Académie de Montpellier
Université Montpellier II
Sciences et Techniques du Languedoc

Étude Bibliographique

Métadonnées et interconnexion de sources de données hétérogènes

Auteur :
Hatim CHAHDI

Encadré par :
J.C. Desconnets
I. Mougenot

4 Mars 2013

Table des matières

1. Introduction et Motivations

De nombreux jeux de données géoréférencés sont aujourd'hui localisables, voire rendus disponibles sur le Web, dans les domaines des sciences de l'environnement et du vivant. Ces jeux de données demeurent cependant souvent décrits au travers de différents standards de métadonnées (Darwin Core, Dublin Core, EML, ISO 19115, CSDGM, Sensor ML, ...) pour les besoins propres de chacune des communautés qui les a produites ; ce qui les rend difficilement exploitables au travers d'outils communs de recherche d'information.
Ce constat nous amène à poser une réflexion en terme d'existant autour du rôle facilitateur joué par les métadonnées dans un contexte d'exploitation de gros volumes de données et autour des démarches menant à l'interconnexion de jeux de données géoréférencés. De nombreux travaux sont conduits depuis plusieurs années afin de rendre interopérables différents standards de métadonnées et de permettre la conversion de lots de métadonnées d'un standard à un autre et d'en assurer un échange et une exploitation fédérée efficace.
Nous avons mené une étude bibliographique qui se concentre sur différents aspects :

- la notion de métadonnée est abordée au travers de ses finalités dans le contexte du web sémantique (ou web de données) [LRC02, PG04, CBT04]. Les apports du web de données, en matière d'exploitation des ressources par des agents logiciels dédiés, sont également exposés.
- différentes approches peuvent s'envisager pour assurer un accès uniforme à diverses ressources organisées et décrites de manière hétérogène. Un article de synthèse [HK10] fait le point sur les méthodes qui s'adossent sur les standards de métadonnées pour garantir l'interopérabilité des systèmes.
- le standard de métadonnées Dublin Core, et ses différentes composantes (DCMI Term, DCMI Type, DCAM, DCSP, DCAP, ...), sont explorés [PN06, DHSW02]. L'importance est donnée au modèle abstrait du Dublin Core (DCAM) qui apparaît comme un modèle ouvert permettant la description de toute ressource (toute entité munie d'une identité) par une collection de métadonnées possiblement définies au travers de multiples schémas de métadonnées.
- la notion de profil d'application [CZ06, ZC06, HP00] est orientée vers les besoins spécifiques des communautés, en terme de description de ressources au travers de vocabulaires de métadonnées appropriés. Il s'agit de ne retenir au sein de ces profils que les termes de différents vocabulaires de métadonnées nécessaires à cette description.

Nous concluons notre étude en proposant quelques suites à ce travail. Ainsi, il nous semble utile de poursuivre par une mise en application des standards de métadonnées des domaines concernés (sciences de l'environnement et du vivant, information géographique), au travers de la définition d'un profil d'application Dublin Core.

2. Métadonnées et annotations dans le contexte du web sémantique

2.1 Introduction aux métadonnées et annotations sémantiques

Le web met à la disposition des internautes, d'innombrables ressources qu'il s'agit de localiser, d'organiser et d'exploiter de manière adaptée. La réflexion qui a conduit au web sémantique, encore appelé web de données, a été d'envisager le web comme une gigantesque base de données, rendant alors les notions de métadonnée et de schéma de métadonnées incontournables. Les métadonnées sont définies comme des données sur des données et permettent d'organiser et de structurer les ressources disponibles afin de faciliter leur indexation pour la recherche, et optimiser leur exploitation par des agents logiciels [PG04]. Le terme ressource est pris dans son acception la plus générale, comme étant toute entité concrète ou abstraite, susceptible d'être identifiée, nommée, et manipulée à travers de multiples représentations. Dans cette vision utilitaire, les métadonnées sont souvent assimilées à des annotations sémantiques qui viennent enrichir les données présentes au sein de ressources, par exemple textuelles, disponibles sur le web. Différents standards de métadonnées, généralistes ou dédiés à une discipline comme l'écologie, l'hydrologie ou la géographie, structurent ces métadonnées. Dans les domaines des sciences de l'environnement et du géospatial, qui nous intéressent au premier chef, nous pouvons citer les standards de métadonnées Darwin Core (biodiversité), EML (écologie) ou ISO 19115 (géographie). Le standard de métadonnées Dublin Core[NJNP06] a une visée généraliste et fournit un socle commun d'éléments comme les éléments `title`, `creator`, ou `coverage` qui vont renseigner sur l'existence de ressources et en faciliter l'exploitation sur le long terme.

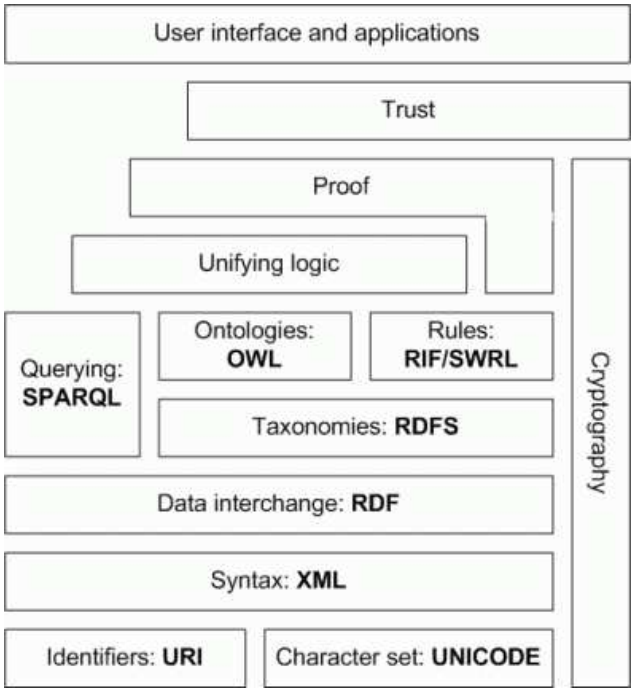

FIGURE 1 – Semantic web stack - Architecture en couches du web sémantique [SBLH06]

Afin de rendre ces ressources exploitables dans le cadre du web sémantique. le W3C [1] propose une panoplie de langages. Ces langages ont pour caractéristique d'être tous exprimables et échangeables en XML. Ils peuvent être vus comme des briques indépendantes, ou chaque langage vient répondre à des besoins d'expressivité spécifiques et peut satisfaire ses propres utilisations. Cependant, ces langages entrent dans le cadre d'une vision globale, chaque langage sert alors de support au langage du niveau supérieur, conduisant à une architecture sous forme d'empilement de couches, appelée "Semantic web Stack" (figure 1). Pour un maximum d'efficacité, les couches doivent répondre à différentes préconisations, actuellement seuls les niveaux du bas de la pile s'étageant jusqu'au langage OWL sont standardisés. Les niveaux supérieurs ne le sont pas encore et contiennent juste des idées et des principes qui ont fait l'objet d'implémentation partielle. Nous nous limitons à la présentation des langages RDF, RDFS et OWL qui servent de socle à l'expression de métadonnées et à leur structuration au travers de schémas dédiés. Ce n'est cependant pas leur seule fonction.

