FROM RAW PUBLISHED DATA TO INTERLINKED SEMANTIC DATA

Author	Pierre-Yves Vandenbussche (Mondeca) Bernard Vatant (Mondeca)
Reviewer	Jérôme David (INRIA-EXMO)
Date	2011-12-28
Reference	Datalift/2011/D2.1/v2
Version	2.0

PROJET DATALIFT

De la donnée brute publiée vers la donnée sémantique interconnectée

Appel ANR CONTINT 2010 ANR-10-CORD-009

RAPPORT DE RECHERCHE

DATALIFT D2.1
METHODES ET INDICATEURS POUR LA SELECTION D'ONTOLOGIES FIABLES ET UTILISABLES

Auteurs: Bernard Vatant (Mondeca)
Pierre-Yves Vandenbussche (Mondeca)
Contributeurs: Lise Rozat (Mondeca)
Bénédicte Bucher (IGN)
Nathalie Abadie (IGN)

Relecteur: Jérôme David (INRIA-EXMO)

Sommaire

Introduction

Le processus de “lifting” des données (conversion, publication et interconnexion) s’appuie sur des vocabulaires contrôlés possédant une sémantique formelle, en d’autres termes des *ontologies* - appelés simplement ci-après des *vocabulaires,* ce choix terminologique étant expliqué plus bas. Ces vocabulaires définissent des types (ou classes) et des propriétés permettant la *description des ressources*, c’est-à-dire des données, mais aussi et surtout des concepts et objets du monde décrits par les données. Le mode de représentation sous-jacent à ce type de description est défini par le standard du W3C Resource Description Framework [RDF], qui fournit comme son nom l’indique un “cadre de description des ressources”. Les vocabulaires (ontologies) auxquels on s’intéresse ici sont exprimés dans des langages d’ontologie standards basés sur RDF, à savoir [RDFS] ou [OWL]. La sélection de tels vocabulaires est une étape critique dans le processus qui nous intéresse, car ce sont eux qui vont structurer les données, préciser leur sémantique, permettre des contrôles d’intégrité logique et des inférences, assurer la liaison et l’interopérabilité avec d’autres données.

Le présent document vise à fournir un **support méthodologique au fournisseur de données** dans son travail de sélection au sens large, qu’il s’agisse de sélectionner et de réutiliser en l’état un vocabulaire “sur étagère”, ou de l’adapter aux besoins spécifiques de son domaine, ou encore de le créer quasi-totalement si cela s’avère nécessaire. Pour mémoire, la tâche correspondante du projet (T2.1) est spécifiée dans le document de soumission de la façon suivante :

Cette tâche s’appuiera sur l’état de l’art en matière de méthodes et d’outils pour l’évaluation et la sélection d’ontologies, et l’adaptera aux objectifs spécifiques du projet. Une attention particulière sera attachée à l’adaptation des ontologies au domaine et à la structure des données à publier, la capacité à supporter l’interconnexion des données, la fiabilité et la persistance des autorités de publication, la conformité aux standards et bonnes pratiques du web sémantique.

Les méthodes et indicateurs seront sélectionnés de façon itérative, par évaluation de leur efficacité pour la construction du catalogue définie par la tâche 2.2, les tâches de conversion de données (WP3) et d’interconnexion (WP4).

Les principaux critères de sélection et d’évaluation sont clairement énoncés dans cette spécification comme des **critères fonctionnels**. Il s’agit de permettre au fournisseur de données de disposer des outils les mieux adaptés à la tâche visée in fine : publication et interconnexion des données. Bien sûr la représentation correcte et cohérente des objets du domaine métier reste un critère incontournable, mais le contexte particulier d’utilisation place en amont de cette exigence quelques critères d’éligibilité spécifiques à l’univers des données liées.

Un pré-requis minimal est que les vocabulaires utilisés pour décrire les données liées appartiennent à la même pile technologique que ces données elles-même, à savoir la famille des standards du web sémantique basés sur RDF [SW].

Nous avons donc mis en tête de liste des critères de sélection le fait que **les vocabulaires soient publiés dans les mêmes formats de données, et facilement accessibles suivant les mêmes protocoles, que les données qu’ils décrivent**. La quantité et la qualité des données publiées se sont accrues ces dernières années dans ce qu’il est convenu d’appeler le Linked Data Cloud [LOD] [Bizer et al, 2009], mais encore trop de ces données utilisent des vocabulaires difficiles, voire impossibles à identifier ou retrouver à partir des données qui

les utilisent, des espaces de noms mal ou pas documentés, des sémantiques non explicitées. Quand les vocabulaires existent et sont publiés, les métadonnées qui les décrivent (date de création et de mise à jour, organisme de publication …) sont elles aussi de qualité variable. Il apparaît donc prioritaire d'examiner les points suivants :

- Dans quelle mesure les vocabulaires utilisés pour publier les données sont eux-mêmes publiés et intégrés dans le Web comme des données liées, ainsi que les différents critères à prendre en compte pour juger de la qualité de ce liage.
- La présence et la qualité des métadonnées disponibles sur les vocabulaires, globalement et/ou sur chacun de leurs éléments. On détaillera les métadonnées indispensables ou souhaitables, et les vocabulaires disponibles pour l'expression de telles métadonnées.

Bien sûr il ne s'agit pas d'ignorer d'excellents vocabulaires qui ne sont pas publiés ou accessibles selon les standards et protocoles du Web de données, et qui sont utilisés pour structurer d'excellents jeu de données liées. On en présentera un exemple typique un peu plus bas. A cet égard la tâche de la communauté du Web de données est d'inciter fortement les éditeurs de tels vocabulaires à les migrer vers les formats adéquats et à les publier de même. On verra que la documentation et les outils permettant une telle migration ne manquent pas.
Une fois les vocabulaires découverts, il reste à évaluer la qualité de leur contenu et la pertinence de leur utilisation dans le domaine des données à "lifter". Cette évaluation sera d'autant plus facile que les vocabulaires seront *simples, compacts, et dédiés* à une tâche de représentation ciblée (domaine métier et/ou domaine fonctionnel clairement défini).
Au-delà des fournisseurs de données, nous espérons que le présent travail sera également utile aux éditeurs de vocabulaires en leur indiquant des pistes pour améliorer la qualité des vocabulaires existants ou futurs afin d'optimiser leur utilisation dans le Web des données.
Enfin, même si la présente tâche se situe en amont de l'élaboration du catalogue qui constitue le livrable 2.2, nous avons de fait procédé à l'élaboration et à la structuration d'un noyau de ce catalogue en appui de l'élaboration des critères, bien sûr sous forme de données liées. C'est le projet Linked Open Vocabularies présenté dans la dernière section [LOV].

