre méthodologique *(travaux de thèse de Marie-Angélique Laporte)* est proposé aux experts de manière à construire par eux mêmes, et de manière communautaire leurs vocabulaires d'intérêt. Le standard W3C de construction de schémas de concepts SKOS est mis à profit. De même les standards de métadonnées du Dublin Core viennent instrumenter les activités d'édition et de validation des éléments du vocabu-

laire. Un modèle intégré s'adossant au modèle abstrait du Dublin Core (DCAM) est au cœur de la démarche et permet également de superviser et de tracer toutes les activités de manière à anticiper les opérations concurrentes conflictuelles. Une application web nommée *Thesauform* a validé la pertinence de l'approche et a permis notamment à une communauté travaillant en écologie fonctionnelle et forte d'une vingtaine d'experts localisés dans différents pays (en particulier France, Allemagne, Pays-Bas, USA, Amérique du Sud et Australie) de définir un thésaurus dédié aux traits fonctionnels. Le thésaurus ainsi construit est nommé TOP (Traits Of Plants) et est riche d'environ un millier de descripteurs (concepts SKOS) qui couvrent une grande majorité des traits fonctionnels habituellement manipulés par les écologues pour établir des liens avec les grandes fonctions des plantes (reproduction, activité photosynthétique, ...).

Définition d'un système de recherche à facettes adossé à un thésaurus [61, 62, 66] Un thésaurus permet aux experts d'un domaine de partager de manière consensuelle les descripteurs (ici traités comme des concepts SKOS) définissant leurs entités d'intérêt. Les descripteurs sont alors organisés au travers d'une taxonomie de généralisation/spécialisation (structuration verticale) et de liens de voisinage (structuration horizontale). Il peut s'avèrer pertinent de proposer des structurations complémentaires qui vont avoir l'avantage de ne pas prétendre à la consensualité ou à l'exhaustivité mais plutôt de rechercher une meilleure adéquation face aux approches scientifiques de communautés d'experts ou encore un meilleur interfaçage au regard d'une source de données en particulier. Le standard SKOS propose à cet effet la notion de collection et de collection ordonnée qui permettent de définir de nouveaux liens méréologiques sur un ensemble de concepts SKOS ou sur un ensemble d'ensembles de concepts SKOS (collection de collections). Une collection SKOS est ainsi assimilée à une facette. Une facette s'attache à décrire une donnée selon un certain point de vue et va donc en filigrane faciliter la structuration pluri-dimensionnelle des données tout en permettant de filtrer l'information relative à une dimension en particulier. Dans la communauté de l'écologie fonctionnelle, il est ainsi pertinent de définir des facettes sur la base des organes qui constituent la plante et qui vont permettre de regrouper les traits fonctionnels en fonction des parties de la plante à partir desquelles ils sont mesurés. Nous définissons ainsi une approche plus naturelle pour les écologues qui ont pour habitude de ne travailler que sur un aspect de la plante en particulier (feuille, tige, racine ou fruit). Prenons un exemple, les traits fonctionnels `SLA (Specific Leaf Area)`, `LMA (Leaf Mass per Area)` et `LBT (Leaf Blade Thickness)` sont tous les trois associés à la feuille et à cet effet sont organisés au sein d'une collection `Leaf`. Un écologue qui travaille sur la feuille d'une plante en particulier pourra filtrer l'information relative aux traits fonctionnels en choisissant de ne s'intéresser qu'aux traits fonctionnels de la feuille. Outre le fait que cette approche satisfait au mieux les besoins des utilisateurs, les avantages de cet ajout de structuration sont multiples et peuvent aller de l'interconnexion de vocabulaires ouverts à l'exemple de ce qui se pratique dans l'initiative LOV (Linked Open Vocabularies) à de l'interopérabilité entre sources de données. Dans les travaux conduits actuellement, le thésaurus TOP est étendu au travers de différents modèles de facettes qui permettent de le lier d'une part au vocabulaire Plant Ontology et d'autre part à la base de données TRY database qui est la base de données de référence consacrée aux traits fontionnels (des millions de mesures de 800 traits différents pour 60000 espèces de plantes).

2.3.2 Standards de métadonnées et définition de profils d'application Dublin Core [30]

L'accent est de plus en plus mis dans nos sociétés sur l'accès, le partage et l'échange de l'information ce qui oblige à repenser la manière de gérer cette information de manière à en décentraliser les accès et à en favoriser l'appropriation et/ou la réutilisation par le plus grand nombre. La volumétrie des données pose également question. Par exemple, les capteurs sont devenus omniprésents dans notre quotidien et font exploser le nombre de données à gérer. Ces capteurs vont de simples instruments de mesure, à l'exemple de capteur de température ou d'humidité, à des capteurs imageurs dédiés à l'observation de la Terre et embarqués sur des satellites. Les dynamiques dans le temps et dans l'espace font en conséquence, également, parties de la compréhension de notre environnement et de l'évolution de la société. Nous nous intéressons ici au domaine de l'observation de la Terre et notamment à l'extraction d'indicateurs sociaux-environnements pertinents dans des images satellites.
Plutôt que de travailler sur la donnée elle même qui est souvent volumineuse à l'exemple des images satellites, l'idée est de travailler sur une *représentation* de cette donnée qui va agir à la manière d'un filtre pour permettre de ne traiter que la partie utile de la donnée face aux usages qui en sont attendus. Les avantages d'une telle approche sont nombreux comme ne partager que l'information nécessaire, alléger la charge en terme de stockage et de traitement des données, protéger les informations sensibles, faciliter une vision distribuée avec la publication de modèles ne visant pas l'exhaustivité mais des fonctionnalités cibles ou encore faciliter le contrôle de qualité sur la donnée.
Cette *représentation* s'appuie sur une collection d'éléments de métadonnées et se manipule sous la forme d'une collection d'enregistrements d'éléments de métadonnées liés. Différents travaux ont été menés autour de la construction de profils d'application. Nous nous intéressons en particulier à la mise en œuvre de profils d'application Dublin Core (DCAP). L'initiative Dublin Core est à l'origine d'une plateforme de définition de profils d'application (framework Singapore) qui donne accès à différents modèles et standards, et offre une articulation modulaire de ces modèles. Dans le contexte des stages de master informatique 2014 (*publication en cours*), nous retenons plus particulièrement le modèle DCAM (pour Dublin Core Abstract Model) qui offre un cadre général à la description de ressources à partir de métadonnées abordées comme des collections de couples propriétés/valeurs, ainsi que le modèle DSP (Description Set Profiles) qui offre un cadre prescriptif aux profils d'application. Nous nous appuyons également sur les bonnes pratiques préconisées autour de la construction d'un profil d'application. Ainsi, nous avons défini des cas d'utilisation qui viennent illustrer les exigences fonctionnelles. Nous avons défini un modèle conceptuel des entités d'intérêt dans le contexte de l'observation de la terre. Un dictionnaire des éléments de métadonnées retenus pour la description des entités d'intérêt et les contraintes qui viennent s'y appliquer, a été également construit. Ces éléments sont choisis à partir de différents standards de métadonnées à l'exemple de Dublin Core Terms ou d'ISO 19115. Sur la base du dictionnaire, un modèle d'éléments de métadonnées conforme au métamodèle DSP a été spécifié. Les langages choisis pour opérationnaliser ce modèle sont RDF/RDFS/OWL. Les perspectives à court terme consistent à alimenter ce modèle au travers d'enregistrements d'éléments de métadonnées et à mettre en place les applicatifs à même de l'exploiter pour faciliter une recherche et un traitement unifiés de ressources distribuées issues de l'observation de la terre. **Ces travaux de recherche sont les seuls développés dans le mémoire d'HDR**

2.3.3 Modèle de workflow scientifique multi-niveaux [70, 76, 78, 73, 75, 74, 71]

Les scientifiques à l'exemple des télédétecteurs ou encore des biologistes font un large usage de dispositifs de traitement de manière à mener leurs recherches au mieux, et en particulier de valider ou invalider leurs théories et hypothèses de travail. Si nous nous intéressons aux sciences du vivant, les biologistes ont l'habitude de désigner par biologie in silico, l'ensemble des expérimentations et simulations réalisées à partir des sources de données et des briques logicielles d'analyse, mises publiquement à disposition. La biologie in silico s'est imposée sur les trois dernières décades notamment en biologie moléculaire et dans la pratique, les biologistes l'intégrent à leurs méthodologies d'expérimentation s'appuyant également sur des dispositifs expérimentaux in vivo, ex vivo and in vitro.
Les ressources, sources de données et outils d'analyse, sont multiples et il est parfois difficile pour un biologiste d'en tirer le meilleur parti. De même, les scénarios d'analyse pour arriver aux objectifs visés, à l'exemple de rechercher les gènes orthologues d'un gène d'intérêt dans une comparaison inter-espèces, sont complexes et nécessitent de chaîner différents traitements et de consulter plusieurs sources de données.
L'objectif est donc de proposer une infrastructure de construction de workflows scientifiques *(travaux de thèse de Yuan Lin)* qui est volontairement générique et qui a été testée avec succès dans les domaines de la télédétection et de la post-génomique. Cette infrastructure s'appuie sur trois niveaux. Le niveau le plus général (**niveau abstrait**) s'appuie sur un environnement organisationnel, qui va permettre à l'expert de construire son plan de travail (squelette du workflow). Il s'agit d'un travail purement conceptuel, l'expert chaîne des activités de calcul et indique les grandes catégories de données en entrée et sortie de ces activités. Le niveau intermédiaire (**niveau concret**) va s'attacher à proposer une concrétisation sur la base du plan de workflow construit au niveau abstrait.
Il n'est nullement demandé à l'expert d'avoir une parfaite connaissance de l'ensemble des outils logiciels ou des sources de données dont il a besoin. La phase intermédiaire va mettre en correspondance d'une part les activités de calcul et les outils logiciels ; et d'autre part les grandes catégories de données / sources de données. Cette phase est semi-automatisée, les experts peuvent effectuer des sélections à partir des listes de choix qui leur sont proposées. Des vérifications de conformité sont effectuées, qui portent notamment sur la validité des types et formats de données en entrée/sortie des logiciels de traitement. La phase de concrétisation prend également en charge les aspects relevant de l'optimisation des calculs. Ainsi les traitements indépendants seront indiqués en tant que tâches parallélisables. Le troisième niveau (**niveau opérationnel**) est le niveau d'exécution proprement dit. Les derniers ajustements notamment au regard de la configuration du serveur de calcul seront réalisés à ce niveau. Notre travail s'est surtout concentré sur les niveaux abstrait et concret avec une modélisation conceptuelle de l'environnement organisationnel, la spécification d'un langage de définition de workflows scientifiques et l'élaboration d'algorithmes pour l'évaluation de la conformité. Le niveau opérationnel n'a a été abordé qu'en partie mais fait actuellement l'objet de travaux de développements dans le cadre de l'implémentation de l'infrastructure de données spatiales du projet GeoSud.

2.3.4 Cadre ontologique pour l'aide au diagnostic dans le contexte des rétinites pigmentaires [49]

Les dystrophies rétiniennes désignent une famille de maladies rares qui causent, d'une manière ou d'une autre, une dégénérescence de la rétine. Les symptômes vont être de diverses natures à l'exemple de la cécité de nuit, de la perte de la vision centrale ou encore de différentes lésions oculaires. De plus ces maladies sont évolutives et très invalidantes, elles entraînent une perte progressive de la fonction à la fois des photorécepteurs (cônes et batonnets) et de l'épithélium pigmentaire de la rétine. Ces maladies sont très hétérogènes sur les plans génétique et clinique. Il a été montré par exemple, que plus de cinquante gènes pouvaient être des facteurs causaux d'une de ces pathologies. De même certains gènes présentent différentes mutations qui pourraient être à l'origine de différentes pathologies. Le gène ABCA4 par exemple, présente plus de 400 mutations et pourrait être en cause dans les maladies de Stargardt, de la dégénérescence maculaire liée à l'âge (DMLA) ou encore de la rétinite pigmentaire autosomique récessive RP19. Il n'existe pas véritablement de signe clinique qui puisse permettre de discriminer chacune des maladies. Le praticien a donc recours à différents examens cliniques et interroge également le patient pour pouvoir à partir de son expertise, poser un diagnostic. Pour une meilleure compréhension des mécanismes génétiques et phénotypiques et pour pouvoir poser des diagnostics cliniques sûrs, nous nous appuyons sur une approche méthodologique à base de représentation de connaissance *(travaux de thèse de Maxime Hébrard)*, de manière à modéliser de manière fine les aspects cliniques liés aux rétinites pigmentaires.
L'approche prend en considération les maladies et l'expertise des cliniciens autour de ces maladies, les patients et une description de leurs signes cliniques. Un module ontologique a été défini en logique de description OWL-DL et prend en charge la représentation des profils cliniques de maladie comme de patient. Un profil est vu comme une collection de symptômes qui peuvent être liés et revêtir différents états susceptibles d'évoluer dans le temps. Nous réutilisons l'existant et des schémas de concepts (ou ontologies terminologiques) à l'exemple de HPO, PATO ou SNOMED sont exploités de manière à venir typer les symptômes et/ou leurs états. Un premier travail a consisté à alimenter le modèle de connaissance avec les profils des patients dont le diagnostic ne pouvait pas être remis en cause. Les profils de maladies, également conformes au modèle de connaissances, ont ensuite été construits par agrégation des profils de patients affectés par cette maladie. Un ratio est ajouté pour chaque couple symptôme/état qui permet d'évaluer la pertinence de ce symptôme à cet état au regard de l'ensemble des patients atteints par cette maladie dans une catégorie d'âge donnée. Les profils d'individus comme de maladies sont ainsi rendus comparables au travers de mesures de similarité sémantique. Nous avons dans une première approche exploité la mesure de similarité basée sur le modèle de contraste de A. Tversky qui prend en considération la présence ou l'absence de caractéristiques communes pour comparer deux objets. Les objets ici sont les profils et les caractéristiques sont les collections de symptômes définis à un état donné. Un premier résultat pour les cliniciens porte sur une nouvelle manière de représenter les rétinites pigmentaires et est à même de laisser transparaître des ressemblances et/ou différences qui auraient pu rester sous silence. Un deuxième résultat est de permettre de faciliter le diagnostic de nouveaux patients en confrontant leur profil aux profils des maladies alimentant le système.

2.3.5 Composants ontologiques autour des images spatiales [2, 102]

La réflexion porte ici sur les apports des ontologies dans la production d'indicateurs socio-environnementaux pertinents, à partir de l'imagerie satellite. Les zones géographiques étudiées sont en particulier, les régions du Sud à l'exemple des ROM-COM, des territoires Amazoniens et des oasis en Afrique du Nord. Les images exploitées sont à haute résolution (Landsat) voire à très haute résolution (SPOT). L'approche reste néanmoins générique et peut s'appliquer à toute problématique liée à l'image satellite [82]. La représentation des connaissances en télédétection relève par essence, d'une approche pluridisciplinaire à l'intersection du traitement d'image, de l'ingénierie des connaissances, de la géographie et de la télédétection. De fait, elle a fait l'objet de travaux à finalités diverses. Une première approche est par exemple d'exploiter les ontologies pour guider les activités de segmentation d'une image classiquement conduites en télédétection et en améliorer le résultat.
Pour ce qui nous concerne, nous avons choisi d'exploiter les ontologies comme conteneurs de connaissances sur lesquels des techniques de classification sémantique ont été appliquées au niveau pixel ou au niveau segment. Les résultats sont encourageants, aux dires des experts du domaine. Nous nous intéressons ici à l'explicitation de connaissances implicites par raisonnement automatique *(travaux de thèse de Samuel Andrès)*, la mise en œuvre de mécanismes permettant d'interpréter le contenu d'une image se heurte à ce qui est désigné par fossé sémantique. L'image satellite est une donnée composite numérique et la connaissance experte est qualitative. Le fossé sémantique est le reflet du manque de concordance entre la représentation matricielle d'une image et l'interprétation que va en tirer un expert dans un contexte donné. Le fossé sémantique est à rapprocher de ce qui est dénommé ancrage de symboles en ingénierie des connaissances.
L'objectif, pour ce qui concerne l'extraction de connaissances à partir d'images, est dans un premier temps, d'associer des "valeurs" qualitatives (ou éléments conceptuels) à des intervalles de valeurs calculés pour un indicateur radiométrique donné à partir de l'image. Les éléments conceptuels sont décrits au sein de l'ontologie image. Nous avons construit un module ontologique autour de l'image spatiale qui vient s'adosser à l'ontologie cadre OBOE. OBOE (ou Extensible Observation Ontology) est dédiée à la description d'observations et mesures scientifiques. Dans OBOE, une attention particulière est portée sur l'explicitation du contexte de l'observation qui est constitué également d'observations. Une image satellite est le produit d'une observation à partir d'un capteur radiométrique. Les images sont souvent disponibles au travers de séries temporelles. Il est ainsi possible au travers du modèle OBOE d'exprimer le fait que d'autres images vont constituer le contexte d'une image étudiée et ainsi en enrichir la caractérisation. La construction du module ontologique s'est faite en trois étapes, avec la définition d'un modèle conceptuel exprimé au travers de la notation UML, la définition d'un modèle ontologique exprimé en logique de description et la spécification d'un modèle opérationnel en OWL DL. La prise en charge de la modularité dans la représentation de la connaissance autour de l'imagerie spatiale est essentielle. La représentation d'une image comme ayant pour parties des pixels, des segments ou des objets, va s'envisager comme une représentation de référence. A l'inverse, les traitements qui vont aboutir à la définition des segments ou des objets au sein de l'image, vont être très fortement dépendants des objectifs visés, et des méthodes de traitement utilisées et donc en quelque sorte du contexte. De même les experts vont devoir faire différents choix à l'exemple de la définition de seuils sur la base des intervalles de valeurs associés

aux indices radiométriques pour tendre vers du qualitatif (ancrage de symboles et fossé sémantique). L'idée est donc de pouvoir interfacer une ontologie image de référence à plusieurs ontologies contextuelles qui vont être spécifiques à un périmètre scientifique. L'approche modulaire va également se révéler pertinente lorsqu'il va s'agir de comparer les segments labellisés ou les objets de l'image avec une vérité terrain. Il est alors question de rajouter un module de connaissance supplémentaire pour prendre en charge cette vérité et faire également appel aux ontologies terminologiques du domaine (Corine Land Cover, LCCS, GHC, ...) de façon à normaliser et à mieux partager les descripteurs venant qualifier les entités d'intérêt décrites au sein d'une image.

2.3.6 Conclusion et perspectives de recherche

Les différents composants structurels de données et de connaissances (standards de métadonnées, thésaurus, ontologies et profils d'application) sont capitalisables et nous permettent d'aborder des projets comme ceux de l'ANR Coclico ou de l'Equipe GeoSud, et de renforcer l'animation pluridisciplinaire.

2.4 Investissements au sein de projets scientifiques

- Responsable pour l'unité ESPACE-DEV en tant que partenaire du projet ANR COCLICO (COllaboration, CLassification, Incrémentalité et COnnaissances) [1] Coordinateur Pierre Gançarski *projet de recherche pour l'analyse multi-échelle de grands volumes de données spatio-temporelles mettant en œuvre une approche multistratégie adossée à différentes méthodes de fouille de données et guidée par des connaissances thématiques (géosciences, géographie, géomatique et télédétection) et du domaine de l'analyse (connaissances sur les méthodes) formalisées sous la forme de composants ontologiques* ANR Modèles Numériques 2012 funded by French Agency for Research ANR (Dec. 2012 - Nov. 2016) ;
- Implication dans le projet européen FP7 EOPOWER (Earth Observation for Economic Empowerment) [2] *le projet EOPOWER se consacre à la définition de solutions innovantes mettant en œuvre les nouveaux moyens d'observer la Terre. Les objectifs en sont de faciliter les développements de produits et de services en particulier écosystémiques dans une vision de gestion durable et économiquement viable de ressources naturelles.* ;
- Co-responsabilité WP3 Equipex GEOSUD (2012-2017) (GEOInformation for SUstainable Development) project funded by "Équipements d'Excellence" Call (2011). [3] Infrastructure d'Information Spatiale sur les Territoires et l'Environnement[82]. Le WP3 est le volet recherche de l'Equipex et je supervise pour l'unité Espace-Dev les recherches effectuées sur les apports méthodologiques, relatifs aux traitements de l'imagerie satellitaire, qui en constituent sa valeur ajoutée.

1. http://icube-coclico.unistra.fr/index.php/Coclico
2. http://www.eopower.eu
3. http://geosud.teledetection.fr/

2.5 Collectif de recherche

2.5.1 Responsabilités d'équipe

J'ai assuré de janvier 2011 (date de la création de l'UMR Espace-Dev) à juin 2013, le rôle de responsable de l'équipe SIC (Systèmes d'Information et de Connaissances). L'équipe SIC, nouvellement renommée MICADO (Modélisation, Ingénierie des Connaissances et Analyse des DOnnées spatiales) est constituée d'un peu plus d'une quinzaine de personnes qui se partagent, de façon classique, entre permanents (1 PR UM2, 1 MCF UM2, 1 DR IRD, 2 CR IRD, 1 IR IRD), et non permanents (doctorants, post-doctorants, ATER et ingénieurs contractuels). Cette responsabilité a correspondu à une gestion administrative et financière ainsi qu'à l'organisation de réunions internes régulières et à la définition du plan de route de l'équipe. MICADO [4] constitue avec les équipes OSE (Observation Spatiale de l'Environnnement) et AIMS (Approche Intégrée Milieux et Sociétés), l'unité Espace-Dev. MICADO se spécialise dans la modélisation à la fois symbolique et numérique de données environnementales, en exploitant pour partie des données issues de la télédétection et des observations in situ. Plus en détail, les axes scientifiques abordés dans MICADO sont en particulier :

- infrastructures de données et gestion de grandes masses de données ;
- métadonnées, ontologies et représentation de connaissances spatio-temporelles ;
- intégration de données multi échelles (images satellites, observations in situ) ;
- théorie des graphes appliquée aux réseaux sémantiques thématiques ;
- couplage de modèles multi paradigmes (en particulier numérique et symbolique) ;
- interprétation automatisée des images satellites pour répondre à des enjeux environnementaux (santé et environnement, écologie, géographie).

2.5.2 Animations scientifiques internes à l'unité

J'anime depuis 2013 l'axe transversal de recherche *"Ontologies et ingénierie de la connaissance"* au sein de l'unité. Depuis environ un an, l'unité Espace-Dev s'est recentrée, sur différents thèmes intégrateurs qui font l'objet d'axes de recherche transversaux. Les activités de modélisation sémantique et en particulier la construction de composants ontologiques dans les sciences du vivant, de l'environnement et de la santé mobilisent différents chercheurs et enseignants-chercheurs appartenant aux trois équipes de l'unité. L'animation de cet axe transversal s'articule autour de réunions de travail en petits groupes pluridisciplinaires. Les ordres du jour sont fixés à l'avance. La dynamique émanant de l'animation autour de l'axe *"Ontologies et ingénierie de la connaissance"* a été mentionnée dans le rapport 2014 du comité d'évaluation de l'AERES, comme étant propice à l'exploration de certaines ruptures théoriques et technologiques. De plus, le rapport du comité d'évaluation de l'AERES souligne en point fort l'existence de ce groupe de travail avec la possibilité de généraliser sa portée au niveau Montpelliérain.

2.5.3 Encadrement de stages

1. Encadrement de stages de master de recherche

4. `http://www.espace.ird.fr/index.php?option=com_content&view=category&layout=blog&id=45&Itemid=66`

- Yuan Lin (co-encadrement avec Thérèse Libourel) - *Ingénierie des composants pour la mise en œuvre de workflows scientifiques* - juin 2008 ;
- Karima Zayrit *Ontologies pour la médiation et l'intégration de données* - juillet 2010 ;
- Franklin Boumda *Médiation de données dans le contexte des agrosystèmes à babaçu* - septembre 2012 ;
- Hatim Chahdi (co-encadrement avec Jean-Christophe Desconnets IRD) - *Métadonnées et interconnexion de sources de données hétérogènes* - juillet 2013 ;
- Mojdeh Soltan Mohammadi (co-encadrement avec Christophe Fagot Intactile Design) - *Gestion distribuée de masses de données spatio-temporelles* - juillet 2014.

2. Encadrement de stages de mémoire d'ingénieur ENIT
   - Ibtihel Rebhi Juillet 2012 Ecole Nationale d'ingénieurs de Tunis (ENIT) *Processus de fouille de texte consacré à la construction d'un thésaurus dédié aux systèmes oasiens tunisiens.*
3. Encadrement de stages de mémoire d'ingénieur CNAM
   - Cédric Bouttes Juillet 2005 *Création d'une application intégrée pour la gestion et l'analyse de données protéomiques* ;
   - Frédéric Gomez Juillet 2007 *Création d'une application serveur pour la qualification et l'échange de données provenant d'appareils de mesure dédiés via les protocoles Ethernet et TCP/IP* ;
   - Martine Hornby Fin 2007 *Conception et développement objet d'une plateforme d'intégration de données écologiques* ;
   - Stéphane George Juillet 2013 *Contribution à la conception d'une ferme de calculs pour la plateforme bioinformatique ATGC.*
4. Encadrement de stages de master professionnel
   - Nordine El Hassouni M2 Aigle *Conception et développement d'une architecture de stockage et d'interrogation de lots de métadonnées au format RDF* (co-encadrement avec Jean-Christophe Desconnets IRD) - prévu septembre 2014 ;
   - Abderrazzak Loukili M2 Aigle *Environnement web pour l'exploitation de métadonnées RDF - application aux données issues de l'observation de la terre* (co-encadrement avec Jean-Christophe Desconnets IRD) - prévu septembre 2014 ;
   - Karina Castillo Perez *Conception de la base de données PhoenixDB dédiée aux marqueurs moléculaires du palmier dattier* M2 Bioinformatique (co-encadrement avec Frédérique Aberlenc IRD) - septembre 2010 ;
   - Arnaud Charleroy M1 BCD *Aspects spatio-temporels dans les métadonnées des fichiers de séquence au format INSDC* - prévu juillet 2014.