2.2 Un langage pour décrire les ressources du web

Le langage RDF (Resource Description Framework)[MM04], à voir plutôt comme un modèle de données, offre une manière simple d'enrichir une ressource par des métadonnées, sans obliger à modifier le contenu de la ressource ni même à accéder véritablement à la ressource. RDF s'appuie sur la notion de triplet ou déclaration {Sujet Propriété Objet} qui va permettre de décrire une ressource (le sujet de la

1. World Wide Web Consortium : www.w3.org

déclaration) au travers d'un couple propriété/valeur (respectivement l'élément de métadonnée et l'objet qui en est la valeur). Un modèle RDF est une collection de triplets et va constituer un graphe. Différents formats sont disponibles pour manipuler un graphe RDF : XML (RDF/XML), N3, Turtle ou encore JSON (JavaScript Object Notation).
Nous donnons un exemple de triplet : un individu, vu comme une ressource, et identifié par l'URI (Uniform Resource Identifier) `exple:individu_1` (définie par la concaténation de l'espace de noms et du nom qualifié) est décrit comme étant de type occurrence du schéma de méta-données Darwin Core. Le vocabulaire Darwin Core[WBG+12] facilite le partage d'informations relatives aux espèces vivantes. Ainsi les specimens identifiés et observés sur le terrain voire collectés au sein de laboratoires de recherche vont pouvoir être renseignés au sujet de différentes considérations : occurrence spatio-temporelle, traits phénotypiques, ...
Nous allons enrichir au fur et à mesure la description de la ressource individu_1 qui nous a été inspirée par le figuier située dans la cour du bâtiment IRD de Lavalette, et qui nous permet de mettre en oeuvre différents schémas de métadonnées provenant des sciences de l'environnement.

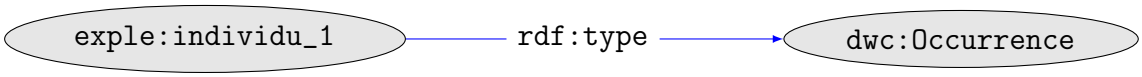

FIGURE 2 – Un exemple de déclaration RDF

Le specimen de figuier est décrit ci-dessous, au travers de la syntaxe N3 . L'espèce végétale concernée (Ficus carica) est référencée et la localisation géographique du figuier est notée au travers de la désignation d'un nœud anonyme (`_:blank1`).

```
<http://www.exemple.net/instance#individu_1>
        a dwc:Occurrence ;
        dcterms:created "2013-2-19" ;
        dcterms:spatial _:blank1 ;
        dwc:taxonConceptID
<http://www.ubio.org/authority/metadata.php?lsid=urn:lsid:ubio.org:namebank:448169> ;
```

2.3 RDFS

Le langage RDFS (Resource Description Framework Schema)[BG04] est un vocabulaire RDF, qui enrichit RDF au travers de différents éléments de modélisation. RDFS introduit notamment la notion de classe, et permet aussi d'organiser les classes et les propriétés au travers de hiérarchies . Dans notre exemple de specimen de figuier, individu_1 est enrichi par une couverture spatiale au travers de la propriété `dcterms:spatial`. `dcterms:spatial` étend la propriété `dcterms:coverage` de la manière suivante : `dcterms:spatial rdfs:subPropertyOf dcterms:coverage`. Il est donc rendu possible de s'intéresser, de manière générale, à la couverture spatiale et/ou temporelle d'un individu.

```
<http://www.exemple.net/instance#individu_1>
        a dwc:Occurrence ;
```

```
        dcterms:created "2013-2-19" ;
        dcterms:spatial _:blank1 ;
        dwc:taxonConceptID
<http://www.ubio.org/authority/metadata.php?lsid=urn:lsid:ubio.org:namebank:448169> ;

_:blank1
    a dcterms:Location ;
    geo:lon "3.8773" ;
    geo:lat "43.61092" ;
    dwc:locality "Montpellier" ;
    dwc:coordinateUncertaintyInMeters "10000" .
```

2.4 OWL

OWL (Web Ontology Langage)[DS04] est un langage de représentation des connaissances, qui vient s'adosser aux langages RDF et RDFS. Il introduit de nouveaux constructeurs qui viennent pallier les manques de RDF et de RDFS, dans la perspective de **la construction d'ontologies**. OWL emprunte de nombreux éléments de modélisation aux logiques de description. Il est ainsi possible de définir de nouvelles classes en appliquant des opérations d'union ou d'intersection sur des classes déjà définies, ou encore par restriction de rôles au moyen des quantificateurs universel et existentiel. OWL offre selon le sous-langage considéré (OWL Lite, OWL DL ou OWL Full [DS04]) différents compromis entre l'expressivité du modèle et la complexité et la décidabilité des mécanismes de raisonnement. En ce sens, OWL est tout indiqué pour structurer les éléments descriptifs (classes et propriétés) d'un schéma de métadonnées de façon formelle, et peut donc être utilisé pour traduire un consensus explicite sur la manière de décrire des métadonnées. Afin de mieux comprendre comment le langage OWL vient s'articuler au dessus de RDF et RDFS, et ainsi aborder les constructeurs qui seront rendus disponibles pour la spécification des éléments descriptifs des schémas de métadonnées. Nous présentons, dans un diagramme de classe UML, une portion du méta-modèle de OWL (figure 2).