1. Ontologies ou vocabulaires?

Attachons-nous d'abord à éclaircir le choix terminologique évoqué en introduction. Le terme "ontologie" est consacré dans les langages du Web sémantique pour désigner un vocabulaire définissant des classes d'objets, les propriétés attachés à ces classes et objets, et les contraintes les reliant. Des définitions plus précises peuvent être trouvées dans les travaux de [Gruber, 1993][Uschold et al., 1996][Charlet, 2002]. Une ontologie pour le Web sémantique est exprimée dans un langage standardisé par le W3C, en général une variante ou « espèce » de OWL, ou de son prédécesseur moins expressif RDFS. Le langage OWL en particulier constitue une adaptation de la famille des langages formels issus des logiques de description [DL] à l'environnement technique du Web (protocole http, syntaxe XML).
Les travaux qui ont conduit à OWL ont de fait réuni des acteurs venus du monde de la logique formelle et de l'intelligence artificielle d'une part, et des technologies du Web d'autre part. Une collaboration pas facile a priori, finalement fructueuse en dépit de débats et de

difficultés de compréhension entre ces deux mondes qui demeurent vivaces encore aujourd'hui, mais dont les spécialistes des vocabulaires contrôlés "classiques" mais non formels (thésaurus, classifications …) ont été longtemps absents. Conséquence peut-être de cette absence des spécialistes des vocabulaires "naturels", en RDFS comme en OWL, le nommage en langage naturel des éléments d'une ontologie (classes et propriétés), s'il est possible et même recommandé, n'est pas indispensable et ne participe pas de la sémantique formelle. Les identifiants des classes et propriétés sont des URI, et les libellés en langage naturels apparaissent en général sous forme d'annotations de type rdfs:label, la plupart du temps en anglais, que le marqueur de langue soit explicite ou non. Rares sont les ontologies utilisant des libellés dans une autre langue, plus rare encore celles qui utilisent les potentialités de SKOS (prefLabel, altLabel) pour indiquer des libellés préférentiels ou alternatifs, dans une ou plusieurs langues.

Cependant, dans la recherche d'une ontologie adaptée à son domaine et à ses données, et dans sa compréhension par un utilisateur humain, le nommage de ses éléments en langage naturel, et si possible en plusieurs langues, est un aspect important dans la sélection. Si la sémantique d'une ontologie OWL est formellement décrite pour les machines, cette sémantique doit aussi pouvoir être approchée par les utilisateurs humains à travers cette interface qu'est le langage naturel. La sémantique formelle doit être autant que possible cohérente avec le sens de leur nom dans le langage naturel.

En bref, le fournisseur de données en quête d'une ontologie adaptée à ses besoins recherche d'abord un *vocabulaire* permettant de décrire ses données. Si ses données sont géographiques, il cherchera sans doute d'abord par exemple si des éléments existent qui portent les libellés "territoire", "pays", "frontière", "population" …avant de vérifier si la définition formelle sous-jacente correspond à la conception de ces concepts que véhiculent ses données. On utilisera donc volontiers le terme "ontologie" si on considère la construction dont on parle comme une structure de classes et de propriétés dont la sémantique formelle est déclarée, dans laquelle les libellés en langage naturel sont optionnels, et dont la fonctionnalité majeure est de supporter des raisonnements logiques. Mais on préférera nommer "vocabulaire" cette même construction si on la considère comme composée de termes du langage naturel dont la sémantique a été formalisée en la "fixant" sur des classes et propriétés liées par des axiomes. Il s'agit donc plus d'une différence de point de vue que d'une différence de nature. On peut parler ici en employant une terminologie venant du monde de l'extraction de concepts et du traitement automatique de la langue, de « ressources termino-ontologiques ».

De plus, même si certains de ces vocabulaires contiennent un certain nombre d'instances (listes contrôlées) on s'intéresse ici essentiellement aux classes et propriétés, autrement dit au niveau appelé "T-Box" dans les logiques de description, le niveau "terminologique" donc.

La pratique du Web sémantique confirme cette tendance à employer souvent "vocabulary" plutôt que "ontology", surtout quand la taille du dit vocabulaire reste modeste. Le standard RDFS lui-même se définit comme un "langage de description des vocabulaires" (RDF Vocabulary Description Language). Même si on dit plus volontiers un vocabulaire RDFS et une ontologie OWL, la frontière n'est pas franche, d'autant que des vocabulaires initialement en RDFS (comme FOAF) évoluent vers plus d'expressivité en introduisant quelques constructions OWL, ou proposent une version OWL. Si on distingue parfois les ontologies légères (light-weight ontologies) par opposition aux ontologies lourdes (full-fledged ontologies), une telle distinction ne recouvre pas exactement celle dont il est question ici, et d'ailleurs n'a pas de définition standard. On considère généralement que les ontologies légères sont essentiellement constituées d'une hiérarchie de concepts (classes et propriétés)

alors que les ontologies lourdes utilisent une expressivité plus forte (classes construites, restrictions, règles d'inférence …).
Un document du W3C consacré à la question [W3C Voc] se garde bien d'ailleurs de trancher le débat, puisque le titre du document est "Ontologies" dans la balise HTML <title> et dans le menu W3C >> Standards >> Semantic Web >> Ontologies, mais le titre est bien "VOCABULARIES" dans le corps du document, et on peut y lire :

> *There is no clear division between what is referred to as "vocabularies" and "ontologies". The trend is to use the word "ontology" for more complex, and possibly quite formal collection of terms, whereas "vocabulary" is used when such strict formalism is not necessarily used or only in a very loose sense.*

A noter que le W3C inclut ici les vocabulaires publiés au format SKOS, c'est-à-dire les vocabulaires de types thésaurus. De tels « vocabulaires de valeurs » (*value vocabularies*), pour reprendre la terminologie retenue par le W3C LLD Incubator Group [VocD] n'entrent pas dans le cadre de notre étude [FLG1] dans un premier temps. Cependant un certain nombre de critères introduits dans la suite s'appliquent sans doute à de tels vocabulaires. Dans le contexte des données liées, les concepts de thésaurus doivent plutôt être considérés comme des ressources de référence (A-Box), utilisables comme valeurs de propriétés de catégorisation ou d'indexation, telles que sujet d'indexation ou catégorie d'un document (dcterms:subject), domaine de recherche d'un organisme ou centre d'intérêt d'une personne (foaf:topic_interest).

Enfin, un dernier argument de type marketing : même s'il commence à être un peu plus familier, le mot "ontologie" reste un peu intimidant, alors que "vocabulaire" a un air connu qui rebutera moins ceux qui abordent le domaine pour la première fois.

2. Quelques principes généraux

Avant d'entrer dans le détail des critères d'évaluation, qualitatifs et quantitatifs, nous examinerons dans cette section comment les principes généraux du Web des données s'appliquent aux ontologies ou vocabulaires utilisés pour la représentation des données.