2.5.4 Encadrement doctoral

- Pierre Larmande (50% co-encadré avec T. Libourel et M. Ruiz) - *Mutaliser et partager, un défi pour la génomique fonctionnelle végétale* - décembre 2007, I2S UM2 ;
- Yuan Lin (40% co-encadré avec T. Libourel) - *Méthodologie et composants pour la mise en oeuvre de workflows scientifiques* - décembre 2011, I2S UM2 ;
- Marie-Angélique Laporte (50% co-encadré avec E. Garnier) -*Définition de standards de données relatifs aux traits fonctionnels des végétaux pour l'étude de la biodiversité* - décembre 2011, SIBAGHE UM2 ;

- Maxime Hébrard (50% co-encadré avec C. Hamel) - *Conception et développement d'un système d'aide au diagnostic clinique et génétique des rétinopathies pigmentaires* - décembre 2012, CBS2, UM1 ;
- Samuel Andrés (50% co-encadré avec T. Libourel) - *Les ontologies dans les images satellitaires* décembre 2013, I2S, UM2 ;
- Hatim Chahdi (60% co-encadré avec L. Berti et Y. Bennani) - *Modélisation et prise en compte des connaissances dans les processus d'analyse exploratoire et de fouille des données incrémentales et collaboratives* en cours de réalisation depuis octobre 2013.
- Mojdeh SoltanMohammadi (60% co-encadré avec L. Berti, T. Libourel et C. Fagot) - *Suivi de trajectoires spatiotemporelles d'objets mobiles et définition de règles sémantiques de supervision de comportements anormaux* en cours de réalisation depuis avril 2015.

2.5.5 Participation en tant qu'examinateur à des jurys de thèse

- Cyril Berthenet - Thèse de doctorat en Biochimie et biologie moléculaire. Sous la direction de Ned Lamb - 2007. à Montpellier 2 ; *Méthode computationnelle pour la prédiction de la mobilité des peptides et l'identification de leurs sites de phosphorylation par empreinte phospho-peptidique bidimensionnelle sur couche mince de cellulose* ;
- Fabien Jalabert - Thèse de doctorat en Informatique - *Cartographie des connaissances : l'intégration et la visualisation au service de la biologie* Université de Montpellier 2, Ecole doctorale Information, Structures et Systèmes. 2007 ;
- Julien Wollbrett - *Génération semi-automatique de Services Web Sémantiques pour des bases de données relationnelles biologiques* CIRAD, Montpellier. Doctorat SIBAGHE 2011 ;
- Sébastien Harispe - Thèse de doctorat en Informatique - *Mesures sémantiques à base de connaissance : de la théorie aux applicatifs* Université de Montpellier 2, Ecole doctorale Information, Structures et Systèmes. 2014.

2.5.6 Participation à des comités de suivi de thèse

J'ai participé à plusieurs comités de suivi de thèse, principalement dans trois écoles doctorales de l'UM2 (I2S, CBS2 et SIBAGHE). Je liste ci-dessous le nom des étudiants en thèse et leur école doctorale d'appartenance, que j'ai eu notamment l'occasion de suivre.

- Fabien Jalabert EERIE Nîmes, I2S ;
- Lucile Soler CIRAD Montpellier, SIBAGHE ;
- Julien Wollbrett CIRAD Montpellier, SIBAGHE ;
- Mehdi Yousfi Monod LIRMM Montpellier, I2S ;
- Benjamin Arnaud LIRMM, Montpellier, I2S ;
- Sébastien Harispe EERIE Nîmes, I2S ;
- Julien Osman CESBIO Toulouse ;
- Louis-Vincent Fichet, Montpellier, SIBAGHE ;
- Hugo Alatrista Tetis IRSTEA Montpellier, I2S.

2.5.7 Animations scientifiques

- Organisation de journées satellites au sein de la conférence nationale JOBIM (Journées Ouvertes : Biologie, Informatique et Mathématiques) ;
  1. OSGB'05 (Ontologie, Grille et Intégration Sémantique Intégration Sémantique pour la Biologie) Lyon 2005 ;
  2. OSGB'06 (Ontologie, Grille et Intégration Sémantique Intégration Sémantique pour la Biologie) Bordeaux 2006 ;
  3. MOQA (Métadonnées, Ontologies et Qualité des Annotations) Montpellier 2010.
- Participation à l'organisation de la session poster au sein de la conférence internationale DILS (Data Integration in the Life Sciences) Evry 2008 ;
- Participation à l'organisation de la session "Ecoinformatique et Ontologie en Ecologie" Ecologie 2010 Montpellier 2010 ;
- Participation à la mission scientifique internationale en écoinformatique, TraitNet [5] ;
- Participation à l'organisation de la journée satellite S4BioDiv 2013 au sein de la conférence européenne ESWC (European Semantic Web Conference) Montpellier 2013 (proceedings disponible `http://ceur-ws.org/Vol-979/`) ;
- Animation axe transversal Ontologies au sein de l'unité ESPACE DEV (en cours) ;
- Animation locale inter-UMR IATE, MISTEA, TETIS et ESPACE DEV autour du séminaire MIAD (Modèles informatiques autour de l'aide à la décision en environnement, agronomie et transformation) (en cours).
- Participation au comité éditorial du numéro spécial *Nouvelles avancées en systèmes d'information pour l'environnement*, Revue Ingénierie des Systèmes d'Information, 2014

3 Activités d'enseignement

Je considère les activités d'enseignement comme étant tout aussi essentielles que les activités de recherche. En conséquence, je cherche à donner du sens et une cohérence à mon travail d'enseignant.

3.0.8 Responsabilités de formations

Le poste d'enseignant-chercheur occupé, à l'interface entre informatique et biologie, m'a naturellement amenée à m'impliquer fortement dans des formations qui relèvent de l'informatique complémentaire. A ma prise de fonction, à la rentrée 2001, le DESS de bioinformatique était à sa toute première année d'existence, et j'en ai assuré la direction d'études pour une promotion d'une vingtaine d'étudiants. J'ai ensuite pris la responsabilité de ce DESS en 2003 et 2004. Avec la réforme LMD, le DESS de bioinformatique est devenu un parcours du mastère spécialité IPS (Informatique Pour les Sciences) et j'en ai conservé la responsabilité jusqu'en 2006. J'ai également dirigé le mastère IPS renommé IC (Intégration de Compétences) de 2006 à 2008. Ces différentes responsabilités

5. `http://traitnet.ecoinformatics.org/traitnet-participants`

ont donné lieu à une forte implication dans les activités qui en découlent de manière habituelle et qui sont listées ci-dessous pour certaines d'entre elles.

- coordination de l'équipe pédagogique (sélection des dossiers de candidature, mise en place des jurys de fin de semestre, ...) ;
- organisation des enseignements ;
- participation à la rédaction des maquettes d'habilitation ;
- interactions multiples avec les étudiants ;
- concertation avec les référents des formations dans les services de la scolarité et du planning.

N'étant plus responsable de formation sur les dernières années, j'ai cependant conservé un rôle dans les équipes pédagogiques du mastère informatique spécialités IPS (Informatique Pour les Sciences), et DECOL (Données, Connaissances, Objets et Langues naturelles) et du mastère spécialité BCD (Biologie, Connaissances, Données). J'interviens également dans la réflexion menée au sein du cycle de licence autour des bases de données et des systèmes d'information (thème BD-SI)

3.0.9 Responsabilités de modules d'enseignement

Je retrace essentiellement ici les responsabilités sur les cinq dernières années écoulées. Ces modules d'enseignement sont pour leur majorité proposés en master, même si je m'efforce actuellement d'enseigner également au niveau licence. Les compétences mobilisées se concentrent sur l'enseignement des bases de données et des systèmes d'information, du web de données, de la programmation objet et de la bioinformatique. Le nombre d'étudiants concernés est indiqué entre parenthèses avec l'intitulé et le code du module (également entre parenthèses).

Public	Intitulé - Code	ECTS	Année
L1 Biologie	Modélisation des données biologiques (GLIN103) (120)	2.5	2011-2013
M2 IPS	Administration BD (GMIN308) (20)	5	depuis 2008
	Programmation avancée (FMIN362) (20 étudiants)	5	depuis 2010
	Technologies du web avancées (FMIN358) (40)	5	depuis 2008
M2 DECOL	Gestion de données complexes (GMIN332) (20)	5	depuis 2012
M1 IPS	Systèmes d'information - BD (FMIN111) (100)	5	depuis 2008
M1 BCD	Information biologique (GMIN206) (15)	5	depuis 2011

3.0.10 Supervision de travaux encadrés

L'apprentissage par la pratique demeure prépondérant dans le processus de capitalisation des compétences orienté vers les étudiants. Plusieurs modules sont proposés à cet effet de manière à ce que les étudiants puissent mettre en œuvre leurs savoirs au travers de stages d'analyse, de projets tutorés, ou encore de stages de fin de cursus. Je participe à ces activités par le biais d'encadrements réguliers au sein des modules :

Public	Intitulé - Code	ECTS	Année
M1 Physique-Informatique	Projet tuteuré (FMPH211)	10	depuis 2009
M2 IPS	Stage industriel ou recherche (GMIN403)	25	depuis 2009
M1 IPS	Stage analyse (GMIN10C)	5	depuis 2009
	Projet tutoré (GMIN20E)	15	depuis 2009
L3 Informatique	Projet (GLIN601)	5	depuis 2012

3.0.11 Implication dans des modules d'enseignement à l'interface avec la biologie

J'interviens également dans des modules dont je n'assure pas la responsabilité. Ces modules sont bien souvent orientés vers de la bioinformatique.

Public	Intitulé - Code	ECTS	Année
L3 Informatique	Algorithmique du texte (GLIN608) (20)	5	depuis 2011
M2 Biologie Fonctionnelle des Plantes	Informatique et mathématique des approches massives (GMBP303) (20)	2.5	depuis 2011
L3 BBB, Microbiologie, BMC	Technologie de l'ADN recombinant (GLBP604) (130)	5	depuis 2008

4 Autres implications collectives

Charges administratives

- Membre élu du conseil d'administration de l'Université Montpellier 2, Octobre 2005- Avril 2012
- Membre de la commission de discipline Université Montpellier 2, 2009-2012
- Membre suppléant de la commission des spécialistes section 27 de l'Université Montpellier 2, 2005-2008
- Comité de sélection poste MCF bioinformatique-biostatistiques Université Montpellier I, 2014

Relecture d'articles

- *BMC Bioinformatics*, revue internationale en bioinformatique.
- *Ecological Informatics*, revue internationale en écoinformatiqe.
- *ICEIS*, conférence internationale en informatique.
- *ERSTI INFORSID*, conférence nationale en informatique
- *CARI*, conférence Africaine en informatique et mathématiques appliquées

Participation à des jurys de concours

– Participation à des jurys de concours de recrutement d'ingénieurs (IE et IR) à l'IRD, à l'INRA, au CNRS et en interne à l'UM2

Participation à des jurys d'évaluation

– Participation à la commission d'évaluation des ingénieurs à l'INRA

Participation à des évaluations de projets

– *IDEX Sorbonne Universités*, Projet de type convergence, 2013
– *PEPII BMI*, Biologie-Informatique-Mathématiques, 2013
– *IRD PARRAF*, Agriculture-Pastoralisme, 2013

Implication dans la vie collective

– Sécrétaire de l'Association Sportive Corporatiste de l'Université Montpellier 2 (ASCUM2), depuis l'automne 2012
– Présidente de l'Association de tennis de l'Université Montpellier 2 (ASCUM2 Tennis), depuis l'automne 2010

5 Liste des publications

Thèse.

– Manuscrit de doctorat de 3ème cycle, décembre 1995, *LIGM-DB un système de gestion intégré des séquences nucléiques des récepteurs d'antigènes et de leurs annotations* `www.lirmm.fr/~mougenot/Bibliographie`

Revues internationales avec comité de lecture

– [63] **ThesauForm - Traits : A web based collaborative tool to develop a thesaurus for plant functional diversity research**, Marie-Angélique Laporte, **Isabelle Mougenot** and Eric Garnier - Ecological Informatics, Vol 11, pp 34-44, 2012.
– [78] **A Framework to Assist Environmental Information Processing**, Yuan Lin, Christelle Pierkot, **Isabelle Mougenot**, Jean-Christophe Desconnets and Thérèse Libourel - Enterprise Information Systems, Lecture Notes in Business Information Processing, pp 76-89, Ed. Filipe, Joaquim and Cordeiro, José, Springer Berlin Heidelberg, 2011.
– [45] **PenBase, the shrimp antimicrobial peptide penaeidin database : sequence-based classification and recommended nomenclature**, Yannick Gueguen, Julien Garnier, Lorenne Robert, Marie-Paule Lefranc, **Isabelle Mougenot**, Julien de Lorgeril, Michael Janech, Paul S. Gross, Gregory W. Warr, Brandon Cuthbertson, Margherita A. Barracco, Philippe Bulet, André Aumelas, Yinshan Yang, Dong Bo, Jianhai Xiang, Anchalee Tassanakajon, David Piquemal and Evelyne Bachère - Developmental & Comparative Immunology, Vol 30, pp 283-288, 2006.
– [68] **LIGM-DB/IMGT : an integrated database of Ig and TcR, part of the immunogenetics database.**, Marie-Paule Lefranc, Véronique Giudicelli, Chantal Busin, Ansar Malik, **Isabelle Mougenot**, Patrice Déhais and Denys Chaume - Annals of the New York Academy of Sciences, Vol 764, pp 47-49, 1995.

Conférences et workshops internationaux avec comité de lecture

– [30] **Application Profile for Earth Observation Images**, Jean-Christophe Desconnets, Hatim Chahdi, **Isabelle Mougenot**, Metadata and Semantics Research - 8th Research Conference MTSR 2014 : 68-82, 2014.
– [66] **A semantic web faceted search system for facilitating building of biodiversity and ecosystems services**, Marie-Angélique Laporte, **Isabelle Mougenot**, Eric Garnier, Ulrike Stahl, Jens Kattge and Lutz Maicher, 10th International Conference on Data Integration in the Life Sciences, DILS 2014 : 50-57, 2014.
– [76] **Method and components for creating scientific workflow**, Yuan Lin, **Isabelle Mougenot**, Thérèse Libourel, ICDE Workshops 2014 : 147-153
– [62] **A Faceted Search System for Facilitating the Understanding and the Prediction of Ecosystem Changes**, Marie-Angélique Laporte, Eric Garnier and

Isabelle Mougenot, TDWG biodiversity information standards 2013, Firenze, Italy, October 2013.

- [61] **A Faceted Search System for Facilitating Discovery-driven Scientific Activities : A Use Case from Functional Ecology**, Marie-Angélique Laporte, Eric Garnier and **Isabelle Mougenot**, The Semantic Web : ESWC 2013 Satellite Events, 1st International Workshop on Semantics for Biodiversity (S4BioDiv'13, May 26-30, Montpellier France, 2013. http ://ceur-ws.org/Vol-979/
- [2] **Ontologies Contribution to link thematic and remote sensing knowledge : preliminary discussions**, Samuel Andrés, Damien Arvor, Laurent Durieux, Marie-Angélique Laporte, Thérèse Libourel, **Isabelle Mougenot** and Christelle Pierkot : In proceeding of : Selper 2012, Cayenne, French Guiana, November 2012.
- [71] **An organizational environment for "in silico" experiments in molecular biology**, Yuan Lin, Marie-Angélique Laporte, Lucile Soler, **Isabelle Mougenot** and Thérèse Libourel, ESWC, 4th International workshop on Resource Discovery (RED'2011), Heraklion, Greece, 2011. http ://ceur-ws.org/Vol-737/paper4.pdf
- [74] **Approach for Verifying Workflow Validity**, Yuan Lin, Thérèse Libourel, **Isabelle Mougenot**, Runtong Zhang and Rongqian Ni, ICEIS (3) 2011, Beijing, China, pp 66-75, 2011.
- [77] **A Framework to Assist Environmental Information Processing**, Yuan Lin, Christelle Pierkot, **Isabelle Mougenot**, Jean-Christophe Desconnets and Thérèse Libourel, ICEIS 2010,Funjal, Madeira,pp 76-89, 2010.
- [70] **A Platform Dedicated to Share and Mutualize Environmental Applications**, Thérèse Libourel, Yuan Lin, **Isabelle Mougenot** , Christelle Pierkot and Jean-Christophe Desconnets, ICEIS 2010, Funjal, Madeira,pp 50-57, 2010.
- [72] **A Workflow Language for the Experimental Sciences**, Yuan Lin, Thérèse Libourel and **Isabelle Mougenot**, ICEIS 2009, Milan, Italy, pp 372-37, 2009.
- [67] **Integration of Data Sources for Plant Genomics**, Pierre Larmande, Christine Tranchant-Dubreuil, Laetitia Regnier, **Isabelle Mougenot** and Thérèse Libourel, ICEIS 2006, Paphos, Cyprus, pp 314-318, 2006.
- [69] **MKBEEM : Ontology Domain Modeling Support for Multi-lingual services in E-Commerce**, Alain Léger, Géraldine Arbaut, Peter Barrett, Sylvain Gitton, Asuncion Gomez-Pérez, René Holm, Aarno Lehtola, **Isabelle Mougenot**, Ana Nistal, Theodora Varvarigou and Jérôme Vinesse, 14th European Conference on Artificial Intelligence (ECAI'00), Workshop on Applications of Ontologies and Problem-Solving Methods, Berlin, Germany, pp 19.1-19.4, August 20-25, 2000.
- [85] **Genetic Sequence Annotation within Biological Databases**, **Isabelle Mougenot**, Thérèse Libourel and Patrice Déhais, Proceedings of the 4th International Conference on Database Systems for Advanced Applications (DASFAA), Singapore, April 11-13, pp 333-341, 1995.
- [28] **An Interactive System for Database in Immunogenetics**, Patrice Déhais and **Isabelle Mougenot**, International Conference on System Sciences (HICSS), Hawai, pp 25-34, 1994.

Conférences et ateliers nationaux avec comité de lecture

- [32] **Construction d'un dictionnaire multilingue de biodiversité à partir de dires d'experts**, Mamadou Dieye, Mohamed Rafik Doulache , Mustapha Floussi, Julie Chabalier, **Isabelle Mougenot** and Mathieu Roche, INFORSID'12, Montpellier, pp 81-88, 2012.
- [75] **Environnement de workflow scientifique Validation et conformités**, Yuan Lin, **Isabelle Mougenot** and Thérèse Libourel, INFORSID'11, Lille, pp 129-144, 2011.
- [59] **Construction collective de standards de données en écologie**, Marie-Angélique Laporte, Eric Garnier and **Isabelle Mougenot**, Journées Francophones sur les Ontologies (JFO 2011), Montréal, Canada, 2011.
- [73] **Autour des chaînes de traitements dédiées aux applications environnementales**, Yuan Lin, Thérèse Libourel and **Isabelle Mougenot**, LMO, Langages et Modèles à Objets, Pau, Mars 2010.
- [33] **Regroupement des définitions de sigles biomédicaux**, Ousmane Djanga, Hanine Hamzioui, Mickaël Hatchi, **Isabelle Mougenot** and Mathieu Roche, Extraction et gestion des connaissances (EGC'2009), Strasbourg, 2009.

Conférences internationales et nationales session poster

- [49] **A knowledge-based system for diagnosis dedicated to inherited retinal dystrophies**, Maxime Hébrard, Gaël Manes, Béatrice Bocquet, Isabelle Meunier, **Isabelle Mougenot** and Christian Hamel, JOBIM, Journées Ouvertes en Biologie, Informatique et Mathématiques, Session Poster, 2012.
- [4] **European Drosophila Sequencing Consortium : sequencing the X chromosome of the fly**, M. Ashburner and B. Barrell and P. Benos and V. Bolshakov and A. Bucheton and S. Cox and P. Deak and J. Demaille and C. Ferraz and F. Galibert and D. Glover and D. Harris and H. Jaeckle and F. Kafatos and C. Louis and E. Madueno and J. Modolell and **I. Mougenot** and L. Murphy and G. Papagiannakis and J. Rogers and C. Salles and R. Saunders and C. Savakis and U. Schaefer and I. Siden-Kiamos and L. Spanos and Y. Zhang, ISMB98, Intelligent Systems for Molecular Biology, Poster Session, 1998.

Deuxième partie

Standards de métadonnées et définition de profils d'application Dublin Core

Sommaire

6 Problématique et objectifs visés

Le web est aujourd'hui appréhendé de manière holistique comme un vaste système distribué facilitant l'accès et l'exploitation de ressources numériques locales comme distantes. Il est aussi parfois nécessaire de définir de nouveaux liens entre les ressources de manière à en élargir les usages ou bien de les réutiliser dans de nouveaux contextes. Les dernières évolutions du web avec le web de données[6] s'orientent vers la mise en place d'une architecture multi-couches [114] permettant d'envisager le web comme une base de données unique, à même d'intégrer de multiples sources de données hétérogènes et distribuées. Le web de données est ainsi un outil de médiation, à savoir qu'il est le principal rouage de la transmission du sens et du contenu à partir des ressources liées [101]. Dans ce contexte de médiation, les standards de métadonnées vont jouer un rôle prépondérant dans l'accès, la gestion et la mutualisation des ressources, en offrant les moyens nécessaires à la description de tout type de ressource de manière externalisée. Les standards de métadonnées sont multiples et répondent à des besoins contextuels. Ainsi, il peut s'avérer nécessaire de faire appel à différents standards pour intégrer au mieux le contenu de diverses sources de données. Les profils d'application permettent de documenter voire de prescrire l'exploitation de différents standards de métadonnées dans un contexte applicatif cible et ainsi de construire des services d'accès à l'information à forte valeur ajoutée. Nous prenons des exemples simples dans les sciences du vivant pour démontrer des potentialités des profils d'application. L'objectif est de montrer comment, à partir d'un profil d'application, recontextualiser et exploiter à façon, des données en provenance du portail Entrez [95][7]. Un travail similaire, conduit actuellement au sein de l'équipe de recherche MICADO, se consacre à faciliter les accès aux images satellites et à leurs traitements, dans le contexte de l'Equipex GEOSUD, en mettant un profil d'application au cœur du dispositif [31].
Nous nous attardons, dans un premier temps, sur différentes notions qui sont associées au web de données. Ainsi, nous introduisons les langages RDF [99], RDFS [88] et OWL [9] et nous revenons sur la notion de métadonnée, de standard de métadonnées et de référentiels de valeurs et de contenus. La vision utilitariste des éléments de métadonnées sur le web, permet en particulier de poser une réflexion sur les architectures et modèles qui vont promouvoir une construction et un usage collectif de l'information et du savoir. Les standards de métadonnées sont, à cet effet, exploités en synergie avec les vocabulaires contrôlés, manipulés comme des référentiels de valeurs, pour venir indexer et lier les jeux de données disponibles sur le web. Dans ce contexte, nous abordons dans un deuxième temps, différentes approches méthodologiques autour de la modélisation abstraite et concrète de ressources et de leurs descriptions, que nous avons mis en œuvre dans nos activités de recherche. L'accent est volontairement mis sur la construction de profils d'application qui nous semblent apporter des éléments de réponse en matière d'utilisation de systèmes interopérables sur le web, au regard des besoins des communautés de pratiques et dans une perspective de sources de données ouvertes.

6. Nous nous en tenons seulement au web de données dans ce mémoire et nous n'aborderons pas les autres évolutions du web à l'exemple du web des objets (WOT)

7. Entrez est le portail web, proposé par le NCBI, qui unifie l'accès à une trentaine de banques de données en sciences du vivant et de la santé

7 Langages du web de données

L'idée conductrice du web de données, est de faire évoluer le web conçu au départ comme un vaste système documentaire vers un système de données encore plus vaste (aussi appelé dataspace) et plus adapté aux besoins sans cesse renouvelés des usagers finaux. L'importance est alors donnée aux traitements et notamment aux agents logiciels qui vont permettre l'accès, la consultation et la manipulation de ces données. Les notions essentielles du web, à l'exemple du protocole HTTP (HyperText Transfer Protocol) ou des URI (Uniform Resource Indentifier) sont conservées. Le concept de ressource revêt une importance particulière et va être assimilé à toute chose qui possède une identité pouvant être explicitée au travers d'une adresse unique [16] ou URI [8]. De plus amples informations sur l'architecture du web de données et notamment sur une de ses visions sous forme d'empilement de couches de formats et de standards sont largement disponibles dans la littérature [114, 19].