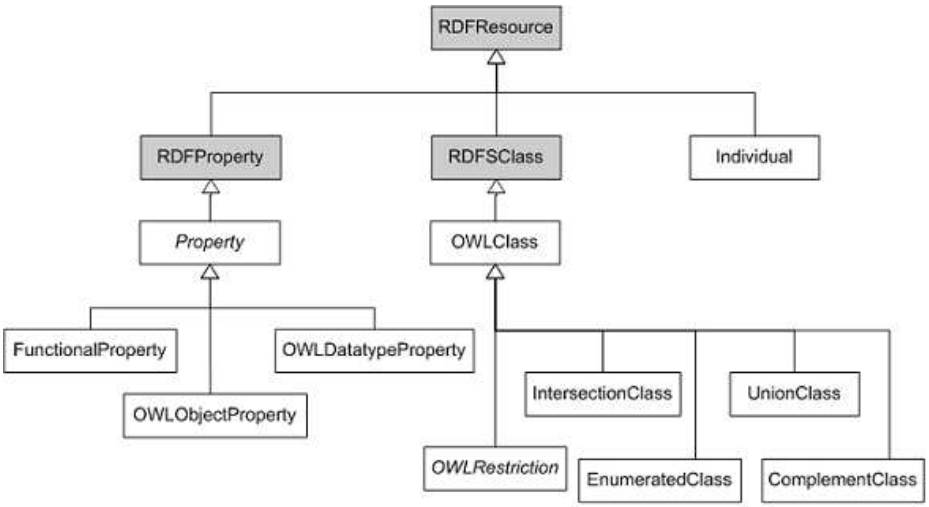

Figure 3 – Portion du Métamodèle OWL

2.5 Métadonnées et exploitation des ressources

Les langages décrits précédemment viennent véhiculer une vision globale du web sémantique. Ils fournissent les moyens nécessaires pour décrire toute ressource disponible, ou tout au moins référençable, au travers du web, et ce de manière normalisée. Disposer de langages commmuns pour doter les ressources de diverses descriptions structurées vues comme autant de métadonnées, est un premier pas vers le partage de ces descriptions et conduit à ce que l'on va pouvoir qualifier d'interopérabilité [2] syntaxique. Nous avons vu également que ces langages n'étaient pas conntraints au seul périmètre des métadonnées mais pouvaient également outiller la construction de vocabulaires contrôlés voire d'ontologies pour le partage et la raisonnement sur de la connaissance d'un domaine en particulier. Nous revenons plus en avant sur la notion de métadonnées dans le contexte du web sémantique, pour en préciser les grandes fonctionnalités. Ces fonctionnalités touchent à une exploitation rationnelle de multiples ressources pouvant être mises en commun par l'intermédiaire du web. Ces différents rôles sont tout d'abord centrés sur la ressource. Il va ainsi s'agir pour un ensemble de ressources donné :

- d'en aider à la découverte et à la localisation,
- d'en permettre la consultation des contenus,
- d'en proposer l'organisation et le partage,
- d'en faciliter l'archivage et la pérennisation
- d'en assurer la visualisation et les traitements ultérieurs
- d'en garantir la propriété intellectuelle
- ...

En se concentrant sur les métadonnées, plutôt que sur les ressources, les métadonnées apparaissent aussi comme un moyen d'enrichir voire de créer par composition, un contenu informationnel, puisque les métadonnées sont aussi des données et viennent donc abonder le socle des ressources. De plus, les métadonnées sont souvent obtenues lors de processus de co-construction avec une participation de nombreux usagers et sont le reflet d'un savoir et d'une capitalisation de connaissances collectives.
Pour illustrer notre propos, nous nous intéressons au rôle de recherche d'information et de consultation du contenu des ressources. Les métadonnées permettent d'indexer et de classifier le contenu des ressources. Un outil de recherche qui va se baser sur la structuration des métadonnées ayant servi à indexer les ressources, va permettre de retourner des résultats exhaustifs, fiables et précis.
Nous supposons que nous recherchons les rivières se situant dans le périmètre de la ville de Montpellier. L'utilisation d'un moteur de recherche classique, avec la requête "rivières de Montpellier" retourne des résultats non pertinents, et ce dès les premières pages de résultats. Ces résultats pointent, par exemple, sur des sites proposant des vacances ou des entreprises de recrutement dans Montpellier, sur des pages rédigées dans différentes langues et seulement quelques ressources sont vraiment associées aux rivières de Montpellier. Le recours à un outil de recherche se basant sur les métadonnées, permettrait de retourner des résultats plus satisfaisants au regard de la requête. Différents schémas de métadonnées peuvent être concernés : des spécifications généralistes comme le vocabulaire Dublin Core, qui vont permettre

2. Nous définissons ultérieurement l'interopérabilité entre systèmes

de renseigner le format de la ressource, le type bibliographique (page web, article de journal, de conférence, . . .), et/ou la langue ayant servi à la rédaction des ressources ; ou bien des standards de métadonnées dédiés comme Darwin Core, GeoML ou EML qui sont associés à une discipline en particulier : bidiversité, géographie ou écologie, . . .
Yannick Prié et Serge Garlatti[PG04] font état de l'utilisation d'un schéma de métadonnées appelé LOM+ pour composer dynamiquement des cours structurés au travers de ce schéma de métadonnées. En ce qui concerne la géographie, le standard de métadonnées ISO 19115 est largement exploité pour la description de ressources géoréférencées et va permettre de répondre aux besoins spécifiques d'aide à la découverte et de catalogage de ressources, qui s'avèrent être des besoins importants notamment dans le cadre des sciences de l'environnement.

3. Interconnexion de sources de données hétérogènes

Un rôle essentiel assuré par les métadonnées porte sur le partage des ressources et nécessite et nécessite de manière sous-jacente de se concentrer sur l'interopérabilité des systèmes qui disposent de ces ressources. Nous reprenons à notre compte, la définition ISO de l'intéropérabilité, qui peut être traduite par : "la capacité pour différents systèmes possiblement hétérogènes (en matière de système d'exploitation, plateforme logicielle, modèle de données et interfaces d'accès) d'échanger et de partager des données".
Nous allons présenter comment les standards de métadonnées peuvent être utilisées afin de garantir l'interopérabilité des systèmes, en nous appuyant sur les travaux de B.HASLHOFER et W. KLAS [HK10] qui proposent une synthèse détaillée des approches existantes. Nous avons choisi d'adopter leur vision de l'interopérabilité et de présenter leurs avancées dans le détail parce que leur synthèse nous a semblé originale et exhaustive par rapport à l'état des travaux actuels.
Comme nous l'avons mentionné au début de cette étude, les sciences de l'environnement et du vivant disposent d'un nombre important de ressources, décrites conformément à différents standards de métadonnées. Ces standards ont été créés pour répondre aux besoins de chaque communauté. Il paraît essentiel aujourd'hui de pallier cette hétérogénéité, pour pouvoir disposer d'accès transparents à toutes ses ressources, sans avoir à se soucier de leur localisation réelle.
Avant de présenter les techniques d'interopérabilité, Nous allons d'abord aborder les différents problèmes d'hétérogénéité pouvant affecter l'exploitation de métadonnées lors du partage de données. Ces problèmes sont analysés à différents niveaux d'abstraction et en considérant différents blocs de construction associés aux métadonnées. La figure 4 donne une illustration des différents niveaux distingués [HK10], cette classification des niveaux d'abstraction est inspirée du standard MOF (Meta Object Facilities) de l'OMG [3]. Toutefois ces niveaux ne prennent en considération que les métadonnées et non pas les données.