2.1. Le lifting commence par les vocabulaires

Formaliser la sémantique d'un vocabulaire initialement en langage naturel est en soi un processus de "lifting", puisqu'il s'agit transformer un ensemble de données semi-structurées mais sans sémantique formelle (des termes du langage naturel) en données liées (des concepts reliés par des relations). Dans la construction d'un vocabulaire formel, on distingue déjà différentes étapes d'un processus de "lifting"

- Inventaire des données : liste de termes pertinents, à partir de glossaire, dictionnaires et autres vocabulaires contrôlés ou non
- Identification des concepts qui sont représentés par les termes, construction d'URI
- Structuration : quels termes désignent des classes, et quels termes des propriétés
- Liaison par formalisation des relations entre concepts : hiérarchie de classes et de propriétés, attachement des propriétés aux classes par des contraintes de domaine et co-domaine etc.

2.2. Ce qui est utilisé est ré-utilisable

L'histoire du Web des données est encore jeune puisqu'on peut dater l'origine de l'expression "Linked Data" à la note originale de Tim Berners-Lee [Berners-Lee, 2006], mais le nombre de jeux de données et de vocabulaires déjà publiés est néanmoins suffisant pour qu'on puisse s'appuyer sur un état de l'art et essayer de dégager des tendances. Une approche naturelle et pragmatique est donc d'examiner quels sont les vocabulaires d'ores et déjà utilisés par les jeux de données existants. Utiliser un vocabulaire largement partagé est en effet un critère important de choix dans une perspective d'interopérabilité.[FLG1] Or, il est souvent difficile pour un utilisateur potentiel d'estimer l'importance de l'utilisation existante d'un vocabulaire dans les jeux de données du LOD, ou de sa réutilisation par d'autres vocabulaires. Il est même à peu près évident que l'auteur ou l'éditeur eux-mêmes n'ont en général qu'une faible connaissance de ce genre d'information, mais en tout état de cause, elle apparaît rarement là où on aimerait « naturellement » la trouver, c'est-à-dire en annexe à la publication du vocabulaire, et si possible sous forme d'éléments de description formelle. Les vocabulaires développés spécifiquement pour la publication d'un jeu de données particulier peuvent parfois indiquer la référence de ce dernier et en extraire, avec des explications, des exemples « pédagogiques » d'utilisation. Mais ce n'est pas loin s'en faut le cas de la majorité des vocabulaires, et de toute manière l'utilisateur potentiel n'a pratiquement jamais de visibilité sur la réutilisation du vocabulaire en dehors de son contexte d'origine.

C'est plutôt du côté des catalogues de données qu'il faut pour le moment aller chercher des indications sur l'utilisation des vocabulaires. On peut s'appuyer par exemple sur les services fournis par l'initiative Comprehensive Knowledge Archive Network [CKAN]. CKAN rassemble les métadonnées et les descriptions de jeux de données ouvertes publiées sur le Web, et en particulier rassemble sous des "tags" les jeux de données utilisant un vocabulaire. On peut d'ores et déjà tenter de comprendre ce qui fait le succès d'un vocabulaire en examinant les plus populaires d'entre eux tels que FOAF, Dublin Core, SKOS, Event …

Au-delà de l'utilisation et de la réutilisation des vocabulaires dans les jeux de données publiées, on s'intéressera aussi à la réutilisation par d'autres vocabulaires. Nous y reviendrons largement dans la section dédiée (Linked Open Vocabularies).

2.3. Les vocabulaires sont des constructions sociales

L'histoire des exemples ci-dessus le montrent, les vocabulaires largement partagés sont ceux qui ont vu leur élaboration et leur évolution appuyées sur des échanges publics et ouverts, leurs auteurs, qu'ils soient individuels ou institutionnels, sont connus et se connaissent. A cet égard on peut voir beaucoup de similarités entre la communauté des développeurs de vocabulaires et celle du logiciel libre, dans l'esprit du slogan original du Web "Let's share what we know".

Source : http://www.flickr.com/photos/danbri/4030764915/

Dans le cadre de l'initiative Linked Open Vocabularies, nous essayons de mettre en évidence ce caractère social des vocabulaires en formalisant les liens qu'ils tissent entre eux: extension, référence, généralisation, spécification … chaque lien d'un vocabulaire à un autre enrichit à la fois la source et la cible, et témoigne de la confiance de l'auteur du lien en son homologue gestionnaire du vocabulaire cible. A cet égard on pourra noter comme exemplaire l'accord récemment conclu entre les organisations DCMI (Dublin Core Metadata Initiative) et les auteurs du projet FOAF, accord destiné à assurer l'évolution conjointe des vocabulaires qu'ils maintiennent, et la pérennité de l'espace de noms et de la disponibilité du vocabulaire FOAF si le projet qui le soutient actuellement venait à son terme [DCMI-FOAF agreement].

3. Critères de sélection d'un vocabulaire

De nombreux travaux ont été consacrés à l'évaluation des ontologies en général [OntoEval] mais ces évaluations concernent essentiellement des aspects de qualité intrinsèque (précision, cohérence structurelle et terminologique ...), indépendantes du contexte applicatif d'utilisation. Le contexte spécifique des données liées nécessite une approche particulière, où les paramètres généraux sont présents, mais où d'autres doivent être pris en ligne de compte.
Nous présentons dans cette section un certain nombre de critères qui peuvent être utilisés par le fournisseur de données aussi bien dans la procédure de sélection de vocabulaires existants, que dans la construction de ses propres vocabulaires.
Il nous semble cohérent d'appliquer ces critères dans l'ordre où ils sont présentés ici, et de rejeter l'emploi d'un vocabulaire qui paraîtrait par exemple intéressant en termes de contenu (critères 4) mais qui ne serait pas conforme aux standards (critères 1), ou dont la pérennité serait douteuse en l'absence totale de métadonnées (critères 2 et 3).

3.1. Conformité aux standards et bonnes pratiques

Puisque nous nous plaçons dans le cadre technologique des standards du Web sémantique, il parait naturel de mettre au premier rang des critères la conformité à ces standards et aux bonnes pratiques de publication et de gestion définis par la communauté des utilisateurs.
Même si ces critères peuvent paraître a priori un peu draconiens, il sont respectés d'ores et

déjà par suffisamment de vocabulaires pour ne pas être trop restrictifs. Les éditeurs qui ne prennent pas la peine de les respecter scrupuleusement sont maintenant en grand risque de voir leurs vocabulaires tout simplement ignorés dans l'avenir.

Vocabulaire dédié à la description des ressources en RDF

Même si certains vocabulaires peuvent avoir été adaptés, avec plus ou moins de bonheur, d'un contexte "pré-Web", on retiendra malgré tout de préférence les vocabulaires développés directement sur les standards du Web sémantique (RDFS ou OWL "natif") et mieux encore ceux qui sont développés spécifiquement dans le contexte des données liées sur le Web - dans la mesure où les auteurs déclarent explicitement cet objectif dans leur documentation. Si le vocabulaire lui-même n'est pas documenté clairement à ce sujet, on pourra trouver un indice dans le fait que le vocabulaire est visiblement développé pour soutenir la publication dans le Web des données d'un ensemble de données clairement identifié. Par exemple l'ontologie développée pour Geonames.org [Geonames] a pour objectif prioritaire la publication et la réutilisation des données de la base d'objets géographiques du même nom. De même un certain nombre d'ontologies sont développées par la BBC pour ses besoins propres de publication de données, telles que Programmes Ontology, destinée à décrire les données concernant les programmes de radio ou télévision [BBC-PO].