7.1 Les langages RDF, RDFS et OWL du W3C

7.1.1 Langage RDF

Le langage **RDF** [9], ou Resource Description Framework (Consortium W3C) [99, 100, 38] est un langage de description de ressources à haut niveau et, qui s'appuie, pour ce faire, sur la notion d'annotation. Une annotation est de manière très simple une étiquette qui vient qualifier une ressource d'intérêt ou une partie de la ressource d'intérêt. Une annotation est très souvent assimilée à de la métadonnée. La seule différence en réalité est que la métadonnée a une acception plus contrainte et ne va pas être associée à une ressource si elle n'en qualifie qu'une partie. Pour RDF, l'intuition de départ, est de définir, pour une ressource donnée, toutes les annotations voulues, sans modifier la ressource en question. Dans un second temps, toutes les descriptions accumulées, de manière décentralisée, vont pouvoir être exploitées pour améliorer la découverte, l'accès et la manipulation de ressources d'intérêt à l'échelle du web. De façon plus générale, RDF est devenu aujourd'hui un moyen de décrire des ressources et des relations entre ressources qu'elles soient présentes ou non sur le web.
Un modèle RDF est représenté sous forme de graphe orienté et étiqueté [10], et vu comme une collection de triplets (encore appelés déclarations ou *statements*) {sujet, prédicat, objet}. Le sujet est mis en relation avec un objet (ou valeur du prédicat) au travers d'un prédicat (ou encore propriété). De manière intéressante, le prédicat est aussi une ressource dotée d'une URI et va pouvoir être manipulé comme tel. De manière générique, la déclaration d'un triplet peut s'énoncer comme suit : {U∪B x U x U∪B∪L}. Le sujet du triplet est obligatoirement une ressource étiquetée (U pour l'étiquette URI) ou anonyme (B pour blank), le prédicat est toujours une ressource étiquetée (U donc) et l'objet est soit une ressource étiquetée, soit une ressource anonyme, soit une valeur terminale ou littérale (matérialisée par un L).
Le standard RDF est un format abstrait et différentes syntaxes permettent la concrétisation de modèles RDF. Ces langages sont par exemple RDF/XML, N3[15] (ou Notation

8. Uniform Resource Identifier, aussi désignée par IRI, International Resource Identifier
9. http ://www.w3.org/RDF/
10. Les nœuds comme les arêtes sont étiquetés

3), Turtle (un sous ensemble de N3) [11] ou encore JSON [108].
RDF possède de plus deux particularités qui lui confèrent des avantages indéniables : il s'auto-décrit et il joue le rôle de métamodèle pour les langages de description que nous abordons ensuite et qui vont en proposer des extensions.

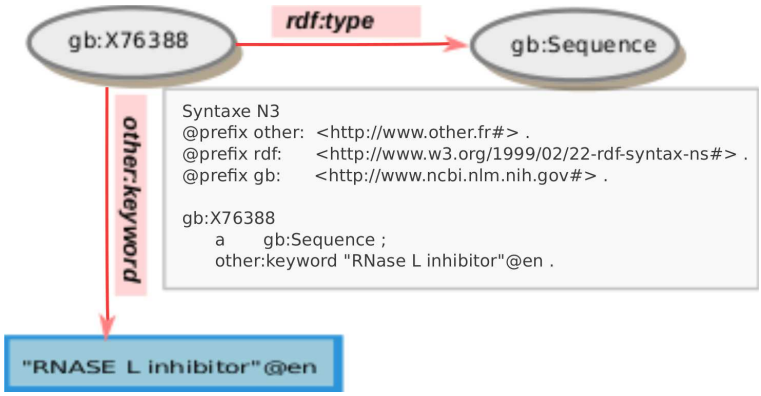

FIGURE 2 – Exemple de modèle RDF

La figure 2 illustre la notion de modèle RDF en biologie avec la description très partielle d'une séquence nucléique de la banque de données GenBank [13, 12]. Le modèle se limite à deux triplets qui possèdent le même sujet, en l'occurrence la ressource `gb:X76388`. Le premier triplet décrit `gb:X76388`, comme une instance de la ressource `gb:Sequence`, en s'appuyant sur le prédicat type du vocabulaire RDF (rdf:type) qui permet de typer des instances. L'objet de ce triplet correspond donc à une valeur non littérale. Le second triplet pointe sur une valeur littérale simple (ou plain literal) complétée par une balise normalisée[11] mentionnant la langue. La valeur ici est une chaîne de caractères en langue anglaise. Le second triplet décrit la ressource `gb:X76388` au travers du mot clé `RNase L inhibitor` qui, comme tous les littéraux, est une feuille terminale du graphe. Trois vocabulaires et donc trois espaces de noms sont exploités dans le modèle : le vocabulaire RDF, et les vocabulaires `gb` et `other` qui ont été créés pour les besoins de l'illustration. La notion d'espace de noms, déjà présente au sein du langage à balisage, XML (eXtended Markup Language), est essentielle à l'interconnexion de vocabulaires et s'avère un préalable à l'intégration de données. Le vocabulaire `gb` s'inspire des descripteurs exploités dans la banque de données généraliste GenBank pour caractériser les séquences nucléiques. Le vocabulaire `other` introduit une propriété `other:keyword` qui permet d'associer le sujet `gb:X76388` au mot clé `RNase L inhibitor`. L'exemple de modèle RDF est également présenté en syntaxe N3 [15]

7.1.2 Langage RDFS

Le langage **RDFS**[12] [17, 88], ou Resource Description Framework Schema, est défini à l'aide de RDF et est orienté vers la construction de modèles (ou encore de schémas). Ainsi, RDFS introduit plusieurs types, à l'exemple de `rdfs:Resource`,

11. Le format d'identification des langages est rfc 3066 Tags for the Identification of Languages
12. www.w3.org/TR/rdf-schema/

`rdfs:Class`, `rdfs:Literal` et `rdfs:Datatype` qui vont faciliter la construction des modèles. Avec le concept de classe (sous-classe de `rdfs:Class`), RDFS donne ainsi la possibilité de distinguer une instance de sa classe d'appartenance. Une classe peut toutefois être aussi une instance, dans la mesure ou elle est une instance d'une métaclasse[13]. RDFS introduit les notions de hiérarchies de subsomption de classes et de propriétés à l'aide des prédicats "sous-classe de" (`rdfs:subClassOf`) et "sous-propriété de" (`rdfs:subPropertyOf`). Les propriétés `rdfs:domain` et `rdfs:range` sont également des ajouts d'importance et permettent de préciser le domaine de définition d'une classe en la caractérisant par une propriété qui va aussi pointer sur une classe cible (ensemble d'arrivée). Les primitives de modélisation définies dans RDFS vont permettre de poser des implications logiques, et de faire émerger par inférence sur la définition des types, des triplets qui étaient jusqu'alors implicites. La figure 3 illustre les éléments de modélisation proposés par RDFS. Ainsi `gb:Sequence` est une classe qui est subsumée par `insd:Sequence`[14]. `gb:X76388` est une instance de `gb:Sequence`, ce qui implique logiquement que `gb:X76388` soit aussi une instance de `insd:Sequence` (transitivité sur la hiérarchie de classes). La propriété `other:hasProv` lie la classe `insd:Sequence` à la classe `insd:Organism` et vient enrichir le domaine de définition de la classe `insd:Sequence` (`other:hasProv rdfs:domain insd:Sequence`) et a pour ensemble d'arrivée la classe `insd:Organism` (`other:hasProv rdfs:range insd:Organism`). La définition d'un domaine d'origine et d'un domaine d'arrivée pour l'association `other:hasProv` va permettre d'interpréter que l'individu `insd:Human` est de type `insd:Organism`, à partir du moment ou le triplet `gb:X76388 other:hasProv insd:Human` est défini au préalable. Enfin `insd:Sequence` possède une autre sous-classe `ena:Sequence` qui est la classe des séquences de la banque de données ENA[15] [23].

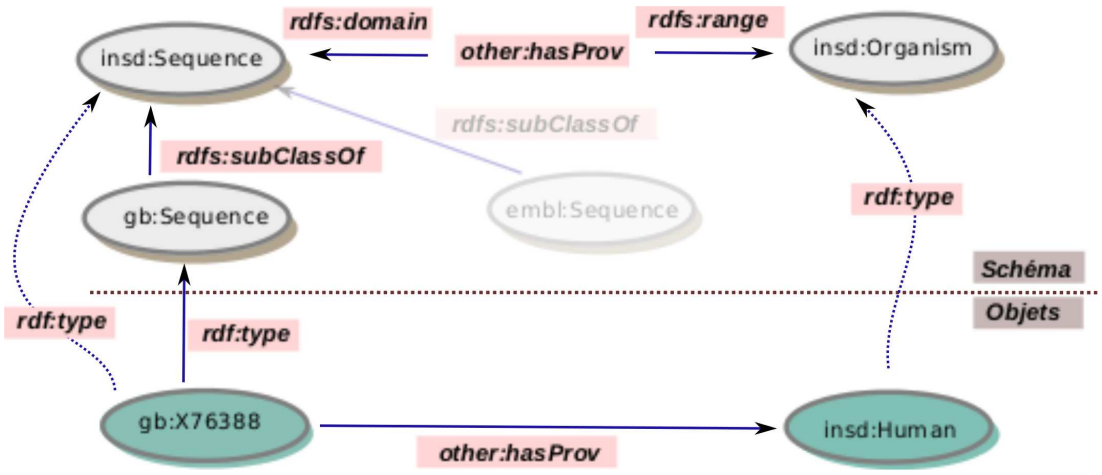

FIGURE 3 – Exemple de modèle RDFS

13. Les activités de métamodélisation dites de *punning* sont également retrouvées dans le langage OWL

14. INSDC ou International Nucleotide Sequence Databases Collaboration est le consortium qui gère l'ensemble des séquences nucléiques au niveau mondial

15. ENA, http ://www.ebi.ac.uk/ena/, European Nucleotide Archive est la banque généraliste de séquences nucléiques en Europe

7.1.3 Langage OWL 2

Le langage **OWL 2** [10, 42], ou Web Ontology Language 2, ajoute de nouvelles primitives de modélisation aux primitives dejà présentes dans RDFS [16] et s'appuie, à cet effet, sur les constructeurs retrouvés dans les logiques de description (opérateurs booléens, classes disjointes, caractéristiques sur les propriétés, restrictions sur les propriétés ...). Le rôle dévolu à OWL 2 est clairement d'être le standard pour la définition de modèles de connaissances, ou ontologies, déployables sur le web. Nous reprenons la définition de [44] pour qui, une ontologie en informatique est une spécification d'une conceptualisation. La spécification exploite un formalisme (à l'exemple d'une logique de description ou d'un profil OWL 2) qui va permettre d'expliciter de manière rigoureuse, le modèle conceptuel décrivant le modèle d'intérêt. Le modèle conceptuel ou conceptualisation est attendu d'être partagé par l'ensemble des experts du domaine considéré. Les ontologies sont devenues des artefacts clés du web alors dit sémantique, en raison de leurs capacités à rendre la connaissance partageable et interprétable. Tout l'enjeu du langage OWL 2 est donc de répondre à ces attentes. Les disciplines et leurs problématiques offrent un éventail de besoins en matière de modélisation de la connaissance. À cet effet, OWL 2 offre trois profils [58] nommés EL, QL et RL pour rendre le partage et l'interprétation de la connaissance les plus fluides possibles dans bon nombre de cas de figures (par exemple, en fonction des constructeurs mis à contribution, du nombre d'individus dans la boîte assertionnelle ou ABox et/ou du nombre de concepts dans la boîte terminologique ou TBox). Nous prenons un exemple très partiel de construction d'une portion d'ontologie. Nous voulons d'abord, exprimer qu'une séquence nucléique du consortium INSD est une séquence qui provient d'une et d'une seule des trois banques de données GenBank, ENA et DDBJ ; puis que cette séquence provient d'au moins un organisme vivant voire de plusieurs s'il s'agit d'une séquence chimérique. Cet exemple est d'abord donné dans le formalisme des logiques de description puis repris dans la syntaxe N3 dans le listing 1 : la classe `insd:Sequence` est une classe OWL et une sous-classe d'une classe anonyme qui correspond à l'union des trois classes `gb:Sequence`, `ddbj:Sequence`, `ena:Sequence` qui sont disjointes deux à deux. De même, tout individu de la classe `insd:Sequence` est associé à au moins un individu de la classe insd:Organism au travers de la relation insd:hasProv et tout individu de la classe `insd:Sequence` ne peut être associé qu'à des individus de la classe insd:Organism au travers de la relation insd:hasProv. L'ensemble de ces expressions sont des conditions nécessaires à la définition d'une séquence INSD mais non suffisantes.

$$\text{insd:Sequence} \subseteq \text{gb:Sequence} \cup \text{ddbj:Sequence} \cup \text{ena:Sequence}$$
$$\text{gb:Sequence} \cap \text{ddbj:Sequence} = \bot$$
$$\text{gb:Sequence} \cap \text{ena:Sequence} = \bot$$
$$\text{ena:Sequence} \cap \text{ddbj:Sequence} = \bot$$
$$\text{insd:Sequence} \subseteq \exists \text{insd:hasProv.insd:Organism} \cap \forall \text{insd:hasProv.insd:Organism}$$

FIGURE 4 – Description partielle de la classe Sequence du consortium INSD en Logique de description

16. http ://www.w3.org/2007/OWL/wiki/RDF-Based_Semantics

```
insd:Sequence a owl:Class ;
  rdfs:subClassOf [
  owl:disjointUnionOf
  ( gb:Sequence
    ena:Sequence
    ddbj:Sequence
  )
] ;
  rdfs:subClassOf
              [ a          owl:Restriction ;
                owl:onProperty insd:hasProv ;
                owl:someValuesFrom insd:Organism
              ] ;

  rdfs:subClassOf
              [ a          owl:Restriction ;
                owl:onProperty insd:hasProv ;
                owl:allValuesFrom insd:Organism
              ] ;
```

Listing 1 – La même description partielle en OWL 2 (syntaxe N3)

8 Autres ressources mobilisables

La définition et la standardisation par le W3C des langages de représentation de contenu RDF, RDFS et OWL ainsi que le rôle central joué par les annotations, ont débouché sur la mise en place progressive de vocabulaires contrôlés partagés qui viennent enrichir le dispositif du web de données. Dans une vision à plus long terme, le succès de cette mise en place passe par une volonté commune et à large échelle, de partager des données en exploitant les langages du W3C et en créant au besoin des extensions à ces langages sous la forme de vocabulaires dédiés. Le point fort du web de données est de permettre à tout un chacun de devenir un contributeur et donc un acteur du web. L'importance est alors de pouvoir mutualiser et intégrer le plus grand nombre de contributions. Le web de données a reçu l'intérêt espéré et de multiples initiatives conduites par de nombreuses communautés de pratique et d'expertise fournissent les standards de métadonnées, les vocabulaires de valeurs et de contenus ainsi que les jeux de données à même de faire évoluer et vivre le web. Ces différentes structures organisant les données, leurs annotations et les valeurs de leurs annotations sont de véritables référentiels réutilisables par tout internaute, et se consacrent à différentes préoccupations tout en restant interopérables. Nous abordons dans cette section, deux grandes catégories de ces référentiels, à savoir les standards de métadonnées et les vocabulaires de valeurs, avant de dresser un panorama général rapide de ce qu'il est convenu d'appeler LOD [48] (pour Linked Open Data ou sources de données ouvertes) et qui est la somme conjuguée de ces référentiels, résultant d'une application collective de bonnes pratiques au sein du web de données.

8.1 Notions de métadonnée et de standard de métadonnées

8.1.1 Définition et typologie des métadonnées

Les métadonnées sont des données sur de la donnée et sont associées à des principes relativement anciens en informatique [6], en particulier, dans les domaines des sytèmes d'exploitation et des bases de données. L'objectif en est d'opérer du contrôle sur les processus et ressources mis en œuvre au sein des systèmes, en venant souvent qualifier des conteneurs de données. Prenons l'exemple des bases de données relationnelles (BDR). Les éléments des schémas de données qui permettent d'organiser les données factuelles, sont à leur tour envisagés comme des données à un niveau d'abstraction supérieur. Ainsi ces éléments sont exploités comme des métadonnées au sein des vues du méta-schéma classiquement appelé dictionnaire de données (voir figure 5). Une telle organisation confère différents avantages ; d'une part elle permet aux administrateurs et gestionnaires de bases de données d'avoir une parfaite connaissance des schémas de données sans avoir à accéder véritablement à ces schémas ; d'autre part elle permet de mettre automatiquement en cache le dictionnaire de données qui va être partagé par l'ensemble des usagers et qui va être exploité par l'optimiseur de requêtes SQL (en particulier dans l'étape de l'analyse syntaxique de la requête) rendant ainsi les mécanismes d'accès aux données plus efficients. Les données du dictionnaire de données sont bien couvent calculées à la volée et le dictionnaire de données est ainsi une collection de vues, à l'exemple des vues Oracle nommées user_tables pour la vue des tables de l'usager ou encore all_constraints pour la vue de toutes les contraintes posées sur la base de données. Nous donnons un exemple de requête SQL qui exploite les éléments du schéma comme de la métadonnée au niveau du méta-schéma. La requête retourne le nom et le type des contraintes ainsi que le nom des attributs sur lesquels elles s'appliquent sur le schéma de la base de données.

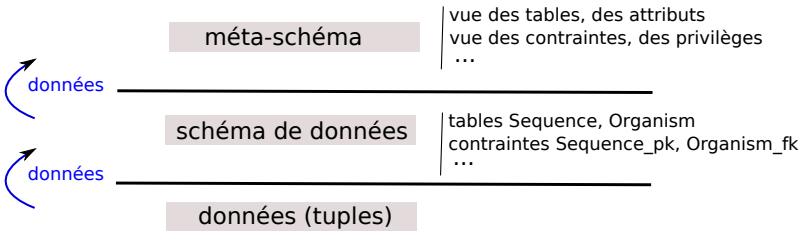

FIGURE 5 – Séparation des niveaux d'organisation au sein des BDR

```
select  uc.constraint_name, uc.constraint_type, cc.column_name
from user_constraints uc, user_cons_columns cc
where uc.constraint_name = cc.constraint_name;
```

Listing 2 – Requête SQL sur le méta-schéma Oracle

Cette constatation amène plusieurs réflexions : le surcoût en matière de stockage des métadonnées est minime, le surcoût en matière d'acquisition des métadonnées est tout aussi minime et enfin les métadonnées considérées sont dites structurelles, en ce sens qu'elles décrivent les éléments structurant les données plutôt que les données elles-mêmes. À cet effet, les métadonnées sont classiquement subdivisées en deux catégories. Les métadonnées dites *structurelles* renseignent les structures présidant à l'organisation des données,

encore appelées conteneurs des données. Les métadonnées dites *descriptives* viennent apporter de l'information supplémentaire aux contenus même des données.
Si l'on se replace dans le contexte du web et de ses dernières évolutions, les métadonnées sont plutôt envisagées comme étant descriptives et viennent enrichir, par de multiples points de vue, les ressources disponibles pour en servir différents usages. Nous reprenons la définition qui en est faite par B. Menon dans [80] et qui met l'accent sur la finalité des métadonnées : une métadonnée est une information associée à une ressource du web permettant d'en favoriser l'utilisation par un agent humain, après exploitation par un agent logiciel. La nature des métadonnées est passée sous silence dans cette définition. Il n'en demeure pas moins différentes typologies organisant ces principales natures pour en faciliter l'exploitation. M.L. Zeng dans [117] passe en revue les principales catégories de métadonnées et leurs usages associés. Nous reprenons dans un premier temps, la typologie de l'association américaine NISO (National Information Standards Organization) [17] [94] qui a l'avantage d'être simple et de bien poser les choses.

1. métadonnées descriptives : la description est bien souvent menée pour pouvoir ensuite faciliter l'identification et l'aide à la découverte des ressources. Les éléments de métadonnées vont par exemple être pour un document, son résumé (abstract) ou des mots clés (keyword) le décrivant au mieux. Les métadonnées peuvent parfois intégrer une dimension sémantique, ainsi l'élément de métadonnées sujet (subject) va puiser ses valeurs parmi les concepts de ressources termino-ontologiques (RTO), facilitant ainsi l'annotation sémantique des ressources.
2. métadonnées structurelles : le propos est légèrement différent de la vision métadonnées structurelles portant sur les conteneurs des données ; ici, les ressources sont parfois des agrégats et leurs sous-parties vont être enrichies par des éléments de métadonnées qui vont en énoncer les relations. Il aurait été peut être plus judicieux de qualifier ces métadonnées de relationnelles.
3. métadonnées administratives : elles recouvrent différentes préoccupations qui sont à la fois techniques, qui visent la préservation à court comme à long terme des ressources (politique d'archivage), ou bien qui permettent d'établir les droits en matière de propriété intellectuelle sur les ressources. Nous retrouvons ainsi parmi ces éléments de métadonnées, le format du fichier associé à la ressource, sa date de création, le propriétaire de la ressource, la licence d'utilisation, ...

A. Gilliland dans [40] propose une autre typologie des métadonnées qui dérive d'une vision générale qui s"applique à tous les objets d'information et qui fait sens dans nos domaines d'application. Les métadonnées sont alors catégorisées selon qu'elles décrivent le contenu, le contexte d'acquisition ou encore la structure de l'objet en question. Nous revenons plus en détail sur ces trois catégories :

1. Contenu : le contenu fait état des caractéristiques inhérentes à l'objet. Les éléments de métadonnées de contenu vont faciliter l'analyse et la compréhension d'un objet.
2. Contexte : le contexte fait état des caractéristiques extrinsèques à l'objet. Ces caractéristiques vont permettre ainsi de faciliter la réponse à des questions de type : Qui, Quand, Quoi, Où et Comment autour des objets d'intérêt. Les éléments de métadonnées de contexte vont faciliter l'aide à la découverte et l'accès aux objets.
3. Structure : la structure décrit les associations qui se nouent soit entre différents objets d'information, soit entre différents éléments du même objet d'information.

17. http ://www.niso.org/home/

Les éléments de métadonnées de structure vont faciliter l'interopérabilité entre sources de données.

8.1.2 Exemple illustratif

Nous prenons, dans le listing 3, l'exemple d'un objet d'information de type fichier textuel de séquence biologique au format GenBank [13] pour illustrer les trois catégories de métadonnées mises en jeu. La figure présentée est un extrait du fichier de la séquence humaine d'ARN messager de numéro d'accès X76388 et qui code la protéine inhibitrice de la RNASE L (Ribonucléase L) [18], désignée par ABCE1 [19] ou RLI [20]. Parmi les enregistrements d'éléments de métadonnées de **contenu**, nous pouvons dégager les informations concernant la définition (H.sapiens mRNA for 2-5A binding protein) et la configuration moléculaire de la séquence (mRNA). Parmi les enregistrements d'éléments de métadonnées de **contexte**, nous pouvons dégager les informations concernant les références bibliographiques avec par exemple le biologiste qui a soumis la séquence (Salehzada,T.) la date de la soumission de la séquence (novembre 1993), et le laboratoire ayant produit la séquence (UMR-CNRS 9942 Montpellier). Enfin, parmi les enregistrements d'éléments de métadonnées de **structure**, nous pouvons dégager l'élément de type clé de caractéristique (ou FEATURE) CDS (CoDing Sequence) qui va décrire la séquence traduite ensuite en séquence peptidique ou encore les informations concernant les relations définies par une étiquette db_xref qui permettent de mettre des éléments de la séquence décrite en relation avec d'autres objets d'information (ici d'autres fichiers) organisés au sein de banques de données biologiques tierces. Ainsi CDS est associé au domaine fonctionnel identifié par IPR001450 dans la banque de données InterPro [86] et la séquence protéique identifiée par P61221 dans la banque de données Swiss-Prot [7]. Dans le fichier présenté, les métadonnées ne sont pas différenciées des données et aucun standard de métadonnées n'est véritablement exploité même si différents efforts sont mis en œuvre au travers de vocabulaires dédiés [36, 5, 20] pour normaliser l'information décrite au sein de la collection des fichiers de séquence.

```
LOCUS       X76388                  3568 bp    mRNA  linear   PRI 12-JUN-2006
DEFINITION H.sapiens mRNA for 2-5A binding protein.
ACCESSION X76388
VERSION  X76388.1  GI:608721
KEYWORDS binding protein; RNase L inhibitor.
SOURCE    Homo sapiens (human)
 [...]
REFERENCE 1 (bases 1 to 3568)
  AUTHORS Bisbal,C., Martinand,C., Silhol,M., Lebleu,B. and Salehzada,T.
  TITLE    Cloning and characterization of a RNAse L inhibitor. A new
           component of the interferon-regulated 2-5A pathway
  JOURNAL J. Biol. Chem. 270 (22), 13308-13317 (1995)
   PUBMED 7539425
REFERENCE 2 (bases 1 to 3568)
  AUTHORS Salehzada,T.
  TITLE    Direct Submission
  JOURNAL Submitted (29-NOV-1993) T. Salehzada, Inst de Genetique Moleculaire
           de Montpellier UMR-CNRS 9942, 1919 Route de Mende BP 5051, 34033
```

18. La ribonucléase L est une enzyme qui clive sélectivement des molécules d'ARN messagers
19. Membre 1 de la famille E de la superfamille des ABC Transporteurs
20. Ribonuclease L Inhibitor

```
            Montpellier, Cedex 01, FRANCE
FEATURES        Location/Qualifiers
     source     1..3568
                /organism="Homo sapiens"
                /mol_type="mRNA"
                /db_xref="taxon:9606"
     CDS        118..1917
                /gene="RNase"
                /codon_start=1
                /product="RNase L inhibitor"
                /protein_id="CAA53972.1"
                /db_xref="GI:987870"
                /db_xref="GOA:P61221"
                /db_xref="HGNC:HGNC:69"
                /db_xref="InterPro:IPR001450"
                /db_xref="InterPro:IPR003439"
                /db_xref="UniProtKB/Swiss-Prot:P61221"
```

Listing 3 – Séquence au format GenBank

L'idée est maintenant d'organiser une partie des informations relevant des trois catégories de métadonnées en s'adossant au standard de métadonnées généraliste Dublin Core[8] et aux référentiels de valeurs Taxonomy et SOFA [21] (sous-vocabulaire de SO, Sequence Ontology [87]). Nous présentons, dans ce sens, une figure 6 qui place la ressource `gb:X76388` au centre des éléments qui viennnent la décrire (disposition visuelle du schéma en étoile ou en hérisson). Les propriétés du standard Dublin Core [22], à l'exemple de `dct:creator` ou de `dct:hasPart`, sont exploitées. Certaines de ces propriétés pointent sur des valeurs normalisées qui sont organisées au sein de référentiels de valeurs. Ainsi `gb:X76388` est qualifiée, au travers de la propriété `dct:subject`, par le terme `so:SO_0000234` du vocabulaire SO (Sequence Ontology) [36], indiquant que cette séquence est une séquence d'ARN messager. Les trois grandes catégories de métadonnées sont représentées sur la figure à l'aide d'un code couleur. Nous reviendrons ultérieurement sur l'intérêt que pourrait avoir l'utilisation de standards de métadonnées pour proposer une structuration appropriée des métadonnées associées aux fichiers de séquences biologiques. Les intérêts directs pourraient en faciliter une remobilisation de l'information pour en proposer de nouveaux questionnements que cela soit d'ordre scientifique (permettant par exemple de mieux appréhender les éléments structuraux et fonctionnels décrits dans les séquences) ou bien d'ordre politique de la science (permettant par exemple d'avoir une vision claire sur quels sont les laboratoires travaillant sur telle ou telle famille génique ou tel ou tel produit d'expression de gène).