3. Object Management Group : www.omg.org/

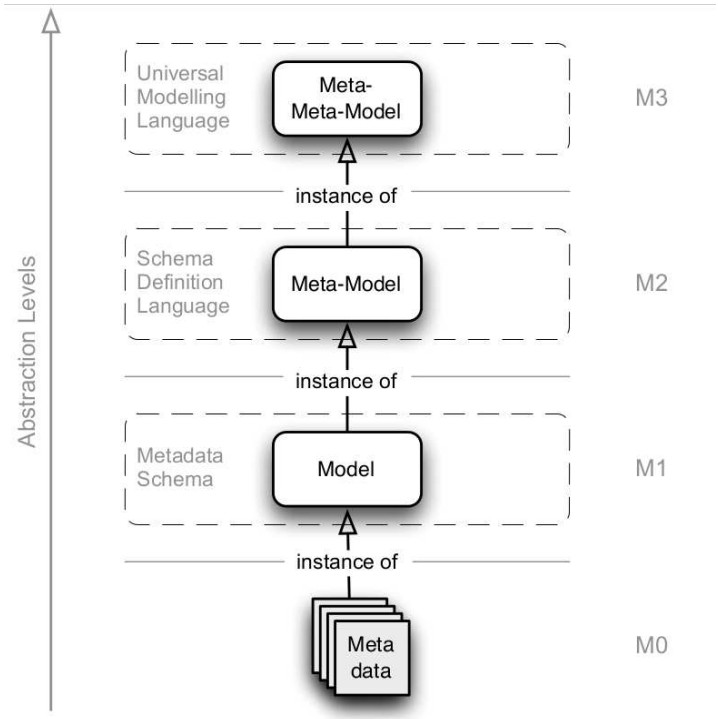

FIGURE 4 – Les différents blocs de construction des métadonnées suivant la vision MOF

Le niveau le plus haut de la figure correspond aux **langages de définition des schémas** : C'est à ce niveau que sont fixées les constructions syntaxiques des métadonnées et leurs définitions sémantiques. Il représente, en d'autres termes, le langage au travers duquel les schémas de métadonnées seront définis. La notation UML, ou le langage OWL se situent à ce niveau. En apparence, UML et OWL ne partagent pas la même vision, et ne sont pas utilisés pour répondre aux mêmes besoins, mais ils peuvent tout les deux être choisis comme langage de représentation de schéma de métadonnées. Et c'est dans cette optique qu'ils se situent tout les deux à ce même niveau d'abstraction.
Le niveau intermédiaire correspond au **schémas de métadonnées**. Ils sont constitués de collections d'éléments dotés d'une sémantique précise et organisant les métadonnées. C'est à ce niveau que porte le consensus sur le vocabulaire pour décrire les métadonnées. Ainsi, la définition des éléments des standards de métadonnées, à l'exemple du Dublin Core ou du Darwin Core sont retrouvés à ce niveau.
Le niveau le plus bas correspond aux **instances de métadonnées**. Les métadonnées et leurs valeurs décrivant les ressources, se retrouvent concrètement à ce niveau.
L'interopérabilité assurée par les métadonnées, est définie dans l'article, comme leur capacité à s'affranchir des frontières entre différents systèmes d'information, de manière à exploiter toute ressource décrite sans entrave liée au contexte ou au système qui l'a produite.

3.1 Hétérogénéité entre métadonnées

Les métadonnées sont donc envisagées sous l'angle des trois niveaux d'abstraction, et nous allons nous intéresser maintenant aux diverses sources possibles d'hétérogénéité et à leur projection sur ces différents niveaux. Cette hétérogénéité peut survenir à n'importe lequel des niveaux et se répercuter par la suite sur les autres niveaux puisqu'il y a une correspondance entre niveaux. Nous rappelons que dans la vision du MOF (figure 4), une relation d'instanciation se noue entre les différents niveaux, et de ce fait, les métadonnées sont l'instance de leur schéma de métadonnées, qui est lui même une instance du schéma des langages de définition.
Les auteurs distinguent deux grands types d'hétérogénéité : structurelle et sémantique.
L'hétérogénéité structurelle concerne l'hétérogénéité liée à des choix différents dans la manière de représenter les métadonnées, ou de définir les éléments de métadonnées. Cette hétérogénéité se trouve au niveau des langages de définition et des schémas de métadonnées. Elle englobe les conflits de définition, d'identification, de contrainte d'intégrité et d'abstraction sur les éléments, suivant le langage de définition choisi. Par exemple, un conflit peut apparaître quand un choix différent est fait dans la manière d'identifier les éléments : l'identification est assurée par le nom de l'élément uniquement ou bien par un identifiant qualificatif complet (espace de nom plus le nom de l'élément).
L'hétérogénéité sémantique est liée aux différences sémantiques au niveau des schémas des langages de définition et des schémas de métadonnées, ainsi qu'aux différences de représentation au niveau des instances de métadonnées. Cette hétérogénéité peut être due à des problèmes d'expressivité dans les langages de description utilisés ou survenir à la suite de l'emploi de différences terminologiques. Par exemple OWL dispose de constructeurs pour exprimer la disjonction entre deux classes, tandis que cette possibilité n'existe pas dans d'autres langages. Ce type d'hétérogénéité peut apparaître aussi au niveau des schémas de métadonnées. Par exemple, le standard Dublin Core a une visée généraliste, il n'est pas doté d'un vocabulaire pour exprimer explicitement des métadonnées spécialisées de l'environnement, comme le ferait le standard ISO 19115 par exemple. Au niveau des instances de métadonnées, l'utilisation de systèmes de mesures différents ou différentes représentations pour la même valeur entraîne aussi des hétérogénéités sémantiques.
Ces différents types d'hétérogénéité sont illustrés dans la figure 5 [HK10]. Cette illustration replace les sources d'hétérogénéité que nous venons d'aborder, sur les différents niveaux d'abstraction relatifs aux métadonnées.