Vocabulaire exprimé dans un standard du Web sémantique

Les vocabulaires retenus doivent être décrits au moyen d'une des espèces du langage OWL ou de RDFS et publiés dans l'une de leurs syntaxes standard. La pratique générale est l'utilisation de RDF-XML, mais certains vocabulaires sont publiés dans la syntaxe alternative n3.
Certains vocabulaires comportent également l'expression de contraintes ou règles exprimées dans des langages complémentaires tels que RIF, SPARQL ou n3, mais ce n'est pas une pratique très répandue.

Vocabulaire accessible sur le Web via son espace de noms

L'espace de noms du vocabulaire (son URI de base) doit permettre d'accéder, dans l'idéal par un processus de négociation de contenu, au vocabulaire formel d'une part (dans l'un des langages ci-dessus) et à sa documentation. A minima, si l'URI ne renvoie qu'au vocabulaire formel, on doit pouvoir naviguer de la représentation formelle à la documentation HTML et inversement. D'ailleurs la mise en place de négociation de contenu ne dispense pas de cette capacité de navigation d'une représentation à une autre. Ainsi la description formelle du vocabulaire (en RDF) devrait contenir l'adresse de la description informelle (documentation HTML) et inversement la documentation (HTML) doit contenir un lien explicite vers la description formelle. Ce mécanisme de référence mutuelle peut prendre différentes formes. Pour le lien de la page HTML vers le document RDF une bonne pratique est de poser le lien sur une des icônes standard proposées par le W3C [RDF Icons]. Pour la référence de la documentation HTML dans le fichier RDF, l'utilisation de la propriété foaf:isPrimaryTopicOf sur l'élément owl:Ontology est une bonne pratique.

Les divers mécanismes de négociation de contenu et de référence mutuelle des différentes versions du vocabulaire, pourtant bien documentés, [SWBP] sont encore souvent mal mis en œuvre, même par des jeux de données qui par ailleurs présentent une excellente qualité de contenu et de forme de la publication. Ainsi l'excellent site legislation.gov.uk consacré à la

publication des données législatives du Royaume-Uni utilise plusieurs vocabulaires dont un vocabulaire du domaine juridique *metalex*, développé dans le cadre d'un projet européen. Visiblement bien adapté aux données publiées, il est malheureusement inaccessible depuis son espace de noms http://www.metalex.eu/metalex/2008-05-02# ...
Sur la page de documentation http://www.legislation.gov.uk/developer/formats/rdf une erreur de copie donne pour espace de noms de metalex celui du vocabulaire FRBR, ce qui fait que découvrir l'URL où est effectivement publié ledit vocabulaire, nécessite de repasser par la racine du site http://www.metalex.eu/ pour découvrir effectivement un lien vers l'ontologie effectivement publiée à http://svn.metalex.eu/svn/MetaLexWS/branches/latest/metalex-cen.owl.
Pour en finir avec cet exemple, et puisqu'on a fini par découvrir ce vocabulaire, on verra que même s'il avait passé avec succès le critère 1, il passerait mal les étapes suivantes, puisqu'il manque de visibilité même à partir des données qui l'utilisent (critère 2), ne contient aucune des métadonnées souhaitables (critère 3) et enfin ne réutilise aucun autre vocabulaire. Il redéfinit en effet des classes telles que "Event", "Document" "Agent" sans les relier à des vocabulaires partagés et connus qui définissent des classes tout à fait équivalentes (FOAF, Dublin Core, Event …)

Vocabulaire documenté

Comme on l'a dit ci-dessus, il est indispensable que le vocabulaire formel publié dans un des formats indiqués ci-dessus soit accompagné d'une documentation HTML. Il existe maintenant un certain nombre d'applications qui permettent de générer la documentation HTML à partir du code RDFS ou OWL, ce qui simplifie les procédures de maintenance. On peut citer des outils [Neologism] [OntoSpec] [LODE] et [Parrot].
Il est toujours bon de trouver en plus de cette reproduction HTML automatique du contenu formel, des documents de présentation générale indiquant le contexte du vocabulaire, comme le cadre de son développement, les données l'utilisant, les perspectives de maintenance ou de développement ultérieur, les articles ou contributions à conférences citant le vocabulaire etc.
Une représentation graphique sous forme de diagramme UML ou équivalent est également une aide appréciée à la compréhension globale du vocabulaire ou de ses aspects les plus importants ou les plus délicats.
Comme on l'a évoqué plus haut, des exemples de données ou d'applications utilisant le vocabulaire sont aussi les bienvenues. Un générateur de données utilisant le vocabulaire à partir d'un simple formulaire d'entrée est également très pédagogique. On peut citer l'application [FOAF-o-Matic] qui permet de créer très simplement un profil FOAF.

3.2. Stabilité, notoriété et usage

Cette deuxième série de critères rassemble des paramètres non explicites et pas toujours mesurables mais néanmoins importants, et qui ne sont pas nécessairement explicites dans le vocabulaire lui-même ni dans ses métadonnées, et nécessitent donc un jugement « humain » de l'utilisateur potentiel et une certaine connaissance des autorités de publication.
La stabilité d'un vocabulaire se mesure de deux façons : la pérennité de l'espace de noms et de l'autorité gérant le DNS associé, et la stabilité des termes du vocabulaire lui-même. S'il est difficile de prévoir de telles évolutions, on peut malgré tout évaluer en général la notoriété et le sérieux des autorités de publication. Or de nombreux vocabulaires ont été ou sont

encore publiés dans le cadre de projets de recherche pas nécessairement pérennes, ou sont le fruit d'initiatives individuelles ou de groupes restreints qui ont été pionniers dans le domaine. Au fur et à mesure que le Web de données mûrit, le relais de ces initiatives doit être assuré par des institutions compétentes et fiables, dont le but devrait être maintenant la pérennité des ressources et des URI associées, une gestion sérieuse des versions, une documentation à jour etc. L'exemple cité plus haut de l'accord entre DCMI et FOAF est exemplaire à cet égard, cette initiative se prolongeant lors de la conférence Dublin Core 2011, qui consacre une session spéciale à la question de l'harmonisation des vocabulaires [DC-2011]. Suite à cette conférence, un groupe de travail dédié à cette tâche a été mis en place.
D'ores et déjà de nombreux acteurs institutionnels prennent des initiatives en ce sens. Bien sûr les premiers d'entre eux ont été les organismes de standardisation comme le W3C ou le DCMI, mais on peut maintenant leur ajouter des grandes bibliothèques comme la Library of Congress, la BNF et d'autres grandes bibliothèques européennes, des organismes internationaux comme la FAO, les institutions européennes à travers l'Office des Publications, etc. Dans le cadre de Datalift, ce rôle sera joué d'abord par l'IGN et l'INSEE, rejoints espérons-le par d'autres acteurs institutionnels comme la DILA.