21. Sequence Ontology Functional Annotation, http ://obo.cvs.sourceforge.net/viewvc/song/ontology/sofa.obo
22. L'alias pour l'espace de noms du Dublin Core est dct pour Dublin Core Terms

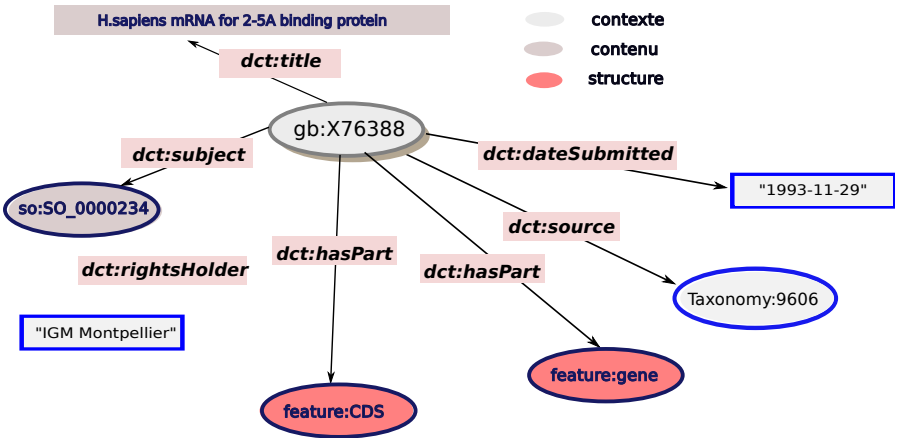

FIGURE 6 – Enrichissement au travers de métadonnées

8.2 Schémas et standards de métadonnées

Les schémas de métadonnées prennent tout leur sens dans un contexte d'accès, d'organisation et de gestion de l'information propice à la diffusion et à la mutualisation de l'information à large échelle. Un schéma de métadonnées est un ensemble systématisé d'éléments de métadonnées et de règles qui s'y appliquent pour en faciliter l'usage [43]. De nombreux schémas de métadonnées sont actuellement disponibles et vont répondre à différents besoins dans différents contextes disciplinaires. Les besoins peuvent par exemple porter sur de la découverte de ressources sur le web, de la consultation, de l'utilisation ou de la recontextualisation de l'information ou encore de la préservation de données sur le long terme. Les schémas de métadonnées font place à des standards de métadonnées dès lors qu'un processus de formalisation s'engage et permet de respecter certaines bonnes pratiques en matière de l'élaboration et du cycle de vie du schéma. Le standard ISO/IEC 11179 a permis notamment de poser des bases unifiant la définition de standards de métadonnées. Un avantage en est en particulier l'enregistrement auprès d'un organisme qui se charge de la diffusion et de la maintenance. Un registre de métadonnées en découle naturellement et est alors accessible à distance et de manière publique. Le standard de métadonnées Dublin Core est dans ce sens rendu disponible au travers d'un registre [23] sous la responsabilité de la communauté DCMI (Dublin Core Metadata Initiative). Le registre propose alors une description de l'organisation de l'ensemble d'éléments de métadonnées dédiés à la prise en charge de ressources web. Les contraintes qui vont s'appliquer à chaque élément sont également décrites. Ainsi les domaines des valeurs autorisées pour les enregistrements d'éléments peuvent être précisés au travers de différentes règles (restriction de valeurs, appartenance à un vocabulaire contrôlé, ...). D'autres standards de métadonnées viennent satisfaire les besoins de différentes communautés d'expertise. Le standard VRA (Virtual Resources Association) [24] fournit les éléments de métadonnées qui permettent de décrire les activités culturelles autour du visuel et les images associées. Le standard EML (Ecological Markup Language) [81, 14] fournit les éléments de métadonnées qui permettent de décrire les jeux d'observations acquises sur le terrain en écologie. Le standard ISO 19115 [1, 53] fournit les éléments de métadonnées qui permettent de décrire les entités géographiques. Nous

23. dcmi.kc.tsukuba.ac.jp/dcregistry/
24. Documentation à l'adresse www.loc.gov/standards/vracore/

revenons dans un premier temps sur le standard généraliste de métadonnées Dublin Core [109, 8] qui demeure le plus connu et qui recouvre en réalité plusieurs standards. Dublin Core a été pensé à l'origine (Dublin Ohio 1995) pour permettre de décrire toute ressource sur le web afin d'en faciliter la découverte. Les domaines de compétence de Dublin Core ont été étendus depuis lors, largement aidés en cela par l'initiative DCMI [113, 97]. Nous retiendrons dans ce sens tous les travaux orientés vers de la médiation de données.
Le premier standard Dublin Core désigné par *Simple Dublin Core*[25] et stabilisé en 1998 est composé de 15 éléments cœur à partir desquels les usagers peuvent rechercher de l'information au sein de sources de données très variées. L'objectif de départ est la simplicité et la capacité d'adaptation à toute problématique. Cependant, un standard de métadonnées ne peut suffire à couvrir l'ensemble des besoins, et un second vocabulaire nommé Dublin Core qualifié ou encore *Qualified Dublin Core*[26] permet d'étendre et qualifier les éléments cœur. Trois éléments ont été ajoutés : dct:audience, dct:provenance et dct:rightsHolder et des extensions de deux natures différentes sont prises en charge :

1. un élément cœur est spécialisé en de nouveaux éléments qui viennent le raffiner, à la manière du principe de l'héritage dans la philosophie des langages à objet. L'élément de métadonnées coverage est par exemple spécialisé par spatial et par temporal qui décrivent respectivement la couverture spatiale et temporelle d'une ressource d'intérêt.

2. les valeurs attendues pour un élément cœur sont restreintes et doivent, par exemple, être prises parmi les termes de vocabulaires contrôlés existants (vocabulary encoding schemes) ou respecter des syntaxes d'encodage normalisées (syntax encoding schemes). Pour ce qui concerne les vocabulaires contrôlés, DCMI propose son propre vocabulaire nommé DCMI Type[27] qui organise différents concepts terminologiques à l'exemple de DCMI Point ou de DCMI Period qui sont particulièrement adaptés pour qualifier respectivement les éléments de métadonnées dct:spatial et dct:temporal.

Le standard Dublin Core qualifié respecte le principe dit du *dumb-down* qui veut que la valeur d'un élément qualifié soit également valide pour ce même élément privé de sa qualification. Ce principe, qui s'apparente à de la rétro-compatibilité, va être précieux pour assurer l'interopérabilité entre les sources de données. Les standards Dublin Core, simple et qualifié, sont pensés pour décrire les ressources en pleine complémentarité et à ce titre sont intégrés depuis 2012 dans le seul et même vocabulaire du Dublin Core qualifié.

25. Espace de noms purl.org/dc/elements/1.1/ et préfixe dc également désigné par Dublin Core Metadata Element Set et préfixe dces
26. Espace de noms purl.org/dc/terms/ et préfixe dcterms ou dct
27. Espace de noms purl.org/dcmitype/ et préfixe dcmitype

LABEL	DÉFINITION	PROPRIÉTÉ
Title ⊡ ⊛	A name given to the resource	dc : title
Description ⊡ ⊛	An account of the resource	dc : description
Coverage ⊡ ⊛	The extent or scope of the content of the resource	dc : coverage
Creator ⊡ ⊛	An entity primarily responsible for making the resource	dc : creator
Rights ⊠	Information about rights held in and over the resource	dc : rights
Date ⊡ ⊛	A point or period of time associated with [...]	dc : relation
Relation ◎	A related resource	dc : relation

FIGURE 7 – Eléments de métadonnées du Dublin Core simple

Le code couleur donne pour pour chaque élément de métadonnées, sa catégorie principale. En rouge, les éléments qui relèvent du contenu, en violet, les éléments qui relèvent du contexte et en bleu, les éléments qui relèvent de la structure. De même, les rôles fonctionnels attribués aux éléments de métadonnées sont désignés par l'iconographie suivante :

1. ⊞ usage de la ressource par les bonnes communautés d'usagers
2. ⊠ utilisation de la ressource
3. ⊡ découverte de ressources
4. ⊛ recherche d'information
5. ◎ mutualisation de ressources

8.2.1 Fonctions attendues

Les fonctions attribuées aux métadonnées dans le contexte du web vont être diverses et variées. Nous revenons rapidement sur les principales fonctions qui se concentrent autour de la découverte, de l'accès, de la mutualisation et de l'intégration de multiples ressources hétérogènes et distribuées. En détail, les métadonnées jouent un rôle facilitateur dans :

- **la découverte, la localisation et l'identification de ressources** : il est difficile d'avoir une parfaite connaissance de toutes les ressources existantes voire accessibles, les métadonnées descriptives auront à charge d'enrichir et d'identifier au mieux les ressources permettant ainsi de les cataloguer à bon escient.
- **l'organisation, l'archivage et la préservation** : il est attendu de fournir des mécanismes soucieux de la préservation des ressources sur le long terme, ainsi que du respect des droits associés à la propriété intellectuelle. Les métadonnées administratives et structurelles sont à même de venir en entrée de ces mécanismes.
- **la diffusion de l'information** : un point d'intérêt porte en particulier sur le caractère non propriétaire des métadonnées [112] alors que les données peuvent être assujetties à un accès contrôlé et limité. Les métadonnées vont permettre ainsi d'aider à la découverte et à la localisation de ressources y compris si ces dernières ne sont pas en accès libre. De même, et nous rejoignons là l'intêrêt de disposer d'un dictionnaire de données au sein d'une instance de base de données relationnelle [28], il ne s'avère pas utile d'accéder à chaque ressource pour en connaître le contenu.

28. Instance = mémoire allouée à la base de données démarrée et processus et processus associés

Il suffit de consulter les métadonnées qui en sont des représentations et qui sont organisées dans leur ensemble au sein d'une structure adaptée. Un tel dispositif va rendre les consultations peu coûteuses et efficaces.

– **l'interopérabilité entre sources de données** : le rôle joué par les métadonnées dans l'échange, le partage et l'intégration de données est particulièrement d'importance. Les sources de données sont alors en capacité d'échanger de l'information qui leur est mutuellement intelligible.

8.2.2 Syntaxes de support

Les métadonnées sont principalement outillées par le biais soit de schémas XML, soit des syntaxes concrètes de RDF, à l'exemple de RDF/XML, N3 et JSON-LD. Le choix entre ces deux grandes orientations pour la gestion et la manipulation des métadonnées n'est pas sans conséquence et offre dans les deux cas, des avantages comme des inconvénients. La granularité de l'organisation ne se situe pas en effet au même niveau. Les schémas XML sont orientés *document* et vont être adaptés à la représentation des collections au sein de ces documents. Un modèle RDF est orienté triplet (ou tuple d'arité 3) ou quadruplet [29] (tuple d'arité 4) quand l'URI du vocabulaire de provenance du triplet est également prise en charge. Un modèle RDF est un graphe orienté construit sur une collection de taille arbitraire de triplets de provenances diverses. Les potentialités à l'interopérabilité sont très fortes, un modèle pouvant être à tout moment agrégé à d'autres modèles ou au contraire séparé en plusieurs modèles ; par contre, l'information est complètement éclatée, et la représentation des collections va être de fait rendue plus difficile. Pour les communautés œuvrant dans les sciences de l'information et de la communication au sens large, le passage des schémas XML à RDF ne relève pas d'une simple formalité. La conversion d'un format à l'autre n'est pas l'élément le plus difficile et des solutions techniques existent au travers notamment des langages XSLT ou GRDDL. La transition des schémas XML vers RDF nécessite par contre de repenser complètement les architectures et d'adopter le tout distribué. De fait, de très nombreux standards sont encore uniquement proposés au format des schémas XML. Il en va ainsi notamment pour des standards comme METS, MODS ou ISO 19115 avec lesquels des jeux de données très volumineux et très détaillés sont structurés. Le standard Dublin Core est lui rendu disponible dans les deux formats, mais le niveau de description rendu possible avec Dublin Core est bien plus simple avec des enregistrements de métadonnées à très peu de niveaux d'imbrication.

```
gb:X76388
        dct:creator  "IGM Montpellier"^^XSD:string ;
        dct:dateSubmitted "1993-11-29"^^XSD:date ;
        dct:hasPart feature:gene  , feature:CDS ;
        dct:source  taxonomy:T016 ;
        dct:subject  obo:SO_0000234 ;
        dct:title  "H. sapiens mRNA for 2-5A binding protein"@en .
```

Listing 4 – Métadonnées du fichier de séquence au format RDF

29. également nommé quad

8.3 Référentiels de valeurs et de contenus

Le web de données cherche à s'affranchir de la vision classique du web par trop centrée sur le document. Le principe est plutôt de faire partager du contenu informationnel, voire du sens, à des agents logiciels en s'appuyant sur la notion de ressource. Comme introduit précédemment, une ressource révèle l'existence de toute chose concrète comme abstraite, à laquelle est attribuée une identité (sous la forme basique d'une URI). Le périmètre du seul document est donc dépassé. À ce titre, un document est une ressource mais est aussi une collection de ressources qui sont susceptibles de s'interconnecter avec d'autres ressources présentes par exemple dans d'autres documents. Nous avons vu que les ressources pouvaient être complétées par des enregistrements de métadonnées venant en faciliter la localisation, l'accès et la consultation.
Nous allons maintenant aborder d'autres modèles qui vont plutôt s'attacher à décrire un univers d'intérêt de manière organisée et consensuelle, et qui vont eux aussi contribuer largement au web de données, en facilitant les usages pouvant s'appliquer sur les ressources disponibles sur le web. Ces modèles s'adossent eux aussi pour leur grande majorité, aux langages de description du W3C et intègrent deux dimensions, souvent à des degrés divers :

- une dimension conceptuelle ou ontologique : l'importance est de décrire les entités existantes dans un domaine de spécialité
- une dimension terminologique : l'importance est alors d'organiser les termes qui désignent les concepts et qui vont faciliter l'indexation des ressources (et donc des documents qui les contiennent)

Ces modèles sont rassemblés sous le vocable ressource termino-ontologique (ou RTO) [111] et comprennent par exemple les taxonomies, les listes de vedettes-matières, les thésaurus et les ontologies. Nous avons pris le parti de parler de référentiels de valeurs et de contenus plutôt que de RTO, pour mieux faire le lien avec les standards de métadonnées qui en exploitent les termes comme étant les valeurs potentielles des éléments de métadonnées.

8.3.1 Référentiels de valeurs

Nous nous contentons ici d'introduire deux formats, nommés SKOS [30] (Simple Knowledge Organisation System) [54] et OBO (Open Biological and Biomedical Ontologies) [31] [104], qui vont proposer des cadres de représentation standards pour la construction de RTO. OBO comme son nom le laisse à penser, est spécialisé dans la construction de modèles liés aux sciences du vivant et de la santé, alors que SKOS est à visée généraliste.

SKOS est un vocabulaire dédié à la construction de RTO, appelés également schémas de concepts ou SOC [32] pour système d'organisation de connaissances. SKOS propose, à cet effet, un cadre normalisé pour la construction de SOC et reprend à son compte bon nombre de travaux autour des normes de conception, gestion et maintenance de thésaurus, ANSI/NISO Z39.19[92], BS 8723 et AFNOR Z 47-100 [22]. SKOS est devenu à l'exemple de RDF, RDFS et OWL sur lesquels il s'appuie, un standard du W3C et est

30. www.w3.org/2004/02/skos/
31. www.obofoundry.org/
32. KOS, Knowledge Organization System en anglais

environné par de multiples développements techniques [56]. De fait, de nombreux référentiels de valeurs [33] sont aujourd'hui disponibles sur le web au format SKOS. Agrovoc [34] [105, 21] et TheSoz [115] en sont des exemples marquants. Dans une première approche, SKOS fait l'amalgame entre descripteur d'un concept (dimension terminologique) et concept (dimension ontologique) mais offre, en contrepartie, quelques flexibilités, pour ce qui touche aux activités de modélisation relatives aux référentiels. Ainsi SKOS offre plusieurs alternatives en terme d'expressivité, avec un premier modèle qui n'exploite que des primitives RDF et RDFS et un second modèle basé sur OWL. De même, SKOS comprend plusieurs modules : un module central nommé SKOS-Core et des modules annexes comme SKOS-XL qui étend la modélisation et chosifie les labels qui désignent les concepts, et qui en conséquence, permet de séparer le concept de ses descripteurs. Autre point d'intérêt, le modèle SKOS est envisageable soit comme un modèle de données (le concepteur d'un nouveau référentiel instancie alors ce modèle), soit comme un métamodèle (le concepteur d'un nouveau référentiel construit alors un modèle de spécialité qui vient se conformer au métamodèle). Pour résumer, SKOS laisse aux concepteurs de SOC la responsabilité de confondre ou bien de séparer dans une certaine mesure, les aspects ontologiques et terminologiques. Des travaux de rétro-conception autour du format SKOS, ont été réalisés par K. Zayrit lors de son stage de master recherche [116]. Le diagramme de classes UML présenté dans la figure 8 reprend en partie ces travaux. Les classes Concept, ConceptScheme et Collection y sont décrites. La classe Element est une classe abstraite qui a été introduite pour factoriser les attributs partagés par l'ensemble des classes SKOS. La classe ConceptScheme correspond de manière élargie à la classe Thesaurus et Concept correspond à la fois à la notion de Descripteur et de Concept. Concept, comme d'ailleurs les autres classes SKOS, est caractérisé par l'attribut prefLabel (label préféré) qui peut s'avérer être le label du descripteur et comprend une collection de altLabel (synonymes) qui peuvent correspondre aux termes non-descripteurs. Les relations hiérarchique et associative sont des associations réflexives de Concept vers Concept. Les concepts les plus génériques dans le schéma de concepts sont modélisés au travers de l'association *topConceptOf*. Enfin, la classe Collection permet d'organiser les concepts au travers de facettes afin de faciliter la navigation dans l'arborescence des concepts d'un vocabulaire.

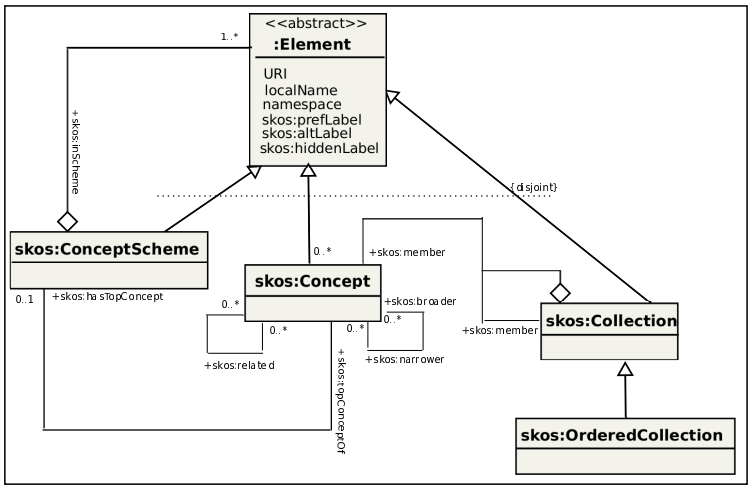

FIGURE 8 – Diagramme de classes autour des éléments principaux de SKOS

33. www.w3.org/2001/sw/wiki/SKOS/Datasets
34. www.fao.org/agrovoc

La définition du concept/terme ARN messager (emprunté à l'ontologie Sequence Ontology) [87] est donnée au format SKOS (syntaxe N3) en guise d'illustration. L'exemple est volontairement très simple. Le concept `obo:SO_0000234` est un individu de la classe `skos:Concept` et possède un label préféré en anglais (mRNA), un synomyme en anglais (messenger RNA) et un label préféré en français (ARNm). Il possède de plus une définition en anglais et est relié à un concept/terme plus général identifié par obo:SO_0000233 (transcrit mature) par le biais de la propriété `skos:broader`

```
obo:SO_0000233
        a           skos:Concept ;
        skos:preLabel "mature_transcript"@en .

obo:SO_0000234
        a           skos:Concept ;
        skos:altLabel  "messenger RNA"@en ;
        skos:broader obo:SO_0000233 ;
         skos:definition   "Messenger RNA is the intermediate molecule between DNA and protein"@en
             ;
        skos:preLabel "ARNm"@fr , "mRNA"@en .
```

Listing 5 – Exemple vocabulaire au format SKOS

OBO Le format OBO est un format d'organisation de référentiels qui s'est très largement répandu depuis le début des années 2000. L'initiative autour de la construction du vocabulaire Gene Ontology (GO) en a été à l'origine et était corrélée à la forte demande de normalisation des termes venant décrire les gènes dans les projets de séquençage de génomes. Gene Ontology permet d'annoter sémantiquement les produits d'expression de gènes au regard de trois intérêts principaux : fonction biologique, localisation cellulaire et processus cellulaire. Depuis lors, des dizaines de vocabulaires sont définis et maintenus au format OBO, et sont rendus disponibles au moyen de registres et de portails (BioPortal, OntoBee et OBOFoundry). Le vocabulaire Sequence Ontology fait partie de ces vocabulaires et permet de normaliser la description des éléments d'intérêt dans le processus d'annotation experte des séquences biologiques. OBO organise les concepts/termes (en particulier au travers de relations hiérarchiques et de synonymie) au moyen d'une structure à base de graphe orienté et acyclique. OBO est proposé dans un format textuel propriétaire également nommé OBO, et dans un second format qui exploite les syntaxes concrètes d'OWL. Le listing 6 reprend au format textuel OBO le même exemple que l'exemple traité dans le listing 5. Le lien de généralisation est ici dénoté par un lien `is_a`. Chacun des vocabulaires OBO est vu comme un module d'une ontologie légère générale venant décrire les sciences du vivant et de la santé. À ce titre, chaque vocabulaire est associé à un espace de noms unique (namespace) qui va faciliter la définition des liens entre des termes provenant de plusieurs de ces vocabulaires.

```
[Term]
id: SO:0000233
name: mature_transcript
namespace: sequence
def: "A transcript which has undergone the necessary modifications [...] "
subset: SOFA
is_a: SO:0000673
relationship: derives_from SO:0000185
```

```
[Term]
id: SO:0000234
name: mRNA
namespace: sequence
def: "Messenger RNA is the intermediate molecule between DNA and protein. It includes UTR and
     coding sequences. It does not contain introns."
is_a: SO:0000233
synonym: "messenger RNA"
```

Listing 6 – Exemple SO au format textuel OBO

Les formats SKOS, dans un contexte d'utilisation large, comme OBO, dans un contexte de spécialité plus restreint, apparaissent comme des solutions privilégiant l'interconnexion de référentiels de valeurs et à même de proposer aux communautés de pratique des moyens d'abonder aux sources de données ouvertes.

Bonnes pratiques de construction de vocabulaires Des conseils de bonnes pratiques sont également fournis aux communautés expertes pour la construction et la publication d'un vocabulaire de spécialité au sein du web de données. Ces conseils viennent rendre encore plus efficace l'adoption des formats SKOS et OBO et facilitent l'interconnexion des vocabulaires et surtout l'intégration de jeux de données qui sont annotés avec ces vocabulaires. Nous listons de manière très concise les conseils donnés pour construire un vocabulaire dit 5 étoiles [55] lorsqu'il répond à tous les critères d'exigence posés.

1. le vocabulaire est rendu accessible sur le web : les éléments du vocabulaire sont dits déréférençables (accessibles via leur URI) et le vocabulaire est doté d'une adresse permanente (mécanismes d'indirection liés au système purl [35] (Persistent Uniform Resource Locators))

2. le vocabulaire est rendu "intelligible" pour une machine : il est construit sur les langages du web de données : RDF, RDFS et OWL en particulier

3. le vocabulaire est lié à d'autres vocabulaires : des correspondances sont définies entre les éléments des vocabulaires

4. des métadonnées viennent compléter le vocabulaire en fournissant des informations complémentaires sur le contenu du vocabulaire, sur son contexte de construction ou bien sur sa propension à être lié à d'autres vocabulaires : des vocabulaires à l'exemple de VOAF [36] (Vocabulary of a Friend) ou OMV [37] (Ontology Metadata Vocabulary) viennent en support de cette tâche de qualification

5. le vocabulaire est suffisamment connu et surtout utile pour être référencé par d'autres vocabulaires : il s'agit en quelque sorte d'une justification des investissements consentis par la communauté dans la construction et la publication du vocabulaire considéré.