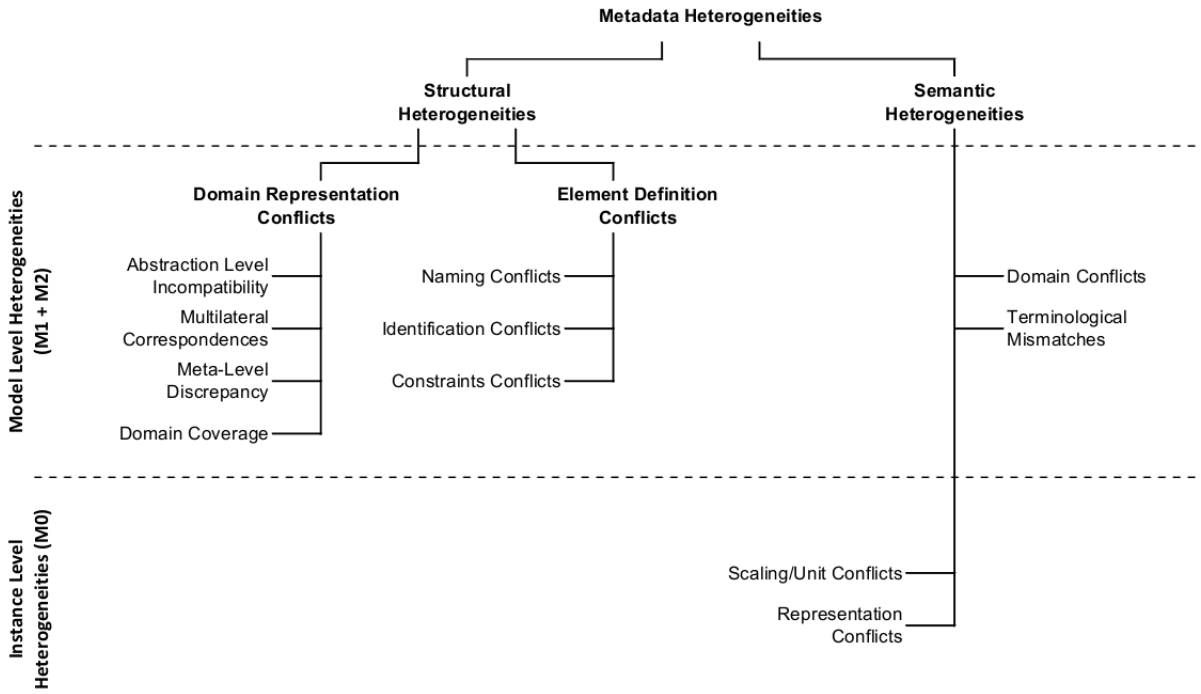

FIGURE 5 – Les différents types d'hétérogénéité des métadonnées par rapport au niveaux d'abstraction

3.2 Interopérabilité des métadonnées

Différentes techniques ont été développées pour mettre en pratique l'interopérabilité à partir des métadonnées. Dans cette partie, nous allons en présenter les plus répandues. D'un point de vue personnel, cela nous permet de nous positionner sur les choix possibles en terme de techniques applicables dans le contexte des sciences de l'environnement et la vie.

Nous commençons par **la standardisation**, appelée aussi **la normalisation** . Cette technique se fait graduellement sur les différents niveaux d'abstraction (Figure 4), respectivement le niveau des schémas de langages de définition, le niveau des schémas des métadonnées et le niveau d'instance des métadonnées. Elle consiste à commencer par adopter un schéma de langage de définition unique pour la description des métadonnées, poursuivi par un accord sur le schéma de métadonnées à utiliser (vocabulaire contrôlé), et peut aller au niveau des instances si un consensus sur les valeurs peut être établi (en définissant des listes ou des intervalles de valeurs possibles).

Une autre technique d'interopérabilité consiste à adopter un méta-modèle : **Meta-model agreement** [HK10]. Cette technique est adaptée pour les institutions qui adoptent déjà leurs propres standards et schémas de métadonnées. Il est en effet difficile d'appliquer un standard unique sur toutes les métadonnées. Dans les sciences de l'environnement et du vivant par exemple, un nombre important de ressources sont décrites avec des métadonnées s'appuyant déjà sur des standards. Il existe différentes approches pour mettre en œuvre cette technique. En adoptant par exemple l'approche du MOF [Gro06] ,ou en définissant un méta-modèle abstrait

référencé sur le niveau des schémas de langages de définition. On peut aussi aller vers un modèle conceptuel global abstrait. Cette technique reste relativement assez difficile à mettre en œuvre. Parce qu'il faut prendre en considération tous les standards employés pour décrire les métadonnées à rendre interopérable. Mais elle permet théoriquement de pallier efficacement l'hétérogénéité et fournir un moyen efficace d'interopérabilité. Nous allons dans la partie suivante présenter le standard Dublin Core et son utilisation pour la mise en œuvre de profils d'application. C'est une approche qui s'inspire beaucoup de cette technique.
Une autre technique pour l'interopérabilité est le **Mapping de Métadonnées**, cette technique englobe une sous-technique de "mapping" de schémas et une autre pour la transformation d'instances, le "mapping" de métadonnées pourrait être découpé en quatre étapes majeures (figure 6) : La première consiste à trouver les relations sémantiques et structurelles entre les éléments des deux schémas. La deuxième étape est la phase de formalisation des relations trouvées. La troisième permet d'établir ces relations. La dernière consiste à la maintenance possiblement évolutive de ces relations sur le long terme.