La notoriété d'un vocabulaire peut se mesurer au nombre de jeux de données qui l'utilisent, tels qu'on peut les découvrir dans le catalogue de CKAN par exemple [CKAN-FOAF] que FOAF est le vocabulaire utilisé par le plus grand nombre de jeux de données dans le catalogue. La réutilisation par d'autres vocabulaires est aussi un moyen viral d'augmenter la notoriété. C'est particulièrement vrai pour les vocabulaires de métadonnées dont nous parlerons dans la section suivante. Le nombre de jeux de données est intéressant à plusieurs titres :

- C'est une indication concernant la facilité de ré-utilisation du vocabulaire ;
- C'est aussi une garantie sur la durabilité et la maintenance du vocabulaire : un vocabulaire très utilisé indique une communauté d'utilisateurs nombreuse et active ;
- Un vocabulaire communément utilisé permet à un plus grand nombre d'utilisateurs de comprendre les données qu'un fournisseur publie.

Le vocabulaire Vocabulary of Interlinked Datasets [VoID] permet de décrire formellement la relation d'utilisation d'un vocabulaire par un jeu de données, par l'intermédiaire de la propriété void:vocabulary. Le catalogue CKAN fournit de plus en plus de descriptions de jeux de données basées sur VoID.

3.3. Métadonnées

Nous examinerons maintenant plus en détail les critères concernant les métadonnées des vocabulaires. Le caractère récursif de RDF permet de l'utiliser pour décrire les vocabulaires eux-même, et de fait on dispose aujourd'hui d'une palette assez vaste de vocabulaires de métadonnées permettant une description fine des vocabulaires aussi bien au niveau global qu'au niveau de chacun des éléments [LOV-Metadata]. Il est à noter que les vocabulaires de métadonnées sont multiples et redéfinissent souvent des propriétés similaires, le Dublin Core restant le plus couramment utilisé. Sur la première centaine de vocabulaires recensés par le projet Linked Open Vocabularies, près de 4 vocabulaires sur 5 utilisent des métadonnées Dublin Core Elements (préfixe dc: ci-dessous) ou Dublin Core Terms (préfixe dcterms:). Sont également fréquemment utilisés (par 1 vocabulaire sur 4 environ) les

vocabulaires de métadonnées spécifiques [VANN] et [Vocab Status]. FOAF et SKOS sont également fortement utilisés comme vocabulaires de métadonnées, nous y reviendrons dans la section consacrée aux « Common Vocabularies ».
Dans une bonne pratique de publication, les métadonnées du vocabulaire lui-même sont contenues dans le RDF comme des propriétés d'une ressource qui représente le vocabulaire lui-même. En général cette ressource utilise le type owl:Ontology. Une bonne pratique est que ces mêmes métadonnées soient dupliquées dans la documentation HTML du vocabulaire. On peut citer comme exemple de bonne pratique à cet égard Music Ontology [MO].

Origines : auteur, éditeur, source

Les origines d'un vocabulaire renseignent le fournisseur de données à différents niveaux :

- Sur la confiance à accorder au vocabulaire : L'éditeur et/ou les auteurs (organisme public, organisme privé, groupe d'experts, personne seule, etc.) font-ils autorité dans leur domaine ? Sont-ils encore en activité ? Ont-ils communiqué ou publié des articles sur le vocabulaire en question ? Les réponses aideront le fournisseur de données à mesurer la confiance avec laquelle il peut aborder le vocabulaire.
- Outre la description ou la présentation générique du vocabulaire, le domaine de connaissances couvert par le vocabulaire est renseigné aussi par l'éditeur : l'activité, la taille, l'origine géographique de ce dernier sont des indications sur la proximité du vocabulaire avec les données à publier.
- L'indication d'une source (dc:source ou dcterms:source) peut renseigner sur le fait que ce vocabulaire n'est pas créé nativement dans le format sémantique courant, mais a été adapté d'un vocabulaire pré-existant, éventuellement dans un format moins formalisé.

Différentes propriétés peuvent être utilisées pour déclarer les origines d'un vocabulaire. Par exemple, on pourra utiliser : dcterms:contributor, dcterms:creator, dcterms:publisher, sioc:has_creator, sioc:has_modifier, sioc:owner_of or foaf:maker, dc:provenance, dc:source, dcterms:source. Comme beaucoup de ces propriétés sont peu ou prou équivalentes et pas toujours ou pas encore formellement alignées, certains éditeurs de vocabulaires n'hésitent pas à dupliquer la même information dans plusieurs dialectes, utilisant par exemple à la fois foaf:maker et dcterms:creator pour désigner l'auteur du vocabulaire. Cet état de choses a conduit les éditeurs de FOAF et Dublin Core notamment à prendre des initiatives communes en vue d'harmoniser les vocabulaires les plus utilisés et éviter les redondances. Nous y reviendrons également dans la section « Common Vocabularies ».

Le vocabulaire "Programmes ontology" (http://purl.org/ontology/po/) est un exemple de bonne pratique concernant l'indication des origines.

Title	Programmes ontology
Contributor	Patrick Sinclair
	Nicholas J Humfrey
Creator	Yves Raimond
Description	This ontology aims at providing a simple vocabulary for describing programmes. It covers brands, series (seasons), episodes, broadcast events,

	broadcast services, etc.
Publisher	BBC

Droits et licence

Les contraintes juridiques associées à un vocabulaire sont importantes pour assurer sa ré-utilisation dans des conditions juridiques correctes. Des licences comme Creative Commons ou GNU indiquent les modalités de réutilisation auxquels le fournisseur de données devra se conformer.

Parmi les propriétés de métadonnées qui indiquent les droits liés à un vocabulaire, nous pouvons citer: cc:licence, dcterms:licence, dcterms:rights and dc:rights

Le vocabulaire “Vocabulary Of A Friend” (http://labs.mondeca.com/vocab/voaf#) est un exemple de bonne pratique concernant l’indication des droits d’utilisation.

cc:license	http://creativecommons.org/licenses/by/3.0/

Version, statut et historique

Les informations de version, statut et historique d’un vocabulaire peuvent intervenir utilement dans le choix des vocabulaires à utiliser. Si un vocabulaire est indiqué avec un statut “in progress”, son utilisation est peut être risquée. Si la première date de publication est très récente, on peut s’interroger sur la pérennité du vocabulaire, si au contrare la dernière date de modification remonte à plusieurs années, on doit se demander si c’est un indice de la stabilité du vocabulaire, ou du fait que sa gestion est à l’abandon. L’accès aux versions antérieures, numérotées et datées, du vocabulaire et notamment à la version directement antérieure, est évidemment souhaitable, de même que la liste des modifications depuis la version précédente : éléments introduits, modifiés, dépréciés ou supprimés..