35. sites.google.com/site/persistenturls/
36. lov.okfn.org/vocommons/voaf/v2.3/
37. omv2.sourceforge.net/

8.3.2 Référentiels de contenus

Les référentiels de contenus ont un objectif légèrement différent des référentiels de valeurs et sont essentiellement à replacer dans le contexte des standards de métadonnées. Il ne s'agit pas seulement d'organiser des termes qui seront mis en complémentarité avec les éléments de métadonnées, mais aussi et surtout d'expliciter les attendus en terme de valeurs pouvant être prises par les éléments de métadonnées. Ces attendus revêtent diverses formes. Ils peuvent s'exprimer au travers de lignes de conduite textuelles sur la manière d'exploiter en complémentarité des éléments provenant de standards de métadonnées et des termes provenant des référentiels de valeurs ou de l'énonciation de règles. Ils peuvent également s'exprimer sous forme de règles portant par exemple sur un élément de métadonnée : l'élément de métadonnée peut être rendu obligatoire à valuer dans le contexte de la description d'une ressource d'intérêt (notion d'obligation), l'élément de métadonnée peut être présent une seule ou plusieurs fois dans la description (notion de répétabilité). Les objectifs avérés sont surtout de satisfaire au mieux des activités d'indexation et de catalogage.
Le standard RDA (Resource Description and Access) [38] [39] [52, 24] est un standard qui facilite la construction de tels référentiels de contenus dans une perspective d'ouverture sur le web. RDA a été développé pour satisfaire les besoins des bibliothèques en matière de règles de catalogage facilitant la médiation de notices bilibiographiques entre catalogues en ligne. À cet effet, les spcécifications de RDA se basent sur le modèle conceptuel (entité-association) de FRBR (Functional Requirements for Bibliographic Records [40]) [27] qui organise les entités d'intérêt associées aux ressources bibliographiques, à l'exemple des entités [41] `Expression`, `Manifestation` ou `Work`. RDA est cependant pensé de manière à pouvoir être réexploité dans d'autres contextes. Les notions d'intérêt définies au sein de RDA sont par exemple, la notion de rôle (RDA role) qui permet de poser des liens entre entités d'intérêt du modèle FRBR et qui aboutissent à l'ajout de liens faisant sens entre notices ; et la notion de déclarations agrégées (aggregated statements) qui sont des groupes d'éléments de métadonnées pré-coordonnés et qui vont faciliter la réutilisation de notices bibliographiques dans différents contextes. Il est à remarquer que ces deux principes sont également pris en charge par le format SKOS qui propose de poser des liens de correspondance (`skos:semanticMapping`) entre termes/concepts pouvant provenir de plusieurs vocabulaires et qui donne la possibilité de définir des collections ordonnées ou non de termes/concepts (`skos:Collection`) pour les utiliser ensuite comme des facettes de manière à étendre les pratiques mettant en jeu les vocabulaires. À ce titre, des collections SKOS ont été exploitées pour la construction de facettes centrées utilisateur, et privilégient une consultation intégrée portant sur différents SOC en écologie fonctionnelle (travaux de recherche de M.A. Laporte [65])

38. metadataregistry.org/rdabrowse.htm
39. www.rdatoolkit.org/
40. En français : spécifications fonctionnelles des notices bibliographiques
41. www.frbr.org/files/entity-relationships.png

8.4 Panorama général, jeux de données et sources de données ouvertes

8.4.1 Panorama général et sources de données ouvertes

Nous avons abordé dans les sections précédentes les différents artéfacts mis à disposition au sein du web de données pour envisager le web comme constitué d'une seule et unique source de données, à même de répondre à des besoins en matière d'accès, de croisement et de traitenent de l'information sans cesse renouvelés. Nous avons vu que la source de données est en fait une myriade de sources de données interconnectées, que ces sources de données sont des collections de ressources dotées chacune d'une adresse unique et inter-reliées au moyen de déclarations ou de triplets. Nous proposons dans cette section, une série de quatre figures, qui positionnent chacun des types de standards abordés et qui les amènent progressivement à jouer chacun leur rôle en toute complémentarité dans la vision des sources de données ouvertes. Une cinquième et dernière figure est plus volontairement orientée sur les travaux que nous menons actuellement.
Dans la série de figures, les éléments de métadonnées qui sont très majoritairement définis comme des propriétés dans les langages du W3C sont présentés visuellement sous forme de rectangles et les standards de métadonnées sous forme d'arborescences de rectangles. À l'inverse, les concepts des référentiels de valeurs sont très majoritairement organisés sous forme de classes dans les langages du W3C et les référentiels de valeurs sont présentés visuellement sous forme d'arborescences d'ovales.
Dans la figure 9, nous revenons sur un exemple de triplet mettant en jeu, une ressource (ici la séquence biologique exemple) provenant du jeu de données GenBank, un élément de métadonnée (ici `dct:subject`) provenant du standard de métadonnées Dublin Core et un concept (ici `obo:SO_0000234`) provenant du référentiel de valeurs Sequence Ontology. Ce seul exemple montre comment les standards de métadonnées et les référentiels de valeurs s'interfacent pour décrire de la donnée de manière normalisée.

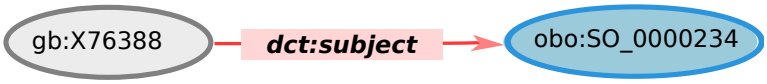

FIGURE 9 – Un triplet en tant qu'élément structurel de base

La figure 10 fait intervenir des éléments de deux standards de métadonnées pour décrire une ressource. L'exemple est générique, les standards de métadonnées et le référentiel de valeurs sont dotés d'espaces de noms arbitraires notés e1, e2 et e3. Il en va de même pour l'identification des ressources et de la valeur du littéral. L'idée à retenir ici est qu'il est possible de faire appel à autant de standards de métadonnées que nécessaires pour décrire au mieux une ressource provenant d'un jeu de données. Une mise en complémentarité des vocabulaires est ainsi attendue de produire des descriptions de la ressource à forte valeur ajoutée.

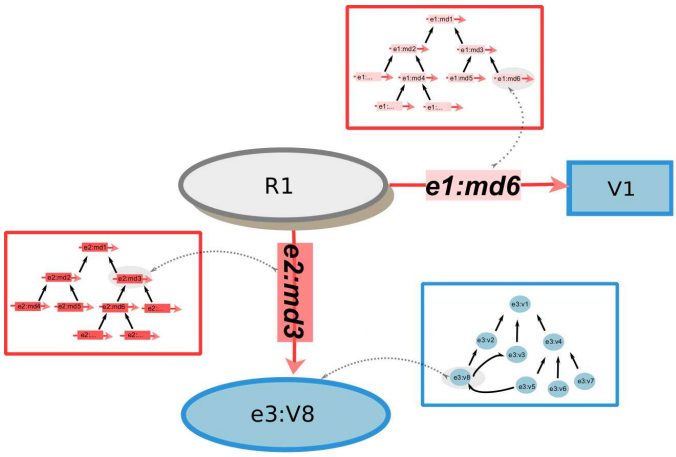

FIGURE 10 – Enrichissement au travers de métadonnées

La figure 11, qui est une extension de la figure 10, met en lumière les potentialités de navigation de terme à terme, ou plus généralement de ressource à ressource, à l'aide de liens de mise en correspondance. Ces liens de mise en correspondance appelés *mapping* ou parfois *crosswalk* lorsqu'il s'agit de liens locaux, sont essentiels à l'intégration des sources de données et à l'émergence de nouveaux savoirs qui se fondent sur la confrontation de multiples pratiques et expertises. Ils font l'objet de nombreux travaux, en particulier sur l'alignement d'ontologies [103] et les mesures de similarité sous-jacentes [46]. La

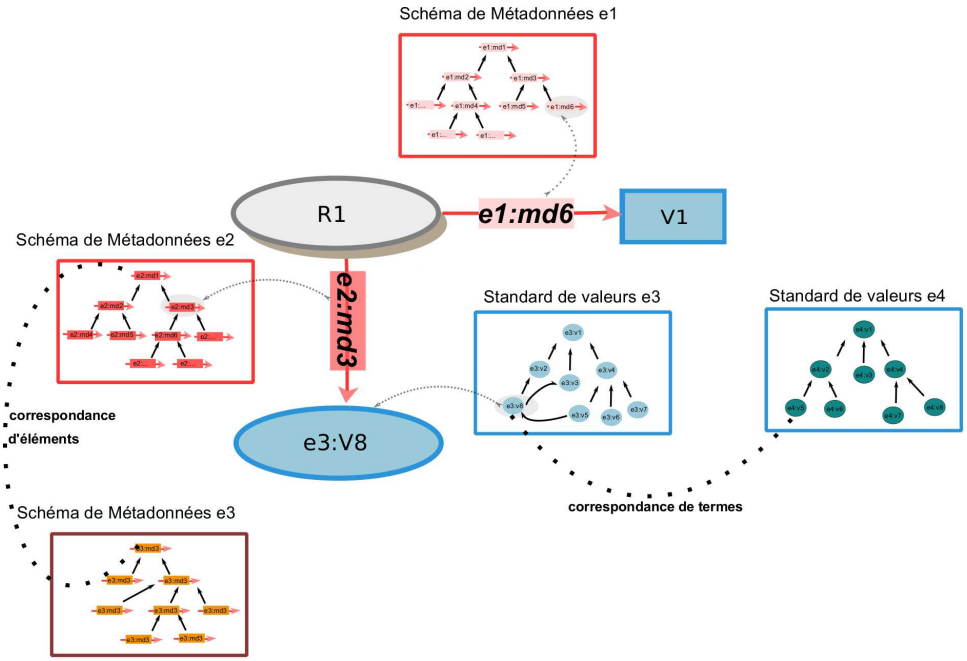

FIGURE 11 – Interconnexion de vocabulaires

figure 12 illustre la mutualisation et la mise en correspondance de sources de données qui résultent de l'application des principes du web de données. Ces sources de données sont alors dites ouvertes et liées (Linked Open Data ou LOD) [48]. Le nombre de ces sources de données augmente de manière continue. Ainsi, le site LODStats [42] fait état

42. http ://stats.lod2.eu/

de 9960 jeux de données ouverts et liés et de 3573 vocabulaires ouverts et liés (Linked Open Vocabularies ou LOV) en mai 2015. Une source de données est dite LOD, à partir du moment où les trois critères suivants sont réunis :

1. taille critique d'au moins mille triplets
2. données liées : au moins 50 liens avec au moins un autre LOD. Des vocabulaires comme VoId [43] (Vocabulary of Interlinked Datasets) permettent de décrire les liens qui se nouent entre jeux de données.
3. données ouvertes : les données sont rendues accessibles et diffusées sous une licence ouverte (licence libre de diffusion) qui en garantit la réutilisation sans restriction, technique, juridique ou financière

Les LOD et LOV s'inscrivent donc dans une vision de partage de l'information à large échelle et sur le long terme.

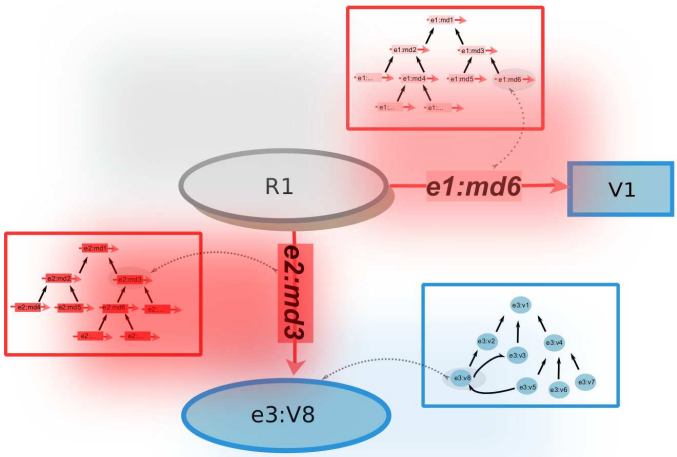

FIGURE 12 – Enrichissement des sources de données liées et ouvertes

8.4.2 Quelques éléments sur les jeux de données

Une organisation normalisée des référentiels de métadonnées, de valeurs comme de contenus, reste un enjeu très accessible à condition que chaque communauté joue le jeu du partage dans une certaine transparence. Il n'en va pas de même pour les jeux de données et ce, pour plusieurs raisons :

- les volumes de données engagés sont bien plus importants et nécessitent des solutions de persistance qui ne sont pas toujours à trouver du côté des formats RDF même si les systèmes de gestion de triplets (ou triplestores) démontrent certaines aptitudes en matière d'efficacité et de robustesse [3].
- l'existant en terme de gestion et de persistance de données est très largement centré sur les systèmes de gestion de bases de données relationnels avec cependant une apparition ces dernières années, des systèmes NoSQL pour ce qui concerne les aspects relatifs aux gros volumes de données distribués [106]. L'idée n'est donc pas

43. http ://www.w3.org/TR/void/

de tout défaire pour tout refaire autrement, mais plutôt de proposer des passerelles pour exploiter au sein du web de données, des données qui sont gérées à l'aide de paradigmes différents, à l'exemple du relationnel, de l'objet ou du semi-structuré. Dans ce sens, des solutions, en particulier le système D2RQ, offrent des mécanismes de transformation, basés sur la définition de correspondance (mapping) entre schémas, pour construire des graphes RDF virtuels.

– le point le plus délicat porte cependant sur les aspects liés à l'ouverture des données et à la propriété intellectuelle. Dans les sciences expérimentales par exemple, les données sont au cœur du savoir-faire et de l'expertise des équipes de recherche. Elles correspondent à la matière première qui est ensuite valorisée au travers de publications. Il est donc naturel de trouver les bonnes manières de protéger les producteurs de données et éviter ainsi tout ce qui peut s'apparenter à du pillage de savoir.

Le web de données a cependant le mérite d'amener les communautés à se poser les bonnes questions. Beaucoup de scientifiques en sont encore à conserver leurs précieux résultats expérimentaux au sein de fichiers tabulaires disparates avec une organisation de colonnes arbitraire, en local sur leur ordinateur et sans aucune politique de sauvegarde ou d'archivage. Un certain nombre de leçons sont à tirer des travaux conduits dans les mondes des bibliothèques, des musées et de l'édition pour tout ce concerne le partage de l'information sur le long terme et la protection du droit de l'auteur.

Nous n'abordons pas dans ce mémoire les aspects liés à la sécurité des données ou encore à la confiance à accorder aux données. Nous n'abordons pas non plus les aspects liés aux droits à la personne et au respect de la vie privée. Il s'agit cependant de sujets qui sont au cœur des préoccupations des acteurs du web. Un rôle panoptique du web de données n'est en effet pas à souhaiter et peut aboutir à des dérives malheureuses. L'idée n'est donc pas de tout ouvrir mais de mieux ouvrir en fonction des besoins et en garantissant la propriété intellectuelle et les droits liés à la personne.

Là encore, des mécanismes ont été mis en place au sein des systèmes de gestion de bases de données relationnelles pour protéger les données. Parmi ces mécanismes, les vues offrent des fonctionnalités d'indirection permettant le contrôle des accès aux données. Ainsi, un usager ne consulte pas directement le schéma d'une base de données mais en interroge seulement les vues qui lui sont accessibles en fonction de ses droits (droits qui sont définis au moyen d'attributions de rôles). La notion de vue est également exploitable en matière d'alimentation d'une table et étend les contraintes assorties à cette alimentation. Enfin les vues sont bien souvent non matérialisées [44] et sont à cet effet des intermédiaires préalables à l'accès et à la consultation des données.

Ces mécanismes de vue peuvent être repensées dans le contexte du web de données. Nous proposons une dernière figure (figure 13) qui va dans ce sens. L'idée est de se contenter de construire des modèles organisant les métadonnées sans chercher à rendre les données disponibles sur le web. Ces données peuvent être rendues accessibles dans un deuxième temps et dans certaines conditions, si cela s'avère nécessaire pour les analyses menées. Nous pouvons faire une analogie dans ce sens avec la nature morte de René Magritte intitulée *la trahison des images* (1929) qui représente une pipe et qui a pour légende "ceci n'est pas une pipe". Tout visiteur de musée y reconnaît une pipe, à ceci près que l'objet "pipe" n'est pas atteignable en tant que tel. Cette même réflexion entre l'objet et sa représentation peut être reprise dans le contexte du web de données. La différence porte,

44. La matérialisation est cependant envisageable pour des raisons de performance dans un contexte par exemple de bases de données distribuées

par contre, sur les objectifs avoués de la représentation. Dans *la trahison des images*, la représentation est orientée vers l'identification de l'objet au travers de ses caractéristiques. Dans le contexte du web de données, l'importance n'est pas de représenter pour identifier [45] mais de représenter pour mieux utiliser. La vision de la métadonnée est, en effet, celle véhiculée par K. Coyle [46], à savoir une information construite, définie pour répondre à des besoins ou à des finalités précis. En d'autres termes, l'objet n'est pas représenté au plus près de ce qu'il est, mais au plus près de ce à quoi il va servir. Les métadonnées sont donc déterminées en fonction de ces visées utilitaristes. Nous en venons à la dernière figure 13 de la série. Cette figure illustre un découplage entre l'objet considéré et la modélisation ou les modélisations qui vont en être faites. Ce n'est plus l'objet qui est modélisé mais un moule ou "template" de cet objet qui en est une sorte de vue fonctionnelle. En fonction de la diversité des besoins informationnels, divers "templates" peuvent être construits pour rendre compte de différents points de vue sur l'objet à partir de templates portant sur des éléments de métadonnées et sur leurs valeurs possibles. Dans l'absolu, la valeur peut être un type de valeur (ou un type de donnée [47]) attendu pour être mis en correspondance avec l'élément de métadonnée. Les référentiels de valeurs sont alors exploités comme des fournisseurs de types qui viennent restreindre le champ des valeurs possibles pour la description d'un élément de métadonnées en particulier et donc d'une caractéristique de la ressource considérée. Cette vision est la vision mise en œuvre dans les profils d'application qui sont au cœur de nos travaux de recherche actuels.

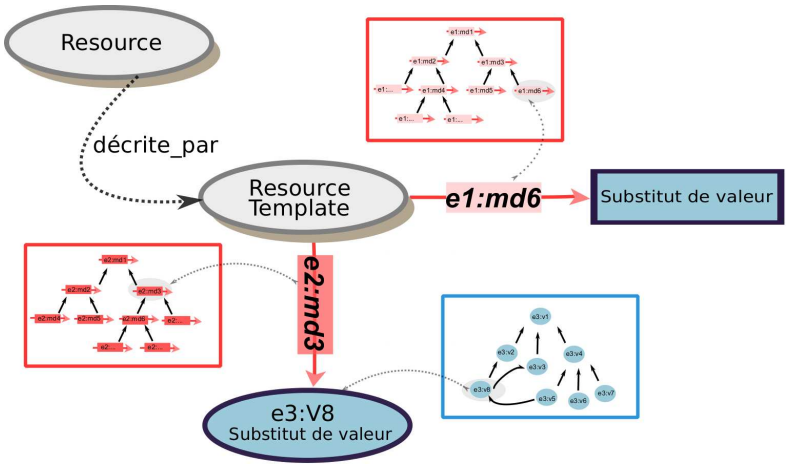

FIGURE 13 – Déporter le modèle sur les éléments de métadonnées

45. L'identification, du reste, est assurée par la définition d'une URI obéissant plus à des conventions de nommage qu'à des considérations sémantiques
46. http ://www.kcoyle.net/meta_purpose.html
47. datatype

9 Profil d'application et derniers travaux menés

9.1 Généralités autour des profils d'application

Les standards de métadonnées jouent des rôles essentiels facilitant la gestion de ressources distribuées au sein d'architectures décentralisées. L'exemple du web montre tout le potentiel de standards tel que Dublin Core ou ISO 19115 dans un contexte applicatif à large échelle. Les standards de métadonnées ont cependant été pensés pour satisfaire des besoins particuliers. De fait, les standards de métadonnées ont été définis indépendamment les uns des autres et ne peuvent pas, seuls, répondre à tous les besoins d'une communauté d'expertise en matière d'usage de l'information. De plus, les standards de métadonnées n'ont pas pour vocation première de contraindre les descriptions qui viennent enrichir une ressource. L'objectif est plutôt de laisser une grande liberté, dans la manière d'exploiter les éléments du standard. Le standard Dublin Core non qualifié en est l'exemple le plus parlant. L'idée initiale était, en effet, de permettre à tout internaute, de décrire tout type de ressource numérique, au travers de quinze éléments très simples à l'exemple de son titre (dc:title) ou de son auteur (dc:author). Cette initiative est en rupture avec les pratiques habituelles des sciences de l'information et de la documentation qui réservent aux seuls documentalistes la mise en œuvre de standards très complets, à l'exemple de MARC [110] . Cette "démocratisation" [48] du standard a été dès lors largement adoptée sur le web.

Il peut s'avérer parfois utile d'exploiter simultanément plusieurs standards de métadonnées et/ou de chercher à restreindre le périmètre d'utilisation des éléments de métadonnées appartenant à ces standards pour répondre à des besoins applicatifs précis. À cet effet, les profils d'application (Application Profile ou AP) [89, 79] réutilisent les standards de métadonnées pour les amener à répondre soit à de nouveaux besoins, soit à des besoins plus ciblés. Il s'agit par exemple de combiner de l'information provenant de différentes sources pour en livrer de nouvelles interprétations ou bien de poser différents filtres sur de l'information pour la recontextualiser. Le principe est alors d'ouvrir les champs du possible en empruntant divers éléments de métadonnées à différents standards et en les articulant (approche dite de "mix et match" [50]) pour en produire une nouvelle organisation d'éléments de métadonnées, particulièrement adaptée à une visée applicative cible. Le principe est donc de réutiliser l'existant au travers d'une nouvelle déclinaison. La figure 14 donne un aperçu simplificateur des pratiques d'intérêt dans la construction d'un profil d'application, qui portent en particulier sur la combinaison et la mise en correspondance d'éléments, ainsi que sur la définition de contraintes sur un élément de métadonnées.

Un profil est un ensemble d'éléments de métadonnées qui ont diverses origines et qui peuvent être mis en correspondance. Dans la figure, les éléments des standards de métadonnées Dublin Core, ISO 19115 et Darwin Core sont exploités. De même, le type de la valeur cible de l'élément de métadonnée peut être précisé de différentes manières. Dans l'exemple proposé, une relation de mise en correspondance (mapping) est posée entre l'élément dct:spatial de Dublin Core et l'élément MD_DataIdentification.extent d'ISO 19115. En outre, la valeur attendue pour un élément dct:spatial est obligatoirement une ressource de type dcmi:Box.

La construction d'un profil d'application répond aux besoins de découverte, de carac-

48. parfois décriée d'où la maxime "worse is better" de Richard Gabriel

térisation et de consultation de ressources distribuées et hétérogènes pour satisfaire les besoins applicatifs d'une communauté experte spécifique. En tant que tel, le profil d'application est soumis à des attentes qui peuvent se révéler contradictoires avec la nécessité de répondre à des besoins précis mais tout en jouant la carte de l'interopérabilité. Un compromis est alors à trouver, les concepteurs d'un profil d'application doivent répondre aux attentes d'une communauté dédiée en ne s'appuyant que sur des standards existants. Leurs seuls leviers se résument alors à définir la meilleure articulation possible entre plusieurs standards de métadonnées et à poser les contraintes sur les éléments de ces standards à même de prendre en charge la spécificité du domaine couvert. Les

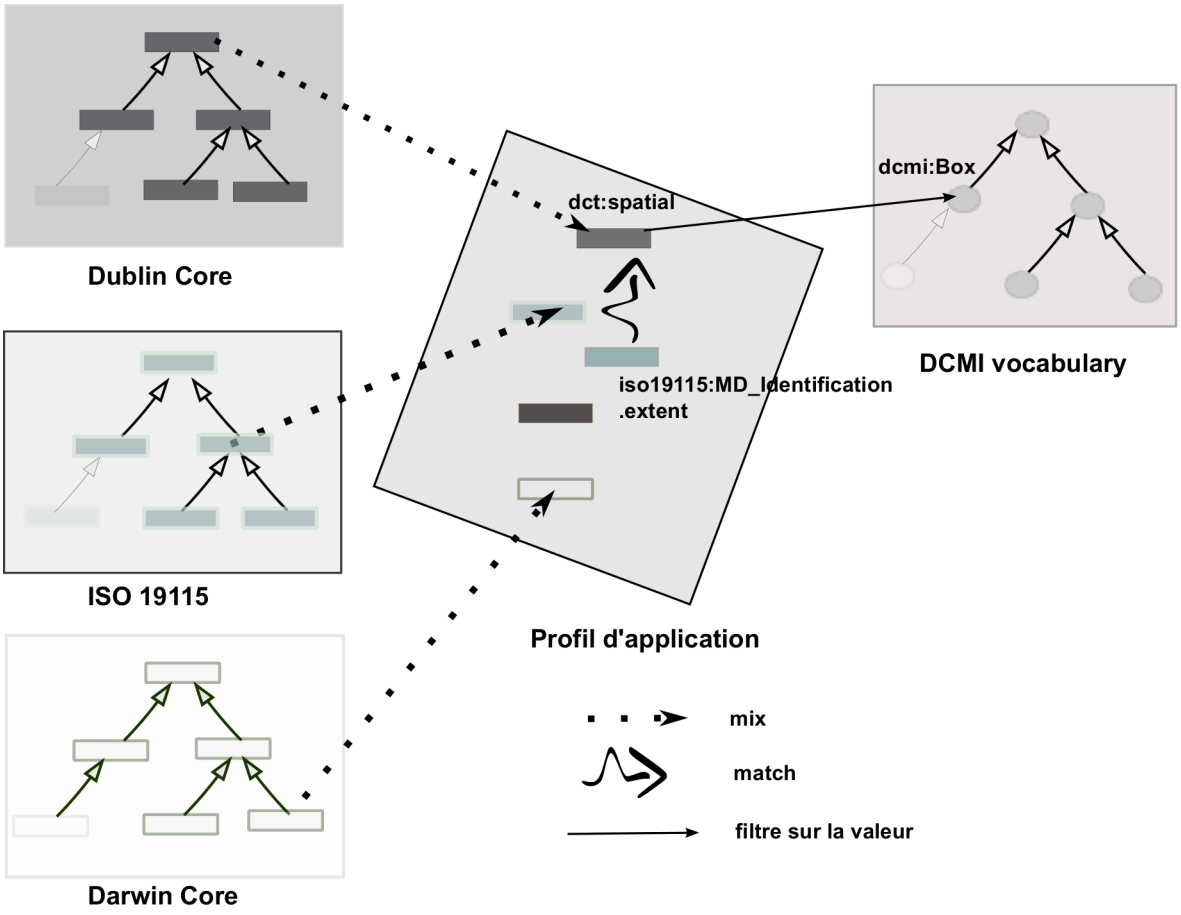

FIGURE 14 – Une illustration simple de l'approche *mix & match*

exemples de profils d'application les plus marquants sont à regarder du côté des sciences de l'information et de la documentation, de l'éducation [41], des sciences humaines ou bien des sciences naturalistes. Cet ancrage n'a rien d'étonnant puisque D. Hillman [51] définit un profil d'application comme un document ou un ensemble de documents [49] qui décrit un consensus posé sur un modèle de domaine. L'apport majeur des profils d'application est de renseigner sur les perceptions et les intentions d'une communauté d'expertise en matière d'usage sur les ressources qu'elle a l'habitude de manipuler. Ces perceptions comme ces intentions vont être livrées au travers d'enregistrements d'éléments de métadonnées et s'avèrent être une valeur ajoutée exploitable en particulier dans un contexte de sources de données interopérables.