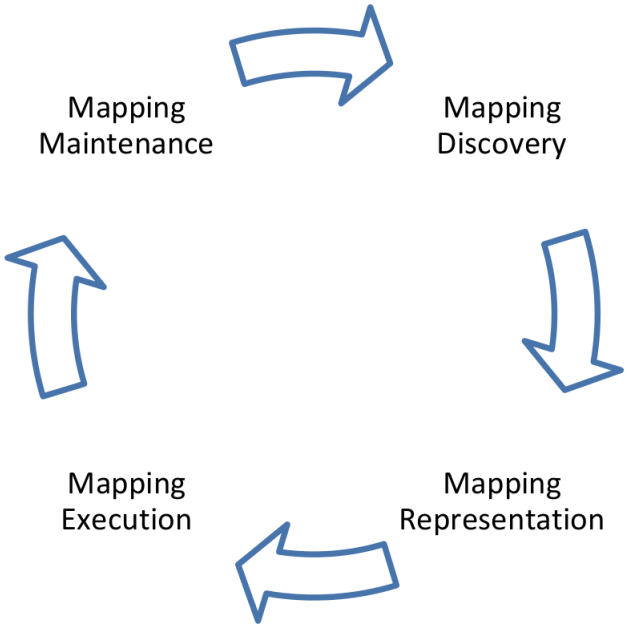

FIGURE 6 – Les quatre étapes majeures de la technique de "mapping" de métadonnées[HK10]

En présentant toutes ces techniques, nous démontrons qu'il existe plusieurs solutions pour aborder l'interopérabilité. Mais néanmoins, le choix d'utilisation d'une de ces techniques doit être fait en étudiant le contexte d'application ainsi que d'autres facteurs. B.HASLHOFER et W. KLAS par exemple n'en retiennent que les techniques **de standardisation** et **de "mapping" des métadonnées**. Car d'après les comparaisons qu'ils ont effectué, seules ces deux techniques couvrent des champs d'hétérogénéité assez larges, et ont l'avantage de pouvoir être utilisées efficacement avec d'autres techniques d'interopérabilité sur le niveau d'instance de métadonnées

pour couvrir toutes les formes d'hétérogénéité (figure 5). Un autre paramètre important intervient dans le choix de la technique d'interopérabilité à utiliser, c'est le coût de la mise en œuvre des techniques. C'est un élément clé et à cet effet, HASLHOFER et W. KLAS [HK10] présentent une analyse des facteurs de coût entre la standardisation et le "mapping" des métadonnées afin d'aider les experts à faire leurs choix.

4. Le standard Dublin Core et les profils d'application

Nous allons nous centrer dans cette section, sur une variante de ce que HASLHOFER et W. KLAS [HK10] désignent par technique du "meta-model agreement". Nous pensons en effet que la méta-modélisation est une piste à explorer. L'initiative Dublin Core (DCMI Dublin Core Metadata Initiative) propose dans ce contexte le modèle abstrait du Dublin Core (DCAM) qui offre une représentation générique de toute ressource décrite par un ensemble de métadonnées. L'idée est de compléter les standards de métadonnées par différentes lignes directrices facilitant la mise en place d'applications tirant parti de métadonnées. Nous revenons dans un premier temps sur le standard Dublin Core.

4.1 Le standard Dublin Core

Avec l'émergence du web sémantique, différents standards de description ont pris une place prépondérante et constituent aujourd'hui des références incontournables pour la production de métadonnées. Ces standards peuvent être distingués en deux types : des **standards spécialisés** pour des domaines de connaissances bien précis, à l'exemple du standard ISO 19115 principalement dédié aux métadonnées géographiques ; et des standards généralistes, appelés **standards de découverte**, visant la description des ressources de manière générale. Ces derniers standards sont souvent constitués d'un ensemble restreint d'éléments. Parmi les plus utilisés, nous retrouvons Dublin Core, c'est un standard générique, visant à décrire des ressources de type quelconque. Dublin Core s'appuie sur quinze éléments de description officiels pour décrire les ressources (Figure 7).

4.2 DCAM : le modèle abstrait du Dublin Core

Le modèle abstrait du Dublin Core [PN06] établit une différence nette entre la ressource décrite et les métadonnées venant décrire cette ressource et intègre, à cet effet, deux modèles nommés "DCMI [4] resource model" et "DCMI description model". La documentation mise en ligne par le DCMI [5] fait état de différents diagrammes de classes UML qui facilitent la compréhension du modèle générique, qui en réalité comprend plusieurs modèles qui se complètent. Un diagramme de classes est également rendu disponible pour la typologie de tous les termes gérés au travers

4. DCMI : Dublin Core Metadata Initiative
5. http ://dublincore.org/documents/abstract-model/

Élément Dublin Core	**Description de l'élément**
Title	Titre principal de la ressource
Creator	Personne morale ou physique à l'origine de la création de la ressource
Subjet	Sujet principal de la ressource ,Mots-clefs ou principales idées
Description	Description du contenu de la ressource décrite
Publisher	Nom de la personne morale ou physique à l'origine de la publication de la ressource
Contributor	Nom d'une personne physique ou morale qui contribue ou a contribué à l'élaboration de la ressource. Chaque contributeur fait l'objet d'un élément Contributor séparé
Date	Date d'un évènement dans le cycle de vie de la ressource
Type	Genre du contenu de la ressource
Format	Type MIME, ou format physique de la ressource
Identifier	Identificateur non ambigu, référençant la ressource dans un contexte donné
Source	Ressource dont dérive la ressource décrite, que ce soit en totalité ou en partie de la ressource en question
Langage	Langue du contenu de la ressource
Relation	Lien avec d'autres ressources.
Coverage	Couverture spatiale (point géographique, pays, régions, noms de lieux) ou temporelle
Rights	Droits de propriété intellectuelle au sujet ou sur la ressource

FIGURE 7 – Dublin Core Metadata Element Set

de l'initiative "DCMI", incluant classes, propriétés, schémas de vocabulaire et de syntaxe. L'ensemble des spécifications sont par par ailleurs disponibles au format rdf/xml. Les attentes du DCMI sont claires : il ne s'agit pas seulement de fournir un schéma de données généralistes mais aussi de fournir des facilités pour l'exploitation de métadonnées dans tout contexte d'application. Ainsi tous les besoins sont pris en considération : schéma de métadonnées, schéma de métadonnées qualifié pour aller plus dans loin dans la structuration des éléments de métadonnées mais aussi dans le contrôle des valeurs prises par ces éléments au travers de vocabulaires contrôlés appropriés et de formats d'encodage adéquats, exploitation conjointe de différents schémas de métadonnées, ... Nous proposons, dans la Figure 8, le diagramme de classes UML décrivant une ressource ("DCMI resource model"), qui exprime, de façon très générale, l'enrichissement d'une ressource au travers de collection de métadonnées. Un tel cadre conceptuel peut servir de socle à l'établissement de références croisées entre métadonnées provenant de différents schémas (mapping de métadonnées ou "crosswalks") ou bien à la définition de profils d'application (approche "mix and match") qui consistent à faire le choix d'éléments de métadonnées pertinents provenant de différents schémas de métadonnées.