Une bonne pratique, encore trop peu répandue, est l’utilisation de la métadonnée vs:term_status qui est une métadonnée au niveau de chaque terme du vocabulaire, dont les valeurs typiques (bien que non contrôlées) sont “stable” et “testing”. L’indication d’une valeur “stable” est un gage de pérennité du terme en question. On trouve encore rarement aussi des indications plus fines de version au niveau des termes, comme la date d’introduction ou de modification, la liste des termes obsolètes et de leur remplaçant éventuel.

On constate que cet aspect de gestion des versions et d’historique est encore très mal gérée de façon globale au niveau du Web des données. On voit encore trop souvent des jeux de données publiés utilisant des éléments de vocabulaires dépréciés. L’exemple le plus notoire étant l’utilisation de la propriété skos:subject, entre autres par DBpedia pendant plusieurs années (jusqu’en 2010 au moins) alors que ce terme n’a été présent que dans des versions préliminaires de SKOS, déclaré “instable” puis abandonnée en 2008 et absente de la recommandation en date de 2009. [SKOS Subject]

Parmi les propriétés de métadonnées qui renseignent sur les versions, statuts et historiques, nous pouvons citer: owl:versionInfo, vs:term_status, dcterms:created et dcterms:modified, sioc:earlier_version or sioc:previous_version.

Le vocabulaire “Vocabulary Of A Friend” (http://labs.mondeca.com/vocab/voaf#) est un exemple de bonne pratique concernant l’indication du statut

vs:term_status	“In progress”

Domaine d'application

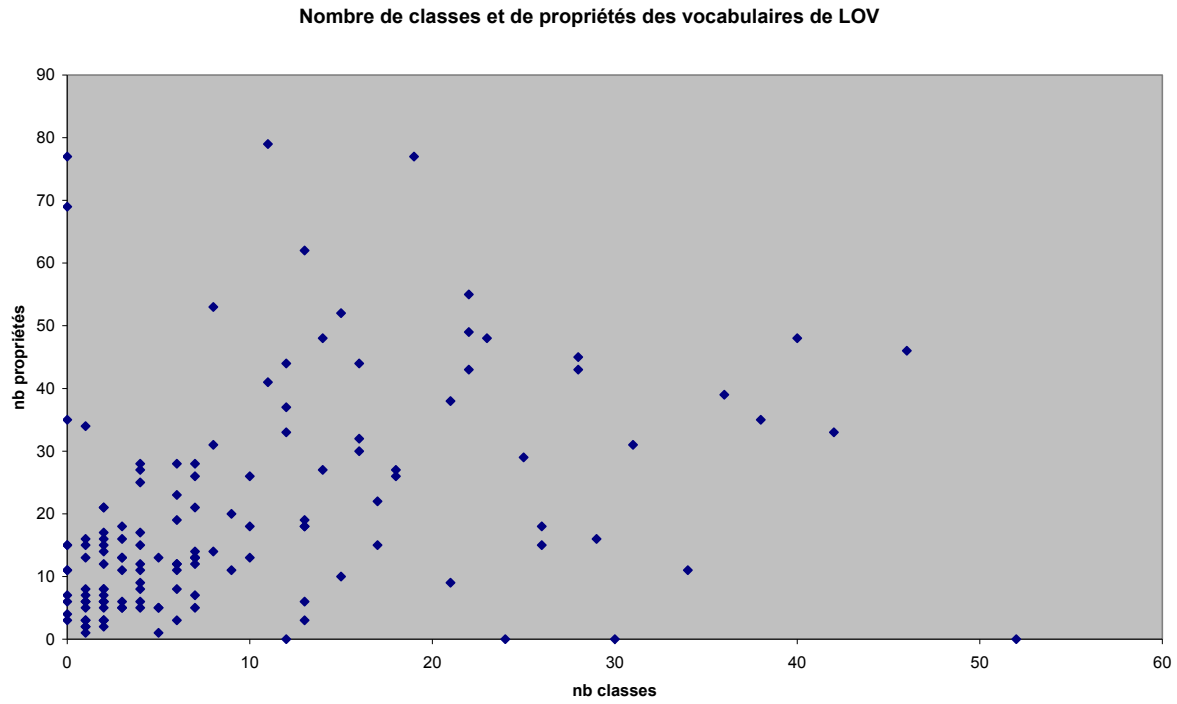
Le domaine d'application d'un vocabulaire est une métadonnée qu'on souhaiterait voir apparaître mais qui est pour le moment quasi-absente, en tout cas des représentations formelles. On peut attribuer cette absence d'une part à la difficulté de définir clairement cette notion de "domaine d'application", et sans doute également de vocabulaire de référence (taxonomie des domaines d'application). On pourra bien sûr se fier à la documentation informelle, mais il est dommage qu'on ne puisse s'appuyer sur un critère formel pour la recherche de vocabulaires dans un catalogue ou via une requête SPARQL sur un entrepôt de vocabulaires.
Dans le cadre de Linked Open Vocabularies, nous avons élaboré une taxonomie à trois niveaux sous forme de ce que nous avons appelé "Vocabulary Spaces", mais cette approche est encore loin de fournir une réponse standard à cette question.

3.4. Qualité du contenu

Expressivité

Bien que les vocabulaires utilisés par les données liées soient majoritairement des ontologies assez simples, l'information de leur expressivité peut être utile. Typiquement l'expressivité peut être limitée à RDFS, OWL-Lite, ou OWL-DL. On tiendra compte de ces informations notamment si les données publiées sont destinées à être utilisées par un outil de raisonnement.

Taille et modularité



Le graphique ci-dessus indique la répartition en nombre de classes et nombre de propriétés des vocabulaires recensés dans LOV, limité aux vocabulaires contenant moins de 100 éléments au total (classes et propriétés), qui représentent plus de 85% des 162 vocabulaires recensés (à la date du 28 décembre 2011). La taille généralement réduite de la majorité de ces vocabulaires confirme que le caractère réutilisable diminue avec la taille et la complexité.

On retiendra donc comme une bonne pratique le fait de découper un vocabulaire en plusieurs modules relativement indépendants, chacun d'eux visant un domaine d'application le plus spécifique possible. Le slogan « petit mais viral » résume assez bien cette approche. Un ensemble de vocabulaires interconnectés publiés par le même éditeur devra être documenté en spécifiant comment ces vocabulaires se relient les uns aux autres. Un bon exemple de telle interconnexion est fourni par le vocabulaire Play Back Ontology [PBO].

Réutilisation et intégration avec d'autres vocabulaires

Le nombre de vocabulaires réutilisés est un indicateur facilement mesurable et intéressant à plusieurs titres. Il peut indiquer une application du principe de modularité ci-dessus, si les vocabulaires réutilisés sont dans le même espace de noms que le vocabulaire courant. Mais plus généralement il témoigne d'une bonne pratique évoquée dans les principes généraux, le souci de l'auteur à intégrer l'existant et à ne pas réinventer. Comme on l'a dit, il s'agit autant d'un principe de simplicité et d'économie que d'une bonne pratique sociale.