49. AP : document of an agreement on a model of how we describe our things in our world in the context of the global web of data

Les disciplines qui manipulent des larges collections de documents à l'exemple de fiches descriptives d'échantillons dans les musées ou de fiches de livres dans les bibliothèques sont plus à même de cerner tout l'intérêt des profils d'application. Il n'en demeure pas moins un rôle applicatif qui peut contribuer à tout domaine d'intérêt, dès lors que des ressources de ce domaine sont gérées de manière distribuée et en masse. Il se révèle alors pertinent de définir un profil d'application à même de tirer parti de l'interopérabilité entre différentes sources de données pouvant être liées. Le profil d'application est envisagé comme un modèle de médiation matérialisé par les enregistrements d'éléments de métadonnées qui sont publiés à partir du contenu des sources de données.

9.2 Approches méthodologiques facilitant la construction d'un profil

Un profil d'application résulte surtout de la mise en pratique de différents principes, et sa construction ne nécessite donc pas de passer par une méthodologie. Il existe toutefois différentes initiatives qui vont en faciliter l'implantation. Ces initiatives restent cependant liées à des contextes applicatifs généraux. Ainsi, une initiative est à mettre à l'actif des sciences documentaires avec les standards METS (Metadata Object Encoding and Transmission Standard) et MODS (Metadata Object Description Schema). METS [25] propose un schéma XML [50] qui rend particulièrement compte de la nature hiérarchique des éléments de métadonnées à même de décrire les objets numériques (microfilm, image, CD audio, ...), les ouvrages présents dans les fonds documentaires voire les specimens et objets de collection présents dans les musées. METS est un format de type conteneur, et permet de fournir les descriptions de différents objets numériques au sein d'un même paquetage lorsque cela est possible. MODS est une version allégée de MARC et fournit un ensemble d'éléments de métadonnées qui viennent renseigner les objets de manière à en faciliter l'accès et la consultation [51]. Des profils d'application peuvent dès lors être construits à partir de METS et de MODS. METS va fournir le cadre général pour organiser et échanger de l'information portant sur un ensemble d'objets numériques sous la forme d'une collection de métadonnées et des éléments MODS vont pouvoir venir s'y greffer pour apporter des enrichissements à visée descriptive. METS et MODS peuvent être complétés par le standard de métadonnées PREMIS [26] qui permet d'enrichir les descriptions faites avec des informations concourant à la préservation des objets numériques sur le long terme (provenance, détails techniques, propriété intellectuelle et droits d'accès, ...). Une autre initiative est à mettre à l'actif des sciences de l'éducation et est associée à LOM (Learning Object Metadata) qui est un standard de description de ressources d'enseignement de d'apprentissage. LOM est également proposé au travers d'un schéma XML [52] et a fait l'objet de différents processus de normalisation dont le dernier en date a abouti à ISO 19788 en 2011. Les profils d'application construits avec LOM sont dits à conformité simple lorsqu'ils contiennent des éléments de métadonnées définis hors du standard LOM. Ils n'ont alors qu'une visée locale. Il peut s'agir par exemple, de supporter les besoins applicatifs en matière de ressources pédagogiques d'une seule structure d'enseignement (une université ou une faculté d'enseignement). Les profils d'application LOM peuvent cependant être construits dans une perspective

50. www.loc.gov/standards/mets/mets.xsd
51. Métadonnées qui sont donc à ranger dans la catégorie descriptive
52. standards.ieee.org/downloads/LOM/lomv1.0/xsd/lom.xsd

d'ouverture et dans ce cadre n'être définis qu'à partir des éléments du LOM. Ils sont alors dits à conformité stricte.
Le listing 7 donne un aperçu de la description en schéma XML, d'un CD audio au travers d'enregistrements de métadonnées provenant de l'instanciation d'un profil d'application METS pour la musique. Cet exemple est extrait de [25] et décrit un enregistrement d'une représentation musicale orchestrée par Léonard Bernstein et contenant des interprétations de mouvements de la symphonie No. 5 de Beethoven. Le standard METS (espace de noms mets) est exploité en tant que structure documentaire principale avec l'élément mets qui joue le rôle de la racine du document. L'élément dmdSec [53] joue le rôle de structure enveloppe pour contenir tous les éléments de métadonnées descriptives du standard MODS (espace de noms MODS). De manière analogue (non illustré ici), l'élément amdSec [54] joue le rôle de structure enveloppe pour des éléments de métadonnées permettant de la gestion et de la pérennisation en exploitant par exemple des standards comme PREMIS. METS rend la description de la structure de la ressource obligatoire au travers de la balise structMap et de ses sous balises à l'exemple de div.

```
<mets:mets>
  <mets:dmdSec>
    <mets:mdWrap>
      <mets:xmlData>
        <mods:mods>
      <mods:titleInfo>
        <mods:title>Bernstein conducts Beethoven and Mozart</mods:title>
      </mods:titleInfo>
      <mods:name> <mods:namePart>Bernstein, Leonard</mods:namePart> </mods:name>
      <mods:relatedItem type="constituent">
        <mods:titleInfo>
          <mods:title>Symphony No. 5</mods:title>
        </mods:titleInfo>
        <mods:name> <mods:namePart>Beethoven, Ludwig van</mods:namePart>
            </mods:name>
        <mods:relatedItem type="constituent">
          <mods:titleInfo> <mods:partName>Allegro con moto</mods:partName>
              </mods:titleInfo>
        </mods:relatedItem>
        <mods:relatedItem type="constituent">
          <mods:titleInfo> <mods:partName>Adagio</mods:partName> </mods:titleInfo>
        </mods:relatedItem>
      </mods:relatedItem>
    </mods:mods>
      </mets:xmlData>
    </mets:mdWrap>
  </mets:dmdSec>
  <mets:fileSec></mets:fileSec>
  <mets:structMap> <mets:div> <mets:div></mets:div> </mets:div> </mets:structMap>
</mets:mets>
```

Listing 7 – Exemple Profil METS (d'après M. Cundiff [25])

La dernière initiative autour des profils d'application est à mettre à l'actif de la communauté Dublin Core et nous allons décrire plus en détail les profils d'application

53. dmdSec est une abréviation de descriptive metadata section
54. amdSec est une abréviation de administrative metadata section

Dublin Core ou DCAP [55] [90] dans les sections qui suivent. Ce sont en effet ces profils qui ont retenu toute notre attention.

9.2.1 Positionnement des profils d'application parmi les référentiels pour le web de données

La figure 15 qui s'inspire largement de [29] apporte un nouvel éclairage sur les liens qui peuvent s'établir entre les différents types de référentiels. Elle positionne notamment les profils d'application face à l'ensemble des référentiels qui vont être des forces de proposition quand à l'interopérabilité entre sources de données et à l'échange et la médiation de ressources. Si l'on suit un sens de lecture de droite à gauche, des modèles très génériques de description de ressources comme le modèle abstrait du Dublin Core (DCAM) ou des modèles entité-relation et ontologique des ressources bibliographiques (FRBRer et FRBRoo) servent de métamodèles à des standards de métadonnées à l'exemple des standards du Dublin Core ou de MARC. Les référentiels de valeurs (tels que Geonames ou AgroVoc) et de contenus (tels que RDA ou AACR2) proposent soit des termes, soit des contraintes qui vont venir qualifier les valeurs attendues pour les éléments de métadonnées et sont dans ce sens associés aux standards de métadonnées. Les référentiels adoptent les mêmes syntaxes support. Les profils d'application ont une position intermédiaire entre les standards de métadonnées et les référentiels de valeurs et de contenus et vont ainsi utiliser à plein les mécanismes rendant les sources de données interopérables. Les profils d'application ne sont que des acteurs indirects de ces mécanismes. C'est parce qu'il les utilisent et qu'ils y contribuent en ouvrant et en liant à leur tour les jeux d'enregistrements de métadonnées construits que les profils sont vus à leur tour comme des solutions à l'interopérabilité.

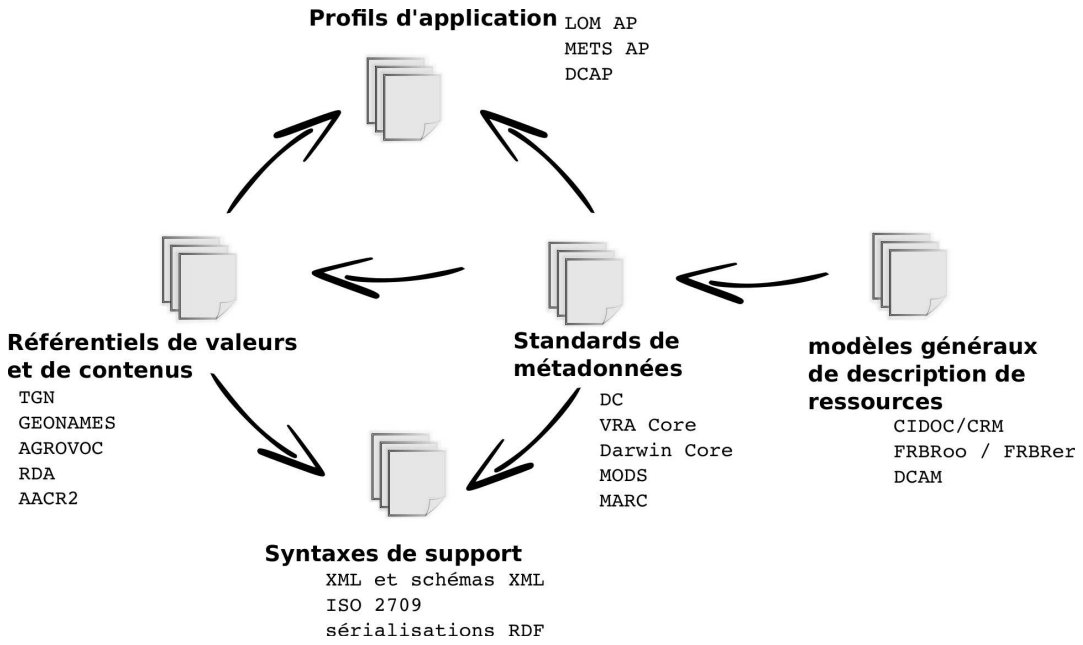

FIGURE 15 – Le positionnement des profils d'application d'après [29]

55. Dublin Core Application Profile

9.3 Principes de construction d'un profil d'application Dublin Core

La définition d'un profil d'application Dublin Core (ou DCAP [56]) est soumise à différentes règles et pratiques qui ont fait l'objet de nombreux travaux et préconisations. Nous ne présentons ici qu'une partie de ces travaux. Un tout premier principe, ciblant l'ouverture et l'interopérabilité des profils, 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 autres principes posés sont plutôt centrés sur la meilleure prise en charge possible des ressources d'intérêt sur lesquelles va s'appliquer le profil. L'idée est alors de se doter d'un modèle conceptuel de domaine qui donne l'organisation des entités qui vont être enrichies par des jeux de métadonnées. La communauté Dublin Core représentée par le DCMI a ainsi défini le framework Singapore [91] pour faciliter la mise en place de profils d'application. Les principaux objectifs du framework Singapore sont de guider les démarches d'enrichissement et d'exploitation des entités du modèle de domaine à partir des éléments de métadonnées provenant des standards, dans des perspectives d'interopérabilité, de flexibilité et donc de réutilisabilité. Le framework

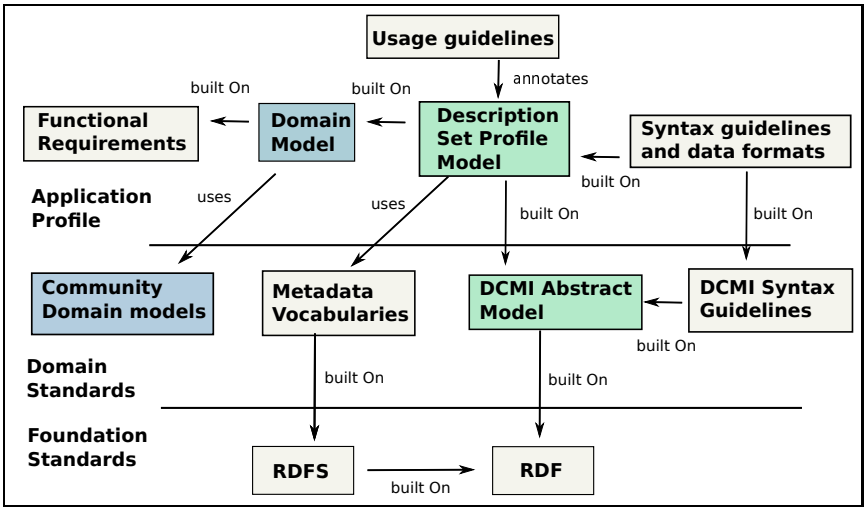

FIGURE 16 – Les modules du framework Singapore et leurs interactions, d'après [91]

Singapore (figure 16) exploite donc les composants habituels du web de données avec les standards de données RDF et RDFS ou encore avec les standards de métadonnées. Il met de plus en valeur deux "pièces maîtresses" dans l'édifice (boîtes vertes dans la figure), qui sont le modèle abstrait Dublin Core (DCAM [57]) et le modèle d'ensemble de descriptions des profils (DSP [58]). Le modèle DSP est défini tout spécialement pour la construction de profils DCAP par le DCMI et va se conformer, à cet effet, au DCAM pour ce qui est de la représentation générale d'une ressource décrite. De plus, un modèle DSP est construit à partir d'un modèle conceptuel de domaine qui peut être un emprunt à des modèles de domaine partagés par des commnautés construits pour répondre aux besoins des communautés de manière large (boîtes bleues dans la figure). Ces modèles

56. Dublin Core Application Profile
57. DCAM pour Dublin Core Abstract Model
58. DSP pour Description Set Profile

conceptuels peuvent être par exemple des ontologies de domaine susceptibles de fournir une connaissance partagée par l'ensemble d'une communauté et sur laquelle des interprétations peuvent s'opérer. Le framework Singapore est construit sur la base de trois niveaux. Le premier niveau est dédié aux syntaxes de support, le deuxième niveau est orienté vers les référentiels qui peuvent être exploités pour le compte d'une thématique d'intérêt et le troisième niveau cible la construction du profil d'application proprement dite. Même s'il n'est fait état que des langages RDF et RDFS pour le premier niveau, il est à remarquer qu'un grand nombre des DCAP en production actuellement s'adossent au langage XML Schema. Nous verrons ultérieurement les raisons de ce choix de langage, cependant moins en phase avec les principes des sources de données liées et ouvertes.
Le framework Singapore est complété par des préconisations d'ordre méthodologique [90] qui viennent faciliter les activités de construction d'un profil d'application. Nous revenons dans l'ordre, sur ces préconisations, sur le modèle DCAM et enfin sur le modèle DSP.

9.3.1 Bonnes pratiques de construction d'un profil DCAP

Un profil d'application se construit progressivement au travers de différentes activités de structuration et de documentation. Ces activités sont soit rendues obligatoires, soit laissées à l'appréciation des concepteurs. Les activités obligatoires comprennent de manière séquentielle :

- les spécifications des exigences fonctionnelles qui posent le périmètre d'utilisation du profil (portée fonctionnelle du profil),
- la définition du modèle de domaine que le profil d'application vient enrichir (portée structurelle du profil). Un diagramme de classes UML est souvent construit à cette occasion. Les classes représentent les ressources à décrire par les éléments de métadonnées présents dans les standards.
- la définition du profil d'ensemble de descriptions (DSP ou Description Set Profile) qui permet de choisir et de contraindre les éléments de métadonnées retenus pour décrire les classes du modèle conceptuel. Un dictionnaire décrivant les éléments de métadonnées retenus et les contraintes qui vont venir s'y appliquer est souvent le mode retenu pour la définition du DSP.

Des activités de documentation annexes peuvent venir compléter de manière optionnelle la spécification du profil d'application pour en faciliter l'exploitation. De même, il peut être utile de fournir certaines préconisations concernant les syntaxes support adaptées dans le contexte du profil.

9.3.2 Modèle de description de ressources DCAM

Un premier modèle structurel nommé DCAM [59] (Dublin Core Abstract Model) [96] (figure 17) explicite la notion de ressource et de sa spécialisation en ressource décrite qui est alors envisagée comme une collection de couples propriété-valeur. La valeur s'envisage parfois aussi comme une ressource et la propriété comme la valeur peuvent être empruntées aux standards de métadonnées et aux vocabulaires contrôlés. La valeur peut également être une valeur terminale ou littérale et associée à un type de données

59. http ://dublincore.org/documents/abstract-model/

primitif (types de données empruntés aux schémas XML). Le modèle DCAM est un modèle très haut niveau et va jouer le rôle de métamodèle pour le modèle DSP. Il est à placer au même niveau que le métamodèle RDF et n'a dans ce cadre qu'une utilité toute relative quand le DSP est défini au travers d'une syntaxe concrète RDF [60].
Nous avons dans des travaux de recherche précédents et déconnectés des travaux de définition de profils, détourné l'utilisation du modèle DCAM pour en faire un modèle d'annotation générique de tout élément structurel d'un vocabulaire au format SKOS. L'objectif était de permettre à une communauté de chercheurs en écologie fonctionnelle travaillant collectivement à l'élaboration d'un thésaurus de converger progressivement vers la définition d'un vocabulaire SKOS qui soit le reflet de toute l'expertise accumulée sur le domaine. La communauté de chercheurs en question est distribuée sur les cinq continents et l'utilisation concertée des modèles DCAM et SKOS a été outillée par le biais de la construction d'un éditeur collaboratif de thésaurus [60, 64].

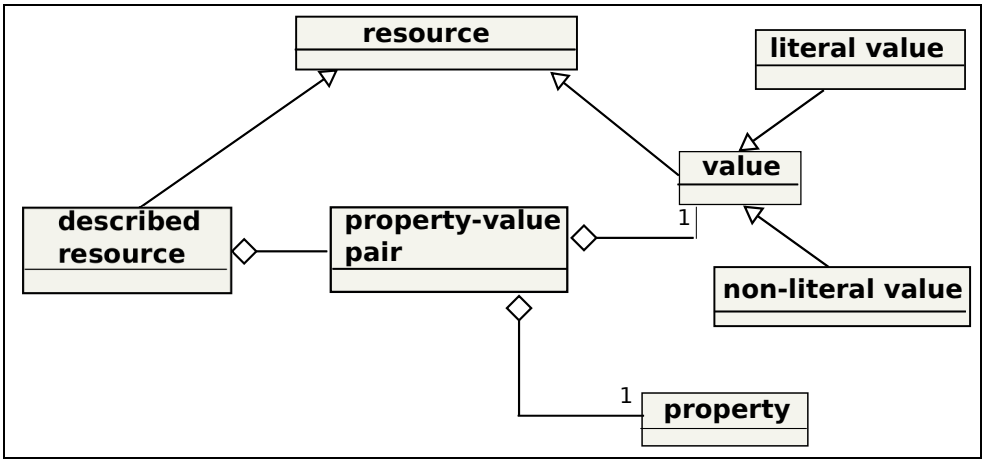

FIGURE 17 – Diagramme de classes UML décrivant le modèle Abstrait de l'initiative Dublin Core (DCAM)

Un second modèle structurel nommé DSP (Description Set Profile) [90] 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 profil d'ensemble de descriptions, et décrit au travers de la notion de `DescriptionSetTemplate`. Chaque description est décrite au travers d'un moule, appelé `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 structurés au sein de déclarations appelées `Statement` et décrites également au travers de moules nommés `StatementTemplate`. Les moules de 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 (`URI_Resource`) provenant de schémas de concepts à l'exemple des vocabulaires 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 19. Ce diagramme de classes UML est très largement inspiré

60. La définition du DCAM est concomitante à la définition de RDF

des diagrammes décrits dans [90, 98]. Le modèle DSP a fait l'objet de spécifications disponibles aux formats XML Schéma et RDF/XML [61]. Les spécifications RDF/XML sont toutefois présentées comme une variante de la version XML Schéma et sous la forme d'un exemple générique de modèle DSP. Des travaux [62] sont en cours pour proposer une version aboutie de vocabulaire RDF pour le modèle DSP.

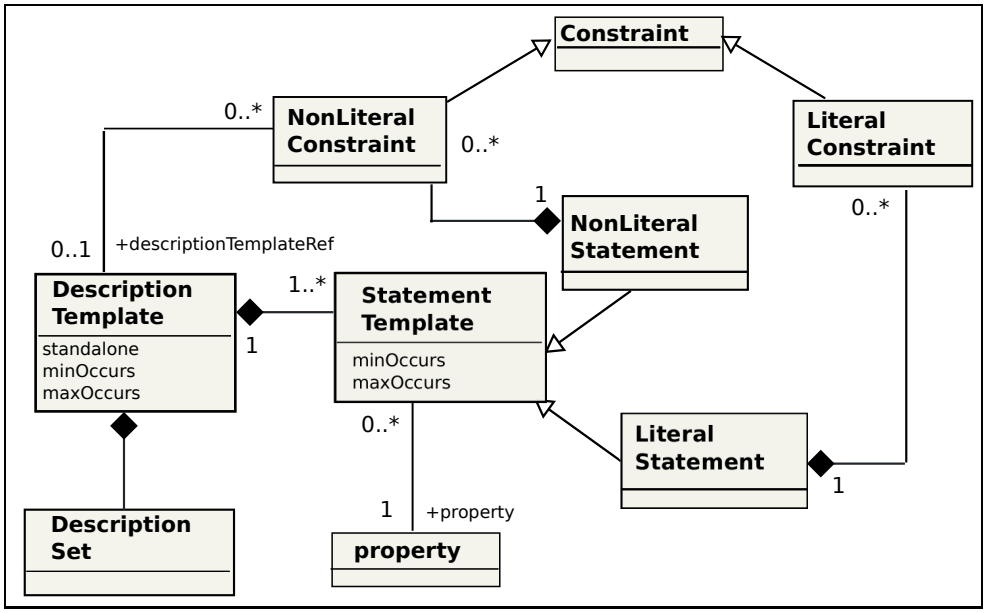

FIGURE 18 – Diagramme de classes UML décrivant le modèle DSP

9.4 Exemple de construction d'un profil DCAP en biologie

Le choix retenu ici est d'adosser la définition du profil d'application pris en exemple, à un modèle DSP défini au travers des langages standards de données RDF et RDFS. L'importance est, ainsi, donnée à des usages du profil d'application dans un contexte de sources de données ouvertes et nous explicitons en conséquence les descriptions des ressources dans les formalismes standards du W3C définis à cet effet. Le DCMI a procédé au même choix, de manière à garantir l'interopérabilité des standards de métadonnées du Dublin Core (à l'exemple de dcterms [57] ou de dcam [96]), sur le web. Les formalismes du web sémantique sont également utilisés pour la production de bon nombre de vocabulaires controlés tels que Gene Ontology [47] ou Sequence Ontology [87].
Nous avons choisi de présenter un exemple très partiel de profil d'application qui pourrait venir faciliter la consultation de séquences nucléiques provenant du portail Entrez. Le cas d'étude choisi porte sur de l'information contextuelle et en particulier sur du traitement de l'information associée aux laboratoires qui ont soumis des séquences nucléiques d'une famille de gènes en particulier. L'objectif est de restituer cette information sous forme de carte géographique après enrichissement à l'aide de référentiels géographiques (ou gazetteer). La figure 20 donne une vue d'ensemble des différents modèles et référentiels mis en jeu pour aboutir à la production du profil d'application. Un modèle conceptuel de domaine (figure 21) vient en support du profil d'application construit et

61. dublincore.org/documents/2008/03/31/dc-dsp/
62. http ://wiki.dublincore.org/index.php/RDF_Application_Profile

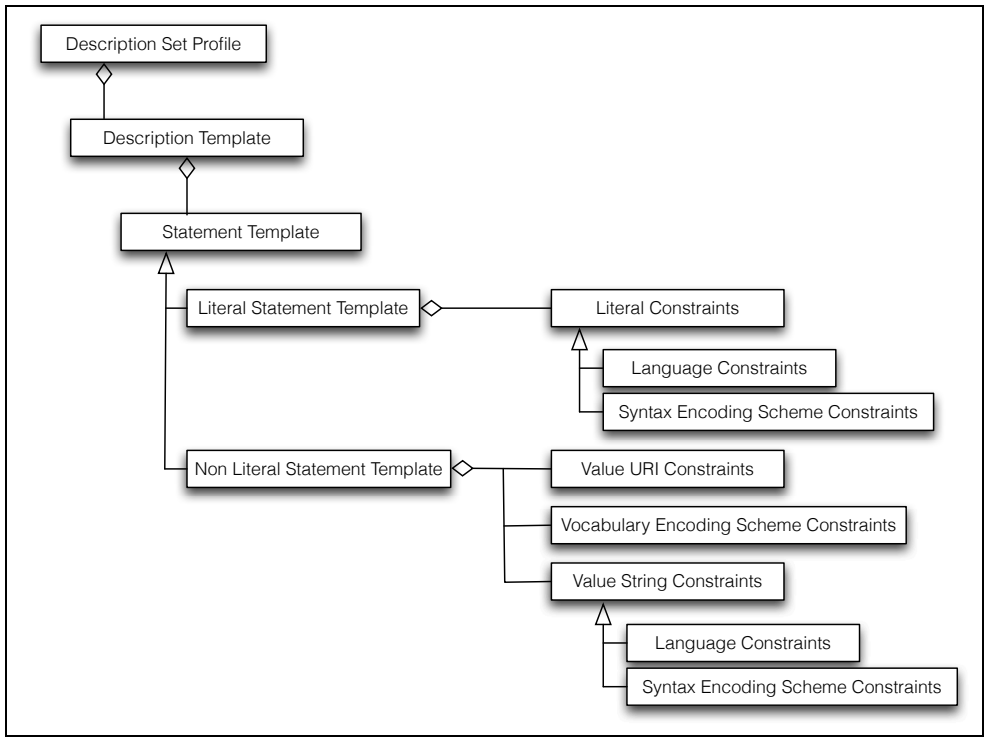

FIGURE 19 – Diagramme de classes UML mettant en exergue les contraintes dans le modèle DSP (extrait de [31]

introduit les entités d'intérêt : Séquence, Biologiste, Laboratoire et Organisme qui seront enrichies par le profil. Les standards de métadonnées et les vocabulaires contrôlés mis à contribution sont respectivement Dublin Core, Darwin Core, Geonames, DCMI Vocabulary et SOFA.