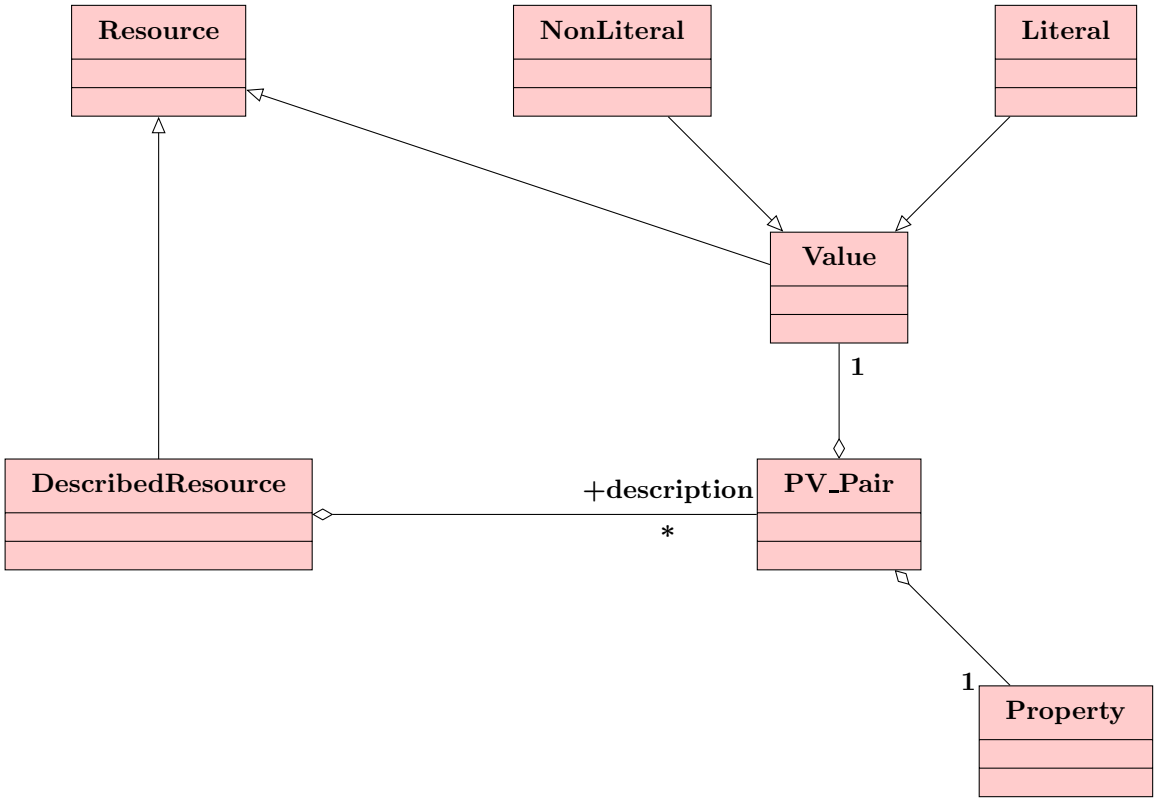

FIGURE 8 – DCMI resource model

Le modèle DCMI spécifie les règles suivantes :
- Chaque ressource est décrite en utilisant une ou plusieurs paires propriété-valeur.
- Chaque paire propriété-valeur est composée d'une et une seule propriété et une et une seule valeur.

- Chaque valeur est une ressource : physique, entité numérique ou conceptuelle ou littérale qui est associé à une propriété lorsqu'un couple propriété-valeur est mis à contribution pour décrire une ressource. Par conséquent, chaque valeur est soit une valeur littérale ou une valeur non littérale.
- Une valeur non littérale est une valeur qui est une entité physique, numérique ou conceptuelle.
- Un littéral est une entité qui utilise une chaîne Unicode en tant que forme lexicale, avec une étiquette de langue facultative ou type de données, pour désigner une ressource.

4.3 Profils d'application

Nous avons vu que le modèle abstrait Dublin Core ne mentionne pas les types de ressources qui peuvent être décrites. Il possède une capacité à décrire de façon générique tout type de ressource. Il peut donc être utilisé comme un préalable à une exploitation conjointe de différents modèles et standards de métadonnées, afin d'implémenter un cadre ("framework") applicatif global. Cette approche fait appel à ce qui est désigné par **profil d'application** et constitue une variante de la technique de **Meta-model agreement** présentée dans la section précédente 3.2.
D'après Heery et Patelles [HP00], les profils d'application sont des schémas de métadonnées constitués d'un ensemble d'éléments provenant d'un ou plusieurs standards. Ces éléments sont combinés de façon optimisée pour répondre aux besoins fonctionnels d'un domaine d'application en particulier(Figure 9).

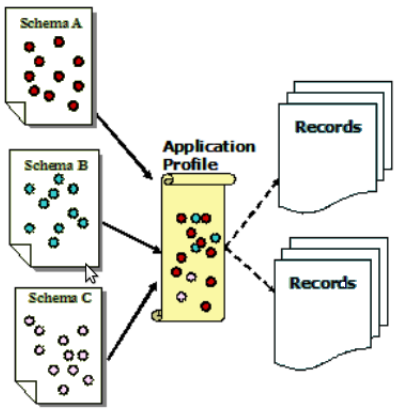

FIGURE 9 – Profil d'application constitué d'éléments provenant de divers schémas de métadonnées [CZ06]

En effet, l'approche des profils d'applications consiste à constituer un schéma de métadonnées en utilisant des éléments d'un ou plusieurs standards existants. Le modèle construit est ensuite optimisé pour répondre aux besoins spécifiques d'utilisation d'une application précise.
Les profils d'application se distinguent par les caractéristiques suivantes [HP00] :

- Les profils d'application peuvent être constitués à partir d'un ou plusieurs schémas de métadonnées.
- Les profils d'application n'introduisent pas de nouveaux éléments. tout les éléments appartiennent à d'autres schémas. Dans le cas ou un nouveau élément non existant doit être introduit, il faut créer un nouveau espace de nom et l'importer sur le modèle.
- Les profils d'application peuvent spécifier les valeurs que les éléments du modèle peuvent prendre.
- Les profils d'application peuvent raffiner [6] les éléments des standards utilisés.