Le vocabulaire Vocabulary of a Friend [VOAF], développé dans le cadre de Datalift et que nous présenterons plus en détail par la suite, permet de décrire ces relations entre vocabulaires, par l'intermédaire de propriétés spécifiques telles que: voaf:extends, voaf:specializes, etc. VOAF lui-même est une spécialisation de VoiD.

Par exemple, le vocabulaire "Event Ontology" étend les vocabulaires FOAF, Geo et Time, comme indiqué dans la description de ce vocabulaire dans la page dédiée de Linked Open Vocabularies : http://labs.mondeca.com/dataset/lov/details/vocabulary_event.html

http://purl.org/NET/c4dm/event.owl# voaf:extends http://xmlns.com/foaf/0.1/
http://purl.org/NET/c4dm/event.owl# voaf:extends
http://www.w3.org/2003/01/geo/wgs84_pos#
http://purl.org/NET/c4dm/event.owl# voaf:extends http://www.w3.org/2006/time

Nommage et définition des termes

A minima, tout terme nommé (possédant une URI) d'un vocabulaire doit déclarer au moins un libellé en langage naturel, utilisant la propriété rdfs:label, avec si possible mention de la langue. Une définition ou description (utilisant les annotations dcterms:description ou skos:definition) est souhaitable, surtout dans les cas où la signification du terme n'est pas triviale dans le contexte.
Il est à noter que la très grande majorité des ontologies utilisées dans le Web des données utilisent exclusivement l'anglais pour les libellés et les définitions, même si l'usage de cette langue n'est pas explicitée (par un attribut xml:lang par exemple). Il est souhaitable que cette pratique monolingue évolue. Dans le cadre de Datalift, nous développerons bien sûr des ontologies utilisant le français pour les libellés et les descriptions, de façon à promouvoir la pratique du multilinguisme dans les vocabulaires.

3.5. Métriques d'utilisation des termes

Les critères mentionnés jusqu'ici se situent au niveau du vocabulaire dans son ensemble. On peut aussi utilement mettre à disposition des utilisateurs des métriques plus fines se situant au niveau des éléments (classes et propriétés). Il sera par exemple intéressant de fournir des statistiques sur l'utilisation de tel élément par d'autres vocabulaires (taux

d'utilisation dans LOV par exemple) et dans les jeux de données du Linked Open Data, ou même de classer les propositions d'utilisation de tel ou tel élément en fonction de telles statistiques.
Il revient à la tâche D2.2 de développer les outils permettant l'intégration de telles métriques dans le module de sélection de la plate-forme DataLift.

4. Inventaire des ressources existantes

Nous rassemblons sous cette section les ressources utilisables dans le cadre de la recherche et de la gestion de vocabulaires, dont certaines ont déjà été introduites dans les sections précédentes

4.1. Catalogues, index et moteurs de recherche

CKAN

Comprehensive Knowledge Archive Network [CKAN]. CKAN rassemble les métadonnées et les descriptions de jeux de données ouvertes publiées sur le Web, et en particulier rassemble sous des "tags" les jeux de données utilisant un vocabulaire.

Schemapedia

[Schemapedia] est une ressource qui référence des schémas de description de données, avec des descriptions et des exemples de données. Ce projet s'intéresse plus à la T-box des ontologies (classes et propriétés) qu'à la A-box (instances).

Watson

[WATSON] est un moteur de recherche de classes ou de propriétés dans des ontologies existantes. [D'Aquin M et al., 2007] . Les options de recherche permettent de chercher un terme désignant une classe ou une propriété, ou utilisé dans la description de ces éléments.

Swoogle

[Swoogle] est un moteur de recherche d'URIs publiées dans les formats du Web sémantique. Une des options permet de rechercher des URI d'éléments d'ontologies (classes ou propriétés) à partir d'un libellé. Mais il s'agit de données brutes, sans information de contexte. Ce service est difficilement utilisable pour la recherche d'information sur les vocabulaires.

Sindice

[Sindice] est également un index général et moteur de recherche des ressources du Web sémantique, et sans doute le plus riche à ce jour. Il permet comme le précédent une recherche multi-critères fine sur des éléments d'ontologies, mais ne présente pas plus que Swoogle d'interface dédiée spécifiquement à la recherche de vocabulaires. Par contre il permet de trouver des instances d'une classe connue, et donc donner des indications quantitatives et qualitatives sur l'utilisation d'un vocabulaire.

4.2. Outils d'édition et publication

Neon Toolkit

Neon Toolkit est un outil d'aide à la construction d'ontologie par la réutilisation d'ontologies existantes [Haase P et al., 2008].

Neologism

[Neologism] est une plate-forme intégrée pour le développement et la publication d'ontologies, développée par le DERI, et utilisant Drupal. Cette plate-forme est utilisée pour la gestion et la publication des vocabulaires développés dans le cadre du projet Datalift.

Protégé

[Protégé] est sans doute l'application la plus connue et la plus utilisée pour l'édition d'ontologies, même si ses fonctionnalités sont un peu surdimensionnées pour la gestion de vocabulaires simples.

TopBraid Composer

[TopBraid] Composer est une application proposée par l'entreprise TopQuadrant, qui se veut l'équivalent "professionnel" de Protégé. En ce qui concerne les vocabulaires simples dont nous traitons ici, sa valeur ajoutée par rapport à Protégé parait minime.

Rappelons aussi l'existence des outils permettant la génération de documentation html à partir de la source RDF du vocabulaire. Ces outils peuvent être utilisés aussi bien par l'éditeur du vocabulaire que par l'utilisateur potentiel qui ne dispose que de la source RDF. [OntoSpec] [LODE] [Parrot]

4.3. Outils méthodologiques

Ontometric

Ontometric est une méthode qui permet à un utilisateur de choisir les ontologies les plus pertinentes face à ses besoins [Lozano-Tello A et al., 2004].

OntoSim

OntoSim est une librairie de distances entre ontologies qui pourrait être utilisée pour rechercher tous les vocabulaires du LOV proches d'un vocabulaire (ou jeu de données).
http://ontosim.gforge.inria.fr/

5. Ressources développées dans le cadre de Datalift

La définition des critères présentés dans ce livrable s'est faite en parallèle avec le début de la création du catalogue d'ontologies qui sera décrit en détail dans le livrable D2.2. Nous pouvons néanmoins présenter les ressources que nous avons d'ores et déjà développées et qui seront détaillées dans le livrable D2.2.

5.1. Ontologies

L'ensemble des ontologies développées dans le cadre du projet Datalift est publiée sur la plate-forme dédiée par le LIRMM [LIRMM-Onto]. Cette plate-forme utilise la technologie Neologism citée plus haut.