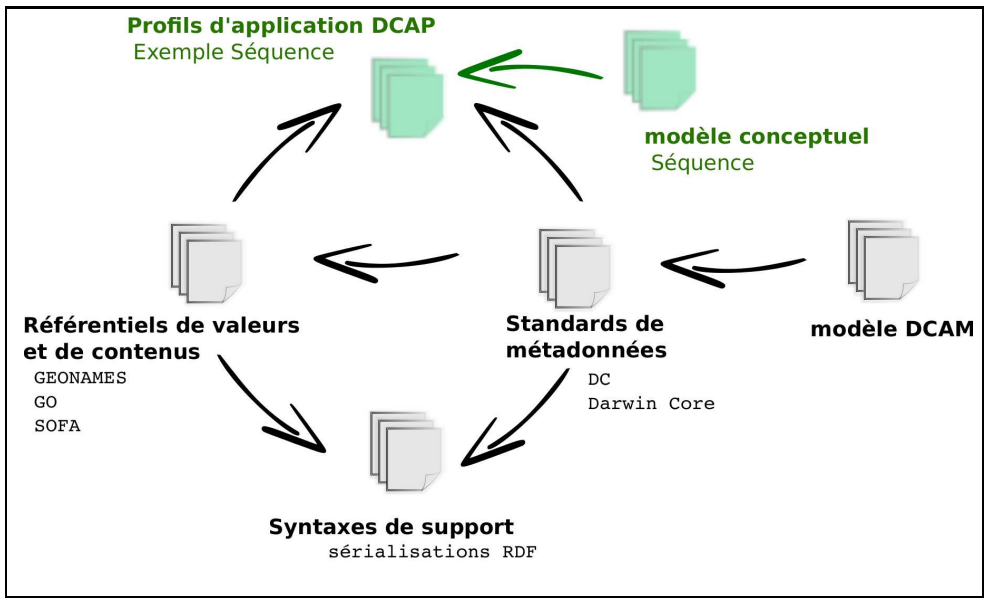

FIGURE 20 – Positionnement du profil d'application exemple

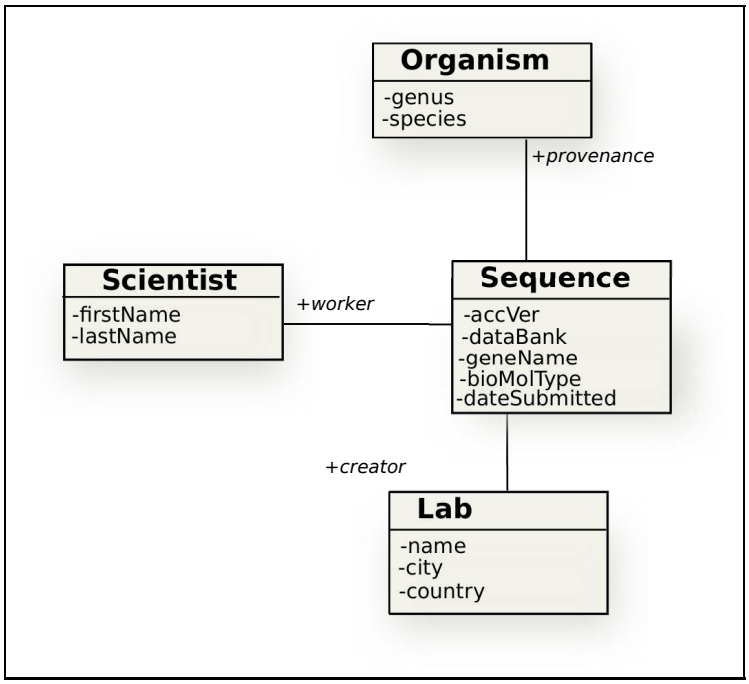

FIGURE 21 – Diagramme de classes du modèle de domaine investi

9.4.1 Les difficultés asociées à un modèle DSP en RDF

Les formalismes RDF/RDFS s'appuient toutefois sur des primitives de modélisation qui ne sont pas toujours en adéquation avec la définition d'un langage de contraintes DSP, ce qui explique que la représentation d'un DSP au travers du langage XML Schema puisse être une solution souvent préférée.
Un premier diagramme de classes UML (figure 22) fait état de StatementTemplate comme étant une classe association qui permet d'attacher, à une description, des couples propriété/contrainte sur les valeurs pouvant être prises par cette propriété. Cette notion de classe association est naturellement prise en charge par le langage XML Schema au travers d'un agrégat de séquences complexes, mais nécessiterait de modéliser des propriétés dites n-aires en RDF/RDFS/OWL alors que ces langages offrent uniquement la représentation d'associations binaires. Une solution est alors de traduire la propriété n-aire au travers d'une ressource anonyme, qui va alors être dotée de n[63] propriétés binaires. Les classes `LiteralConstraint` et `NonLiteralConstraint` sont également représentées comme des classes anonymes. Elles correspondent en réalité non pas à une contrainte, mais à une collection de contraintes qui s'appliquent à l'ensemble des valeurs littérales comme non littérales, pouvant être prises par une propriété. Cette notion de collection est encore une fois bien prise en charge par le langage XML Schema, mais nécessite la traduction de l'agrégat sous forme de classe anonyme ou bien encore sous la forme de l'énonciation de multiples propriétés binaires entre `StatementTemplate` et `Constraint`.
Pour tenter de remédier à la lourdeur de mise en œuvre du modèle DSP en RDF liée à la définition nécessaire de nombreuses ressources anonymes, nous sommes revenus sur la démarche de modélisation. À notre sens, le modèle DSP se prête bien à une repré-

63. arité de la propriété n-aire

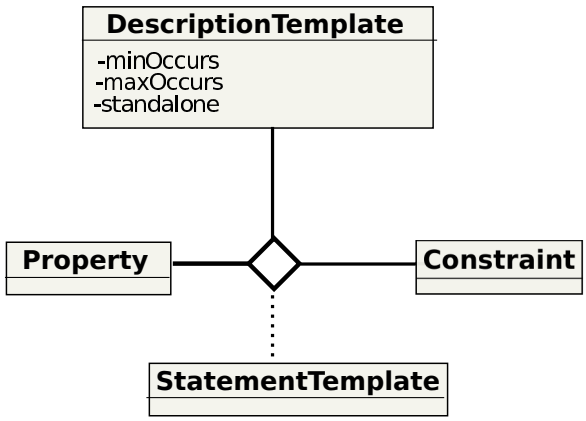

FIGURE 22 – Illustration de la classe association StatementTemplate

sentation UML organisée en trois couches : un méta-niveau, un niveau modèle et un niveau instance. Nous empruntons à cet effet, les stéréotypes définis au sein du profil UML [39, 18] associé au métamodèle des langages RDF/RDFS. Ces stéréotypes sont repris dans la figure 23 et les métaclasses `RDFSClass`, `BlankNode` ou `RDFProperty` apportent un complément de description à chaque classe du DSP.
Par ailleurs, les langages du web de données s'appuient sur l'hypothèse du monde ouvert, avec une organisation des données flexible qui favorise le partage et l'intégration de données hétérogènes, mais qui, a contrario, ne facilite pas la définition de contraintes. Il est alors nécessaire de replacer les modèles de données produits dans un cadre plus proche de l'hypothèse du monde fermé si l'on veut pouvoir tirer parti des contraintes définies. Les choix souvent faits aujourd'hui sont soit d'emprunter au langage OWL pour définir les axiomes venant circonscrire les données (toutes les contraintes ne peuvent cependant pas être posées) ; soit d'assortir le modèle DSP de modèles de vérification de données a posteriori, à l'exemple de ce qui peut être fait au travers du langage SPIN (Sparql Inference Notation) [107] qui étend le langage SPARQL pour dégager des contradictions potentielles sur les données.
Le diagramme de classes simplifié du DSP met l'accent sur les deux grandes difficultés évoquées : les classes anonymes *StatementTemplate* et *Constraint* (stéréotype "BlankNode") proviennent de la traduction des propriétés n-aires ; les contraintes posées sont multiples et posent un cadre sur la définition syntaxique et structurelle d'une description standard et partageable qui peut venir enrichir une ressource d'intérêt.

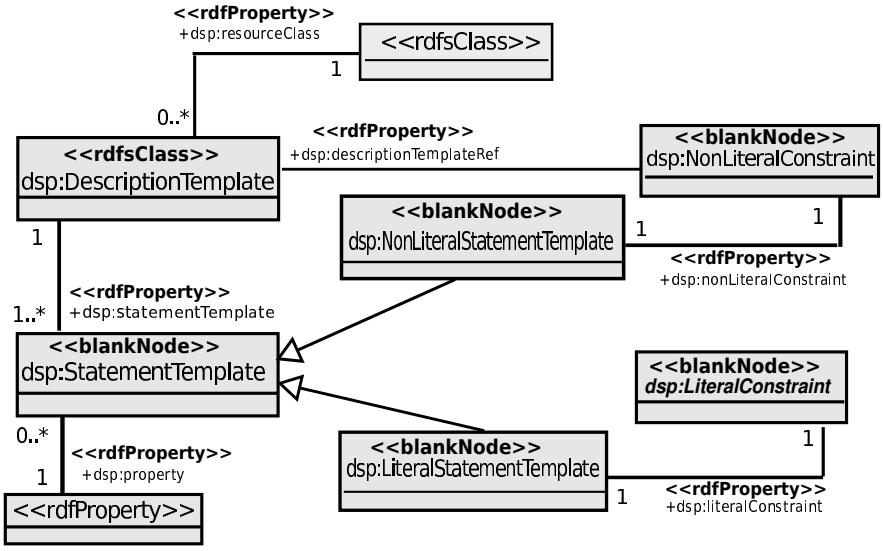

FIGURE 23 – Diagramme de classes DSP simplifié utilisant les stéréotypes du profil UML du métamodèle RDF/RDFS

9.4.2 Définition du modèle DSP pour le profil d'application exemple

Le modèle DSP est envisagé ici comme un métamodèle et nous lui associons un profil UML qui permet d'envisager les éléments du langage DSP comme des stéréotypes qui sont des sous-types des stéréotypes du profile UML de RDF/RDFS. L'intérêt de la démarche, tirant parti des profils UML et de l'ingénierie des modèles, est de disposer d'un emboîtement de modèles, chaque modèle proposé étant un modèle instance du modèle du niveau au-dessus. Ainsi, il est plus facile de dégager ce qui relève de la modélisation des éléments des langages RDF/RDFS, de ce qui relève des éléments proprement dits du langage du DSP et enfin de ce qui relève des éléments spécifiques du modèle de description des séquences biologiques. Un autre intérêt est de pouvoir exploiter le langage d'expression de contraintes OCL [93] (Object Constraint Language) classiquement utilisé au sein des diagrammes structurels UML. OCL est affranchi de toute considération opérationnelle et les contraintes décrites pourront être ensuite traduites indifféremment au travers d'axiomes OWL [42] ou bien de requêtes SPIN. Nous proposons ici une illustration du modèle spécifique pour la description des séquences nucléiques. Les stéréotypes définis pour le langage DSP sont mis à constribution. Deux entités de type `DescriptionTemplate`, `Sequence_T` et `Lab_T`, sont représentées. L'entité `Sequence_T` est attachée à la description d'une séquence nucléique et est relayée par une collection d'éléments de type `StatementTemplate` à l'exemple de l'élément présenté dans la figure 24, qui permet d'associer à la séquence, une description destinée à fournir des renseignements sur le laboratoire au travers du moule (`Lab_T`). Le moule (`Lab_T`) est associé à une propriété `dct:spatial` empruntée au standard `dcterms` qui pointe sur un objet de la classe `dct:Location` et qui dans notre exemple, correspondra à la ville dans laquelle se trouve le laboratoire de biologie.

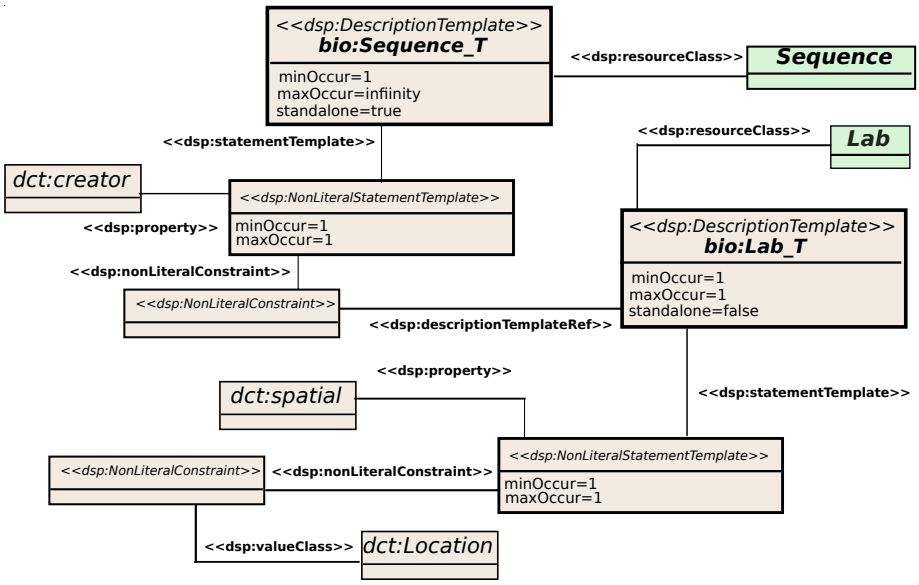

FIGURE 24 – Court extrait du diagramme de classes du modèle de description des séquences conforme à DSP

Le même extrait du profil pour les séquences est proposé au format RDF dans la syntaxe N3.

```
@prefix bio:      <http://test.fr/bio_ap/> .
@prefix foaf:     <http://xmlns.com/foaf/0.1/> .
@prefix dct:      <http://purl.org/dc/terms/> .
@prefix xsd:      <http://www.w3.org/2001/XMLSchema#> .
@prefix dsp:      <http://purl.org/dc/dsp/> .

bio:Sequence_T
      a         dsp:DescriptionTemplate ;
      dsp:maxOccur "infinity"^^xsd:nonNegativeInteger ;
      dsp:minOccur "1"^^xsd:nonNegativeInteger ;
      dsp:resourceClass bio:Sequence ;
      dsp:standalone "true"^^xsd:boolean ;
      dsp:statementTemplate
            [ a         dsp:NonLiteralStatementTemplate ;
              dsp:maxOccur "1" ;
              dsp:minOccur "1"^^xsd:nonNegativeInteger ;
              dsp:nonliteralConstraint
                    [ a         dsp:NonLiteralConstraint ;
                      dsp:descriptionTemplateRef
                            bio:Lab_T ;
                      dsp:valueURIOccurrence
                            "mandatatory" ;
                      dsp:vocabularyEncodingSchemeOccurrence
                            "disallowed"
                    ] ;
              dsp:property dct:creator
            ] .

bio:Lab_T
      dsp:statementTemplate
            [ a         dsp:NonLiteralStatementTemplate ;
              dsp:maxOccur "1" ;
              dsp:minOccur "1"^^xsd:nonNegativeInteger ;
              dsp:nonliteralConstraint
                    [ a         dsp:NonLiteralConstraint ;
                      dsp:valueClass dct:Location
                    ] ;
              dsp:property dct:spatial
            ] .
```

Listing 8 – Court extrait du modèle DSP des séquences

Pour ce qui relève des contraintes exprimables au travers du langage OCL, nous donnons l'exemple de règles qui s'appliquent à l'entité `Sequence_T`. Les contraintes sont définies au niveau de l'entité mais vont concerner les individus de cette entité. Ainsi, les contraintes posées dans le listing 9 vérifient qu'un objet `Sequence_T` renseigne obligatoirement un objet `Sequence` et renseigne un et un seul objet. De même, un objet `Sequence_T` ne dépend d'aucune autre description.

```
context eit : Sequence_T
inv: eit.resourceClass.oclIsTypeOf (Sequence)
inv: eit.minOccur = 1
inv: eit.maxOccur = 1
inv: eit.standalone = true
```

Listing 9 – OCL constraints

Le modèle DSP pour les séquences biologiques est à instancier une fois construit. Nous avons exploité à cet effet les services web d'Entrez nommés E-Utilities et notamment la fonction eFetch qui permet de retourner des documents XML de séquences qui satisfont les critères de sélection passés en arguments dans l'URL. Les triplets instance du modèle DSP ont été rendus persistants au sein du triplestore TDB de Jena et une consultation du triplestore à partir de la librairie JavaScript OpenLayers permet de proposer la viualisation résultat.

10 Conclusion et perspectives

Nous avons essayé de montrer autour d'un exemple simple en biologie, l'intérêt de la construction et de la mise en œuvre de profils d'application dans le contexte du web de données. Dans les sciences expérimentales, les jeux de données sont multiples et peuvent être exploités dans de nombreuses hypothèses pour la construction ou la validation de modèles théoriques. Ces jeux de données peuvent être de plus enrichis par de multiples descriptions expertes faisant appel à des ressources termino-ontologiques de spécialité. Les profils d'application apparaissent comme à même d'apporter des solutions méthodologiques pour la construction de modèles intégrés favorisant le croisement d'informations et exploitant les mécanismes d'interopérabilité mis en place au sein du web de données. À notre sens, un profil d'application Dublin Core est une solution de médiation qui permet de poser des articulations entre le savoir collectif réprésenté aujourd'hui par les sources de données liées et ouvertes et les besoins applicatifs à l'échelle d'une communauté de pratique. Il n'en demeure pas moins une certaine difficulté à la mise en pratique de profils d'application qui nécessite de manipuler de nombreux référentiels et plusieurs niveaux de modélisation. Dans nos travaux actuels, nous nous intéressons aux activités de méta-modélisation pour également proposer des applicatifs web qui facilitent la construction, l'instanciation et la consultation de modèles DSP. Le développement de ces applicatifs ont d'ores et déjà fait l'objet de stages de master en informatique dans le contexte de l'EQUIPEX GEOSUD [37]. L'importance est donnée à la mise en place d'un profil d'application pour faciliter l'accès et la consultation de milliers d'images satellites acquises à l'échelle du territoire national. À très court terme, nous nous intéresserons aussi à la définition de contraintes sur les modèles DSP ainsi qu'à la manière de définir ces contraintes. Un profil d'application emprunte des éléments de modélisation à de nombreux vocabulaires et y applique de nombreuses contraintes. Toutefois la vérification de ces contraintes n'est pas prise en charge. Il est ainsi nécessaire de définir des mécanismes de vérification qui garantissent que l'usage des éléments de modélisation des référentiels ne soient pas dévoyés ou encore que les données des jeux de données soient correctement décrites par les instances du modèle DSP. Une solution possible est de déléguer la vérification à des mécanismes de post-traitement en exploitant par exemple le

formalisme SPIN (Sparql Inference Notation) qui est un langage de définition de règles qui s'appuie à cet effet sur SPARQL. Les règles SPIN peuvent également venir s'intégrer à la défintion des moules de description (DescriptionTemplate) [107].

Références

[1] Geographic Information – Metadata. ISO 19115 :2003, May 2003.

[2] Samuel Andrés, Damien Arvor, Laurent Durieux, Marie-Angélique Laporte, Thérése Libourel, Isabelle Mougenot, and Christelle Pierkot. Ontologies Contribution to link thematic and remote sensing knowledge : preliminary discussions. In *Selper 2012*, 2012.

[3] Marcelo Arenas, Claudio Gutierrez, and Jorge Pérez. Foundations of RDF Databases. In Sergio Tessaris, Enrico Franconi, Thomas Eiter, Claudio Gutierrez, Siegfried Handschuh, Marie-Christine Rousset, and RenateA. Schmidt, editors, *Reasoning Web. Semantic Technologies for Information Systems*, volume 5689 of *Lecture Notes in Computer Science*, pages 158–204. Springer Berlin Heidelberg, 2009.

[4] M. Ashburner, B. Barrell, P. Benos, V. Bolshakov, A. Bucheton, S. Cox, P. Deak, J. Demaille, C. Ferraz, F. Galibert, D. Glover, D. Harris, H. Jaeckle, F. Kafatos, C. Louis, E. Madueno, J. Modolell, I. Mougenot, L. Murphy, G. Papagiannakis, J. Rogers, C. Salles, R. Saunders, C. Savakis, U. Schaefer, I. Siden-Kiamos, L. Spanos, and Y. Zhang. European drosophila sequencing consortium : sequencing the x chromosome of the fly. http ://www-lbit.iro.umontreal.ca/ISMB98/anglais/accposters.html.

[5] Michael Ashburner. Ontologies for Biologists - A Community Model for the Annotation of Genomic Data. In *Fourth International IEEE Computer Society Computational Systems Bioinformatics Conference, CSB 2005, Stanford, CA, USA, August 8-11, 2005*, page 7, 2005.

[6] Philip R Bagley. Extension of programming language concepts. Technical report, University City Science Center, Philadelphia, PA, 1968.

[7] Amos Bairoch and Rolf Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28 :45–48, 2000.

[8] Thomas Baker. Maintaining Dublin Core as a Semantic Web Vocabulary. In Matthias Hemmje, Claudia Niederée, and Thomas Risse, editors, *From Integrated Publication and Information Systems to Information and Knowledge Environments*, volume 3379 of *Lecture Notes in Computer Science*, pages 61–68. Springer Berlin Heidelberg, 2005.

[9] Jie Bao, Elisa Kendall, Deborah McGuinness, and Peter Patel-Schneider. OWL 2 web ontology language : quick reference guide, 2009. `http://www.w3.org/TR/owl2-quick-reference/`.

[10] Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lynn Andrea Stein. OWL Web Ontology Language Reference, 2004.

[11] Beckett, David and Berners-Lee, Tim and Prud'hommeaux, Eric and Carothers, Gavin . RDF 1.1 Turtle - Terse RDF Triple Language. Technical report, W3C, 2014.

[12] Dennis A. Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. GenBank. *Nucleic Acids Research*, 41(D1) :D36–D42, 2013.

[13] Dennis A. Benson et. al. The GenBank Genetic Sequence Databank. *Nucleic Acids Research*, 35 :21–25, 2007.

[14] Chad Berkley, Shawn Bowers, Matthew B. Jones, Joshua S. Madin, and Mark Schildhauer. Improving Data Discovery for Metadata Repositories through Semantic Search. In Leonard Barolli, Fatos Xhafa, and Hui-Huang Hsu, editors, *2009 International Conference on Complex, Intelligent and Software Intensive Systems, CISIS 2009, Fukuoka, Japan, March 16-19, 2009*, pages 1152–1159. IEEE Computer Society, 2009.

[15] Tim Berners-Lee, Dan Connolly, Lalana Kagal, Yosi Scharf, and Jim Hendler. N3Logic : A logical framework for the World Wide Web. *Theory and Practice of Logic Programming*, 8 :249–269, 5 2008.

[16] Tim Berners-Lee, Roy Fielding, and Larry Masinter. RFC 3986 - Uniform Resource Identifier (URI) : Generic Syntax, January 2005.

[17] Dan Brickley and Ramanathan V. Guha. RDF Vocabulary Description Language 1.0 : RDF Schema. http ://www.w3.org/TR/rdf-schema/, February 2004.

[18] Saartje Brockmans, Peter Haase, Pascal Hitzler, and Rudi Studer. A Metamodel and UML Profile for Rule-extended OWL DL Ontologies. In York Sure and John Domingue, editors, *The Semantic Web : Research and Applications*, volume 4011 of *LNCS*, pages 303–316, Budva, Montenegro, June 2006. Springer.

[19] Robert Cailliau and Helen Ashman. Hypertext in the Web - a history. *ACM Comput. Surv.*, 31(4es) :35, 1999.