5. Conclusion et futurs travaux

L'étude bibliographique conduite, avait pour ambition de poser un état de l'art autour des métadonnées et de leur capacité à apporter des solutions en matière d'interconnexion de sources de données hétérogènes dans le cadre du web sémantique. Notre point de départ a consisté en un ensemble d'articles de publications scientifiques en anglais et en français qui nous ont permis de mieux comprendre le contexte de l'étude et de saisir les différentes facettes du sujet. Une des difficultés a justement porté sur les multiples facettes des métadonnées qui jouent de nombreux rôles et qu'il n'est pas toujours simple de caractériser au mieux. Les métadonnées font actuellement l'objet de nombreux travaux, notamment dans le contexte du web sémantique et nous nous sommes efforcés d'établir une synthèse des avancées les plus significatives, afin de nous positionner sur les tâches à réaliser pour répondre aux objectifs du stage.
Nous avons donc présenté dans un premier temps l'infrastructure du web sémantique, expliqué ses usages et les différents langages qui en constituent le socle. Nous avons poursuivi en montrant les rôles clefs que pouvaient jouer les métadonnées dans l'interconnexion des sources de données hétérogènes. Nous avons par la suite présenté les différentes solutions qui s'offraient à nous pour pallier l'hétérogénéité qui se fait jour lorsqu'il s'agit de partager des ressources. Une motivation forte est de garantir un accès efficace et complet aux différentes ressources disponibles sur le web pour ce qui concerne les sciences du vivant et de l'environnement.
Au terme de cette étude, les enseignements en sont que la modélisation et la métamodélisation de métadonnées sont un préalable à la mise en place d'une approche visant à l'interconnexion de différents sources de données hétérogènes. Les travaux qui doivent être entrepris dans un premier temps devront porter tout d'abord sur une étude détaillée des standards des sciences de l'environnement existants (ISO 19115, Darwin Core, EML, FGDC CSDGM, Water ML, ...). Différentes approches peuvent être adoptées : "mapping" de métadonnées, profiles d'application, registres de métadonnées, schémas de métadonnées dérivés, ... Nous souhaitons nous orienter vers une démarche empruntant aux profils d'application et au "mapping" de métadonnées. Il nous semble en effet judicieux de nous adosser au socle défini par le DCMI (vocabulaires Dublin Core et modèle abstrait DCAM) et de faire un choix

6. Un raffinement restreint la signification d'un élément, mais sans la changer fondamentalement.

parmi les éléments de métadonnées des standards spécialisés précédemment cités, afin de définir un profil d'application adapté. Différents travaux autour des profils d'application sont là encore disponibles comme le "framework" de Singapour, toujours à l'initiative du Dublin Core qu'il nous faudra nous approprier. L'objectif est ensuite de définir un composant logiciel qui va s'articuler sur le profil d'application afin de proposer des conversions de description de ressources d'un standard spécialisé à un autre et ainsi faciliter le partage de ressources à l'origine décrités au travers de différents standards de métadonnées.

Références

[BG04] Dan Brickley and R.V. Guha. RDF vocabulary description language 1.0 : RDF schema. W3C recommendation, W3C, February 2004. Published online on February 10th, 2004 at `http://www.w3.org/TR/2004/REC-rdf-schema-20040210/`.

[CBT04] Jean Charlet, Bruno Bachimont, and Raphaël Troncy. Ontologies pour le web sémantique. *Revue I3 Information-Interaction-Intelligence, Numéro Hors-série Web sémantique, Edition Cépaduès*, 30(6) :69–99, 2004.

[CZ06] Lois Mai Chan and Marcia Lei Zeng. Metadata Interoperability and Standardization - A Study of Methodology, Part I : Achieving Interoperability at the Schema Level. *D-Lib Magazine*, 12(6), 2006.

[DHSW02] Erik Duval, Wayne Hodgins, Stuart Sutton, and Stuart L. Weibel. Metadata principles and practicalities. *D-Lib Magazine*, 8(4), 2002.

[DS04] Mike Dean and Guus Schreiber. OWL Web Ontology Language - Reference. W3C recommendation, W3C, 10 feb 2004.

[Gro06] OMG : Object Management Group. Meta object facility (MOF) core specification. MOF Specification 2, OMG, Needham Heights, USA, January 2006.

[HK10] Bernhard Haslhofer and Wolfgang Klas. A survey of techniques for achieving metadata interoperability. *ACM Comput. Surv.*, 42(2) :1–37, March 2010.

[HP00] Rachel Heery and Manjula Patel. Application profiles : mixing and matching metadata schemas. *AGI - Information Management Consultants*, 10(1) :1–10, January 2000.

[LRC02] Philippe Laublet, Chantal Reynaud, and Jean Charlet. Sur quelques aspects du web sémantique. In *2ème Assises Nationales du GdR I3, Nancy, France*, December 2002.

[MM04] Franck Manola and Eric Miller. RDF primer. W3C recommendation, World Wide Web Consortium, February 2004. `http://www.w3.org/TR/rdf-primer/`.

[NJNP06] Mikael Nilsson, Pete Johnston, Ambjörn Naeve, and Andy Powell. Towards an interoperability framework for metadata standards. In *Proceedings of the 2006 international conference on Dublin Core and Metadata Applications : metadata for knowledge and learning*, DCMI '06, pages 13–23. Dublin Core Metadata Initiative, 2006.

[PG04] Yannick Prié and Serge Garlatti. Métadonnées et annotations dans le web sémantique. *Revue I3 Information-Interaction-Intelligence, Numéro Hors-série Web sémantique, Edition Cépaduès*, 24(6), 2004.

[PN06] Sarah Pulis and Liddy Neville. Using the DC Abstract Model to support application profile developers. In *DC-2006 : International Conference on Dublin Core and Metadata Applications, Colima (Mexico), 3-6 October 2006*, pages 227–234, Colima, Mexico, October 2006.

[SBLH06] Nigel Shadbolt, Tim Berners-Lee, and Wendy Hall. The semantic web revisited. *IEEE Intelligent Systems*, 21(3) :96–101, May 2006.

[WBG+12] John Wieczorek, David Bloom, Robert Guralnick, Stan Blum, Markus Döring, Renato Giovanni, Tim Robertson, and David Vieglais. Darwin core : An evolving community-developed biodiversity data standard. *PLoS ONE*, 7(1) :e29715, 01 2012.

[ZC06] Marcia Lei Zeng and Lois Mai Chan. Metadata interoperability and standardization : A study of methodology. Part II : Achieving Interoperability at the Record and Repository Levels. *D-Lib Magazine*, 12(6), 2006.