[20] Evelyn Camon, Daniel Barrell, Vivian Lee, Emily Dimmer, and Rolf Apweiler. The Gene Ontology Annotation (GOA) Database - an integrated resource of GO annotations to the UniProt Knowledgebase. *In Silico Biol*, 4(1) :5–6, 2004.

[21] Caterina Caracciolo, Armando Stellato, Ahsan Morshed, Gudrun Johannsen, Sachit Rajbhandari, Yves Jaques, and Johannes Keizer. The agrovoc linked dataset. *Semantic Web*, 4(3) :341–348, 2013.

[22] Dominique Chichereau, Odile Contat, Danièle Dégez, Alina Deniau, Michèle Lénart, Claudine Masse, and Dominique Ménillet. Les normes de conception, gestion et maintenance de thésaurus. *Documentaliste-Sciences de l'Information*, 44(1) :66–74, 2007.

[23] Guy Cochrane, Blaise Alako, Clara Amid, Lawrence Bower, Ana Cerdeño-Tárraga, Iain Cleland, Richard Gibson, Neil Goodgame, Mikyung Jang, Simon Kay, Rasko Leinonen, Xiu Lin, Rodrigo Lopez, Hamish McWilliam, Arnaud Oisel, Nima Pakseresht, Swapna Pallreddy, Youngmi Park, Sheila Plaister, Rajesh Radhakrishnan, Stephane Rivière, Marc Rossello, Alexander Senf, Nicole Silvester, Dmitriy Smirnov, Petra ten Hoopen, Ana Toribio, Daniel Vaughan, and Vadim Zalunin. Facing growth in the European Nucleotide Archive. *Nucleic Acids Research*, 41(D1) :D30–D35, 2013.

[24] Karen Coyle. RDA in RDF. *Library Technology Reports*, 46(2) :26–37, 2010.

[25] Morgan V. Cundiff. An introduction to the Metadata Encoding and Transmission Standard (METS). *Library Hi Tech*, 22(1) :52–64, 2004.

[26] Angela Dappert and Markus Enders. Using METS, PREMIS and MODS for archiving eJournals. *D-Lib Magazine*, 14(9/10), 2008.

[27] Ian Davis and Richard Newman. Expression of Core FRBR Concepts in RDF. http ://vocab.org/frbr/core.html, 2009.

[28] Patrice Déhais and Isabelle Mougenot. An interactive system for database in immunogenetics. In *HICSS (5)*, pages 25–34, 1994.

[29] Lorcan Dempsey. Interoperability : the value of recombinant potential, 2004. http ://fr.slideshare.net/lisld/interoperability-the-value-of-recombinant-potential.

[30] Jean-Christophe Desconnets, Hatim Chahdi, and Isabelle Mougenot. Application profile for earth observation images. In *Metadata and Semantics Research - 8th Research Conference, MTSR 2014, Karlsruhe, Germany, November 27-29, 2014. Proceedings*, pages 68–82, 2014.

[31] Jean-Christophe Desconnets, Hatim Chahdi, and Isabelle Mougenot. Application Profile for Earth Observation Images. In Sissi Closs, Rudi Studer, Emmanouel Garoufallou, and Miguel-Angel Sicilia, editors, *Metadata and Semantics Research*, volume 478 of *Communications in Computer and Information Science*, pages 68–82. Springer International Publishing, 2014.

[32] Mamadou Dieye, Mohamed Rafik Doulache, Mustapha Floussi, Julie Chabalier, Isabelle Mougenot, and Mathieu Roche. Construction d'un dictionnaire multilingue de biodiversité à partir de dires d'experts. In *INFORSID*, pages 81–88, 2012. http ://liris.cnrs.fr/inforsid/sites/default/files/2012_2_1-MamadouDieye.pdf.

[33] Ousmane Djanga, Hanine Hamzioui, Mickaël Hatchi, Isabelle Mougenot, and Mathieu Roche. Regroupement des définitions de sigles biomédicaux. In Jean-Gabriel Ganascia and Pierre Gançarski, editors, *EGC*, volume RNTI-E-15 of *Revue des Nouvelles Technologies de l'Information*, page 487. Cépaduès-Éditions, 2009.

[34] Patrice Duroux, Isabelle Mougenot, Jacques Divol, and Jean Sallantin. Developement of adaptive interfaces for use in conjunction with the ORIEL project, 2002. E-BioSci/ORIEL Partners Summer Meeting, Wiesloch Germany.

[35] Patrice Duroux, Isabelle Mougenot, Laetitia Régnier, and Jean Sallantin. Adaptive Interfaces and BiODS tools, 2004. E-BioSci / ORIEL Annual Workshop, Hinxton England.

[36] Karen Eilbeck, Suzanna E Lewis, Christopher J Mungall, Mark Yandell, Lincoln Stein, Richard Durbin, and Michael Ashburner. The Sequence Ontology : A tool for the unification of genome annotations. *Genome Biology*, 6(5) :R44, 2005.

[37] Nordine El Hassouni. Environnement web pour l'exploitation de métadonnées rdf : application aux données issues de l'observation de la terre. Master's thesis, Montpellier, France.

[38] Fabien L. Gandon, Reto Krummenacher, Sung-Kook Han, and Ioan Toma. Semantic Annotation and Retrieval : RDF. In John Domingue, Dieter Fensel, and JamesA. Hendler, editors, *Handbook of Semantic Web Technologies*, pages 117–155. Springer Berlin Heidelberg, 2011.

[39] Dragan Gaševic, Dragan Djuric, and Vladan Devedžic. The Ontology UML Profile. In *Model Driven Engineering and Ontology Development*, pages 235–243. Springer Berlin Heidelberg, 2009.

[40] Anne J. Gilliland. *Setting the Stage*, pages 189–197. Getty Information Institute., 2008.

[41] Diny Golder, Les Kneebone, Jon Phipps, Steve Sunter, and Stuart A. Sutton. A configurable rdf editor for australian curriculum. In Gobinda G. Chowdhury, Chris Koo, and Jane Hunter, editors, *ICADL*, volume 6102 of *Lecture Notes in Computer Science*, pages 189–197. Springer, 2010.

[42] B. Grau, I. Horrocks, B. Motik, B. Parsia, P. Patelschneider, and U. Sattler. OWL 2 : The next step for OWL. *Web Semantics : Science, Services and Agents on the World Wide Web*, 6(4) :309–322, November 2008.

[43] Jane Greenberg. Understanding metadata and metadata schemes. *Cataloging & Classification Quarterly*, 40(3/4) :17–36, 2005.

[44] Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2), 1993.

[45] Yannick Gueguen, Julien Garnier, Lorenne Robert, Marie-Paule Lefranc, Isabelle Mougenot, Julien de Lorgeril, Michael Janech, Paul S. Gross, Gregory W. Warr, Brandon Cuthbertson, Margherita A. Barracco, Philippe Bulet, André Aumelas, Yinshan Yang, Dong Bo, Jianhai Xiang, Anchalee Tassanakajon, David Piquemal, and Evelyne Bachère. PenBase, the shrimp antimicrobial peptide penaeidin database : sequence-based classification and recommended nomenclature. *Developmental & Comparative Immunology*, 30 :283–288, 2006.

[46] Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. The semantic measures library and toolkit : fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics*, 30(5) :740–742, 2014.

[47] M A Harris, J Clark, A Ireland, J Lomax, M Ashburner, R Foulger, K Eilbeck, S Lewis, B Marshall, C Mungall, J Richter, G M Rubin, J A Blake, C Bult, M Dolan, H Drabkin, J T Eppig, D P Hill, L Ni, M Ringwald, R Balakrishnan, J M Cherry, K R Christie, M C Costanzo, S S Dwight, S Engel, D G Fisk, J E Hirschman, E L Hong, R S Nash, A Sethuraman, C L Theesfeld, D Botstein, K Dolinski, B Feierbach, T Berardini, S Mundodi, S Y Rhee, R Apweiler, D Barrell, E Camon, E Dimmer, V Lee, R Chisholm, P Gaudet, W Kibbe, R Kishore, E M Schwarz, P Sternberg, M Gwinn, L Hannick, J Wortman, M Berriman, V Wood, N de la Cruz, P Tonellato, P Jaiswal, T Seigfried, and R White. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, 32(Database issue) :258–261, Jan 2004.

[48] Tom Heath and Christian Bizer. *Linked Data : Evolving the Web into a Global Data Space*, volume 1 of *Synthesis Lectures on the Semantic Web : Theory and Technology*. Morgan & Claypool, 1 edition, 2011.

[49] Maxime Hébrard, Gaël Manes, Béatrice Bocquet, Isabelle Meunier, Isabelle Mougenot, and Christian Hamel. A knowledge-based system for diagnosis dedicated to inherited retinal dystrophies. In *JOBIM, Journées Ouvertes en Biologie, Informatique et Mathématiques, Session Poster*, 2012. http ://jobim2012.inria.fr/sources/p72.pdf.

[50] Rachel Heery and Manjula Patel. Application profiles : mixing and matching metadata schemas. *Ariadne*, 25, 2000.

[51] Diane I. Hillman, Jon Phipps, and Karen Coyle. Introduction to application profiles, 2010.

[52] Diane Hillmann, Karen Coyle, Jon Phipps, and Gordon Dunsire. RDA Vocabularies. *DLib Magazine*, 16(1/2), 2010.

[53] International Organization for Standardization. Geographic information – Metadata - Part 1 : Fundamentals. ISO 19115 :2014, 2014.

[54] Antoine Isaac, John Phipps, and Daniel Rubin. Skos use cases and requirements. W3c working draft, W3C, May 2007.

[55] Krzysztof Janowicz, Pascal Hitzler, Benjamin Adams, Dave Kolas, and Charles Vardeman. Five stars of linked data vocabulary use. *Semantic Web*, 5(3) :173–176, 2014.

[56] Simon Jupp, Sean Bechhofer, and Robert Stevens. A flexible api and editor for skos. In Christian Bizer and Anupam Joshi, editors, *International Semantic Web Conference (Posters & Demos)*, volume 401 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.

[57] Stefan Kokkelink and Roland Schwänzl. Expressing Dublin Core metadata using the Resource Description Framework (RDF). Available via the World Wide Web at `http://dublincore.org/documents/dc-rdf/`.

[58] Markus Krötzsch. Owl 2 profiles : An introduction to lightweight ontology languages. In Thomas Eiter and Thomas Krennwallner, editors, *Reasoning Web. Semantic Technologies for Advanced Query Answering*, volume 7487 of *Lecture Notes in Computer Science*, pages 112–183. Springer Berlin Heidelberg, 2012.

[59] Marie-Angélique Laporte, Eric Garnier, and Isabelle Mougenot. Construction collective de standards de données en écologie. In *Journées Francophones sur les Ontologies (JFO 2011), Montréal, Canada*, 2011.

[60] Marie-Angélique Laporte, Eric Garnier, and Isabelle Mougenot. Construction collective de standards de données en écologie. In *Journées Francophones sur les Ontologies (JFO 2011), Montréal, Canada*, 2011.

[61] Marie-Angélique Laporte, Eric Garnier, and Isabelle Mougenot. A Faceted Search System for Facilitating Discovery-driven Scientific Activities : A Use Case from Functional Ecology. In *The Semantic Web : ESWC 2013 Satellite Events, 1st International Workshop on Semantics for Biodiversity (S4BioDiv'13)*, May 26-30 2013. http ://ceur-ws.org/Vol-979/.

[62] Marie-Angélique Laporte, Eric Garnier, and Isabelle Mougenot. A Faceted Search System for Facilitating the Understanding and the Prediction of Ecosystem Changes. In *TDWG biodiversity information standards 2013*, October 2013.

[63] Marie-Angélique Laporte, Isabelle Mougenot, and Eric Garnier. Thesauform - traits : A web based collaborative tool to develop a thesaurus for plant functional diversity research. *Ecological Informatics*, 11 :34–44, 2012.

[64] Marie-Angélique Laporte, Isabelle Mougenot, and Eric Garnier. Thesauform - traits : A web based collaborative tool to develop a thesaurus for plant functional diversity research. *Ecological Informatics*, 11 :34–44, 2012.

[65] Marie-Angélique Laporte, Isabelle Mougenot, Eric Garnier, Ulrike Stahl, Lutz Maicher, and Jens Kattge. A Semantic Web Faceted Search System for Facilitating Building of Biodiversity and Ecosystems Services. In *Data Integration in the Life Sciences - 10th International Conference, DILS 2014, Lisbon, Portugal, July 17-18, 2014. Proceedings*, pages 50–57, 2014.

[66] Marie-Angélique Laporte, Isabelle Mougenot, Eric Garnier, Ulrike Stahl, Lutz Maicher, and Jens Kattge. A semantic web faceted search system for facilitating building of biodiversity and ecosystems services. In *Data Integration in the Life Sciences - 10th International Conference, DILS 2014, Lisbon, Portugal, July 17-18, 2014. Proceedings*, pages 50–57, 2014.

[67] Pierre Larmande, C. Tranchant-Dubreuil, L. Regnier, Isabelle Mougenot, and Thérèse Libourel. Integration of data sources for plant genomics. In Yannis Manolopoulos, Joaquim Filipe, Panos Constantopoulos, and José Cordeiro, editors, *ICEIS (1)*, pages 314–318, 2006.

[68] Marie-Paule Lefranc, Véronique Giudicelli, Chantal Busin, Ansar Malik, Isabelle Mougenot, Patrice Déhais, and Denys Chaume. LIGM-DB/IMGT : an integrated database of Ig and TcR, part of the immunogenetics database. *Annals of the New York Academy of Sciences*, 764 :47–49, 1995.

[69] Alain Léger, Géraldine Arbaut, Peter Barrett, Sylvain Gitton, Asuncion Gomez-Pérez, René Holm, Aarno Lehtola, Isabelle Mougenot, Ana Nistal, Theodora Varvarigou, and Jérôme Vinesse. MKBEEM : Ontology Domain Modeling Support for Multi-lingual services in E-Commerce. In *ECAI'00, Workshop on Applications of Ontologies and Problem-Solving Methods, Berlin, Germany*, pages 19.1–19.4, August 2000.

[70] Thérèse Libourel, Yuan Lin, Isabelle Mougenot, Christelle Pierkot, and Jean-Christophe Desconnets. A platform dedicated to share and mutualize environmental applications. In Joaquim Filipe and José Cordeiro, editors, *ICEIS (1)*, pages 50–57. SciTePress, 2010.

[71] Yuan Lin, Marie-Angélique Laporte, Lucile Soler Isabelle Mougenot, and Thérèse Libourel. An organizational environment for "in silico" experiments in molecular biology. In *ESWC, 4th International workshop on Resource Discovery (RED'2011), Heraklion, Greece*, 2011. http ://ceur-ws.org/Vol-737/paper4.pdf.

[72] Yuan Lin, Thérèse Libourel, and Isabelle Mougenot. A workflow language for the experimental sciences. In José Cordeiro and Joaquim Filipe, editors, *ICEIS (3)*, pages 372–375, 2009.

[73] Yuan Lin, Thérèse Libourel, and Isabelle Mougenot. Autour des chaînes de traitements dédiées aux applications environnementales. In *LMO, Langages et Modèles à Objets, Pau, Mars 2010*, 2010.

[74] Yuan Lin, Thérèse Libourel, Isabelle Mougenot, Runtong Zhang, and Rongqian Ni. Approach for verifying workflow validity. In Runtong Zhang, José Cordeiro, Xuewei Li, Zhenji Zhang, and Juliang Zhang, editors, *ICEIS (3)*, pages 66–75. SciTePress, 2011.

[75] Yuan Lin, Isabelle Mougenot, and Thérèse Libourel. Environnement de workflow scientifique validation et conformités. In *INFORSID*, pages 129–144, 2011. http ://liris.cnrs.fr/inforsid/sites/default/files/2011_55.pdf.

[76] Yuan Lin, Isabelle Mougenot, and Thérèse Libourel. Method and components for creating scientific workflow. In *Workshops Proceedings of the 30th International Conference on Data Engineering Workshops, ICDE 2014, Chicago, IL, USA, March 31 - April 4, 2014*, pages 147–153, 2014.

[77] Yuan Lin, Christelle Pierkot, Isabelle Mougenot, Jean-Christophe Desconnets, and Thérèse Libourel. A framework to assist environmental information processing. In Joaquim Filipe and José Cordeiro, editors, *ICEIS*, volume 73 of *Lecture Notes in Business Information Processing*, pages 76–89. Springer, 2010.

[78] Yuan Lin, Christelle Pierkot, Isabelle Mougenot, Jean-Christophe Desconnets, and Thérèse Libourel. A Framework to Assist Environmental Information Processing. In Joaquim Filipe and José Cordeiro, editors, *Enterprise Information Systems*, volume 73 of *Lecture Notes in Business Information Processing*, pages 76–89. Springer Berlin Heidelberg, 2011.

[79] Mariana Curado Malta and Ana Alice Baptista. State of the art on methodologies for the development of a metadata application profile. In Juan Manuel Dodero, Manuel Palomo-Duarte, and Pythagoras Karampiperis, editors, *MTSR*, volume 343 of *Communications in Computer and Information Science*, pages 61–73. Springer, 2012.

[80] Bruno Menon. Le web sémantique : de nouveaux enjeux documentaires ? *Documentaliste-Sciences de l'Information*, 40(6) :387–391, 2003.

[81] William K Michener, James W Brunt, John J Helly, Thomas B Kirchner, and Susan G Stafford. Nongeospatial Metadata for the Ecological Sciences. *Ecological Applications*, 7(1) :330–342, 1997.

[82] Isabelle Mougenot, Samuel Andrès, and Marie-Angélique Laporte. Recherche méthodologique pour la conception d'une ontologie métier pour l'aide à l'interprétation d'images satellites, 2014. Livrable du WP3.1 ANR Equipex GEOSUD, Investissements d'Avenir (2011).

[83] Isabelle Mougenot, Patrice Duroux, Francis Daumas, and Jean Sallantin. Development of interactive interfaces, 2004. Deliverable D5.2 of EU IST project 2001-32688 E-Oriel : Online Research Information Environment for the Life Sciences.

[84] Isabelle Mougenot, Patrice Duroux, Jean Sallantin, Jacques Divol, Mireille Gay, and Francis Daumas. D5.1 : Requirements, choice and justification of core technologies and concepts to be employed for implementation and demonstration, 2002. Deliverable D5.1 of EU IST project 2001-32688 E-Oriel : Online Research Information Environment for the Life Sciences.

[85] Isabelle Mougenot, Thérèse Libourel, and Patrice Déhais. Genetic sequence annotation within biological databases. In Tok Wang Ling and Yoshifumi Masunaga, editors, *DASFAA*, volume 5 of *Advanced Database Research and Development Series*, pages 333–341. World Scientific, 1995.

[86] Nicola J Mulder, Rolf Apweiler, Teresa K Attwood, Amos Bairoch, Alex Bateman, David Binns, Paul Bradley, Peer Bork, Phillip Bucher, Lorenzo Cerutti, Richard Copley, Emmanuel Courcelle, Ujjwal Das, Richard Durbin, Wolfgang Fleischmann, Julian Gough, Daniel Haft, Nicola Harte, Nicolas Hulo, Daniel Kahn, Alexander Kanapin, Maria Krestyaninova, David Lonsdale, Rodrigo Lopez, Ivica Letunic, Martin Madera, John Maslen, Jennifer McDowall, Alex Mitchell, Anastasia N Nikolskaya, Sandra Orchard, Marco Pagni, Chris P Ponting, Emmanuel Quevillon,

Jeremy Selengut, Christian J A Sigrist, Ville Silventoinen, David J Studholme, Robert Vaughan, and Cathy H Wu. InterPro, progress and status in 2005. *Nucleic Acids Res*, 33(Database issue) :201–205, Jan 2005.

[87] Christopher J. Mungall, Colin R. Batchelor, and Karen Eilbeck. Evolution of the Sequence Ontology terms and relationships. *Journal of Biomedical Informatics*, 44(1) :87–93, 2011.

[88] Sergio Muñoz, Jorge Pérez, and Claudio Gutierrez. Simple and efficient minimal RDFS. *Journal of Web Semantics*, 7(3) :220–234, 2009.

[89] Mikael Nilsson, Pete Johnston, Ambjörn Naeve, and Andy Powell. Towards an interoperability framework for metadata standards. *International Conference on Dublin Core and Metadata Applications, DC-2006*.

[90] Mikael Nilsson, AlistairJ. Miles, Pete Johnston, and Fredrik Enoksson. Formalizing dublin core application profiles - description set profiles and graph constraints. In Miguel-Angel Sicilia and MiltiadisD. Lytras, editors, *Metadata and Semantics*, pages 101–111. Springer US, 2009.

[91] Thomas. Nilsson, Mikael. Baker and Pete Johnston. The singapore framework for dublin core application profiles, 2008.

[92] NISO. *ANSI/NISO Z39.19 - Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*, 2005.

[93] Object Management Group (OMG). Object Constraint Language (OCL). Version 2.3.1. http ://www.omg.org/spec/OCL/2.3.1/, 2012.

[94] National Information Standards Organization. Understanding Metadata. Technical report, University City Science Center, Philadelphia, PA, 2004.

[95] James M. Ostell. Entrez : The NCBI Search and Discovery Engine. In Olivier Bodenreider and Bastien Rance, editors, *Data Integration in the Life Sciences*, volume 7348 of *Lecture Notes in Computer Science*, pages 1–4. Springer Berlin Heidelberg, 2012.

[96] Andy Powell, Mikael Nilsson, Ambjörn Naeve, Pete Johnston, and Thomas Baker. DCMI Abstract Model. DCMI Recommendation, June 2007. http ://dublincore.org/documents/2007/06/04/abstract-model/.

[97] Andy Powell, Mikael Nilsson, Ambjörn Naeve, Pete Johnston, and Thomas Baker. DCMI Abstract Model, 2012.

[98] Sarah Pulis and Liddy Nevile. Using the DC Abstract Model to support application profile developers. *International Conference on Dublin Core and Metadata Applications*, 2006.

[99] Resource description framework (RDF). Available via the World Wide Web at `http://www.w3.org/RDF/`.

[100] RDF/XML Syntax Specification. Available via the World Wide Web at `http://www.w3.org/TR/rdf-syntax-grammar/`.

[101] Maryse Rizza. Document et musée : du discours sur l'œuvre à la médiation culturelle. *Documentaliste-Sciences de l'Information*, 50(2), 2014.

[102] Samuel Andrès and Damien Arvor and Thérèse Libourel and Isabelle Mougenot and Marie-Angélique Laporte and Laurent Durieux. Vers un paradigme ontologique d'interprétation des images de télédétection. In *Conférence internationale de Géomatique et Analyse Spatiale SAGEO'14*, Grenoble, France, 2014.

[103] François Scharffe and Jérôme Euzenat. MeLinDa : an interlinking framework for the web of data. Rapport de recherche 7691, INRIA, July 2011.

[104] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, The OBI Consortium, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H Scheuermann, Nigam Shah, Patricia L Whetzel, and Suzanna Lewis. The OBO Foundry : coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, 25(11) :1251–1255, Nov 2007.

[105] Dagobert Soergel, Boris Lauser, Anita C. Liang, Frehiwot Fisseha, Johannes Keizer, and Katz Stephen. Reengineering thesauri for new applications : the AGROVOC example. *Journal of Digital Information*, 2004.

[106] Mojdeh SoltanMohammadi. Définition de modèles de connaissances pour la gestion et l'exploitation de données spatio-temporelles. Master's thesis, Montpellier, France.

[107] spinrdf.org. SPIN - SPARQL Inferencing Notation, 2012.

[108] Manu Sporny, Gregg Kellogg, and Markus Lanthaler. JSON-LD 1.0 - A JSON-based Serialization for Linked Data. Technical report, W3C, 2014.

[109] Shigeo Sugimoto, Thomas Baker, and Stuart L. Weibel. Dublin Core : Process and Principles. In Ee-Peng Lim, Schubert Foo, Chris Khoo, Hsinchun Chen, Edward Fox, Shalini Urs, and Thanos Costantino, editors, *Digital Libraries : People, Knowledge, and Technology*, volume 2555 of *Lecture Notes in Computer Science*, pages 25–35. Springer Berlin Heidelberg, 2002.

[110] Jason Thomale. Interpreting marc : Where's the bibliographic data ? *Code4Lib Journal*, September 2010.

[111] Anis Tissaoui, Nathalie Aussenac-Gilles, Philippe Laublet, and Nathalie Hernandez. Evonto : un outil d'évolution de ressource termino-ontologique pour l'annotation sémantique. *Technique et Science Informatiques*, 32(7-8) :817–840, 2013.

[112] Genoveva Vargas-Solar and Anne Doucet. Médiation de données : solutions et problèmes ouverts. Available via the World Wide Web at http ://www.irit.fr/GDR-I3/fichiers/assises2002/papers/12-MediationDeDonnees.pdf, 2002.

[113] Stuart Weibel, Thomas Baker, Tod Matola, Eric Miller, and Pete Johnston. Namespace Policy for the Dublin Core Metadata Initiative (DCMI). DCMI Recommendation, July 2007.

[114] Wikipedia. Semantic web stack, 2012.

[115] Benjamin Zapilko, Johann Schaible, Philipp Mayr, and Brigitte Mathiak. Thesoz : A skos representation of the thesaurus for the social sciences. *Semantic Web*, (3) :257–263.

[116] Karima Zayrit. Modèles de données adaptés à la construction partagée d'un thésaurus dédié aux traits fonctionnels. Master's thesis, Mémoire de Stage de Master DECOL, UM2, Montpellier, France, 2010.

[117] Marcia Lei Zeng. Metadata Basics. Available via the World Wide Web at `http://marciazeng.slis.kent.edu/metadatabasics/types.htm`.