Ontologies géographiques de l'INSEE

L'INSEE avait développé dès 2006 en collaboration avec Mondeca et XML.fr une première publication des éléments du Code Officiel Géographique aux formats du Web sémantique. Cette première publication comprenait une ontologie au format OWL des entités géographiques de la France jusqu'au niveau de la commune, et des fichiers RDF packagés par départements [COG]. Les données ont été mises à jour par l'INSEE en 2011.
Dans le cadre de Datalift, l'ontologie de 2006 a été révisée et enrichie de la description des structures infra-communales et d'éléments d'historisation, ainsi que d'une liaison plus forte avec les ontologies "populaires" du Web des données (FOAF, Dublin Core, Geonames) [INSEE-2011].

Ontologies géographiques de l'IGN

Dans le domaine des données géographiques, la sémantique des données - compris comme le lien entre les données et la réalité - est porté par un ensemble d'informations particulier : "les spécifications d'acquisition, de saisie et de contenu." Celles-ci décrivent de façon détaillée quelles catégories d'objets du monde réel sont observés (les routes, les forêts) et comment ces objets sont représentés dans la base (classe route avec un attribut "géométrie" qui correspond à une suite de points indiquant l'axe de la route). C'est ainsi que pour comprendre des données géographiques et les intégrer avec d'autres données, il est important de comprendre deux niveaux d'information.
1. D'une part le domaine représenté dans les données. A cet égard il y a souvent des subtilités venant du fait qu'il est difficile de définir des catégories universelles. A partir de quand cesse-t-on de considérer une étendue d'eau comme un lac? A partir de quand, du moins, cesse-t-on de s'intéresser à des étendues d'eau et à les mettre sur la carte? D'autres subtilités viennent de la notion de niveau de détail. Il est important de préciser, si on s'intéresse à un phénomène comme les villes à quel niveau de détail on considère les villes : comme des entités agrégées, qui seront représentées par un point et par des statistiques globales.
2. D'autre part la représentation, c'est-à-dire la structure de données et les modalités de documentation de cette structure (comment est-ce que je dessine la géométrie d'une maison sur la carte : un polygone correspondant aux murs ou correspondant à l'aplomb du toit?).
Une thèse conduite au COGIT s'est concentrée sur l'analyse des catégories de connaissance rentrant dans les spécifications des produits IGN et a proposé un premier modèle formel de représentation de ces spécifications [Gesbert 2005]. A la suite, des travaux ont porté dans un premier temps sur la semi-automatisation de l'acquisition des spécifications formelles depuis les documents textuels grâce à des outils de TALN et une thèse se consacre à proposer un formalisme qui permette les inférences nécessaires à l'intégration de données [Abadie 2010].

Le maintien de ces deux niveaux est pertinent dans le contexte de Datalift quand bien même le but n'est pas d'intégrer différentes bases de données géographiques.
1. pour faciliter le traitement automatique de certaines hétérogénéités dans les modèles. Pouvoir détecter automatiquement la correspondance entre la classe "chef lieu" du modèle IGN et la relation "chef lieu" du modèle INSEE (les méthodes à base de clé permettent aussi de le faire mais sur certains cas sont probablement plus coûteuses).
2. pour exprimer des informations de précision qui s'expriment comme l'écart maximum de la position de l'objet dans la représentation à la position de l'objet dans le monde réel.

Dans le cadre du projet ANR [GeoOnto] une ontologie a été proposée qui reprend les principes de modélisation, c'est-à-dire qu'elle comprend une ontologie du monde réel, une ontologie des structures de données et un alignement entre les deux rendant compte de la représentation du monde réel dans la base. La partie "ontologie du monde réel" de l'ontologie GeoOnto est reprise dans Datalift (elle est en cours de finalisation).

Une nouvelle version de la [BD TOPO] de l'IGN étant disponible, une nouvelle ontologie rendant compte de la structure de cette nouvelle version a été produite en s'attachant aux classes d'objets considérées d'intérêt dans le projet : routes, bâtiments, limites administratives, points d'intérêt.

Ontologie de l'Assemblée Nationale (OAN)

Cette ontologie représente la description du fonctionnement de l'Assemblée Nationale française. Elle est en cours de développement et publiée sur le site du LIRMM [OAN]. Les données de test pour cette ontologie seront celles du fournisseur "Regards Citoyens", actuellement publiées sous http://www.nosdeputes.fr/

Vocabulary of a Friend (VOAF)

Le catalogue des ontologies pour la création de jeux de données nécessite lui-même l'utilisation d'un vocabulaire pour décrire ces ontologies et leurs relations. De même que le vocabulaire Vocabulary of Interlinked Datasets (VoID) décrit les relations entre jeux de données [Alexander et al., 2009], nous avons étudié les vocabulaires qui permettent de décrire des vocabulaires et les relations qu'ils entretiennent. Afin de respecter les bonnes pratiques du web de données et de ce projet, nous réutilisons des classes et propriétés existantes parmi les vocabulaires:

- Vocabulary Of Interlinked Datasets (VoID): un vocabulaire pour décrire les jeux de données. Il apporte des informations de statistique ou encore la localisation du fichier de contenu d'un jeu de données;
- Dublin Core Terms (dcterms): un vocabulaire pour la description de métadonnées sur des documents électroniques [Weibel S. 1997].
- Vocabulary for ANNotating vocabularies [VANN] est un vocabulaire pour la description de vocabulaires. Nous utilisons ses propriétés vann:preferredNamespacePrefix et vann:preferredNamespaceUri pour représenter le préfixe à utiliser et le namespace d'un vocabulaire.
- BIBliographic Ontology (bibo) est un vocabulaire pour la description de données bibliographiques. Nous utilisons sa propriété bibo:shortTitle afin de décrire le titre abrégé d'un vocabulaire [D'Arcus B et al., 2010].

En ce qui concerne les relations entre vocabulaires, nous n'avons pas les éléments à disposition dans les vocabulaires existants. Nous avons donc créé un nouveau vocabulaire VOAF, qui repose sur ceux existants, à cet effet.

5.2. Linked Open Vocabularies (LOV)

Le Linked Open Vocabularies est un jeu de données exprimé en langage RDF, qui inventorie les vocabulaires de description des jeux de données du web sémantique mais également les relations qui lient les vocabulaires. Le travail de recensement au sein du LOV n'est pas

exhaustif, c'est un travail en cours. Le web de données est par définition non contrôlé et par nature en constante évolution. De même que son environnement et ce qu'il décrit, ce travail a comme caractéristique intrinsèque d'être inachevé et incomplet.

Méthode et outils

Le travail initial de construction du LOV est manuel. Nous avons de par notre expérience et l'analyse des jeux de données, identifié plus de 100 vocabulaires satisfaisant nos critères. Le travail de mise à jour et d'évolution dépend maintenant en partie de nos recherches (semi-automatiques), mais également des remarques des utilisateurs du web de données.
Pour assister ce travail de mise à jour et compte tenu du fait que les vocabulaires peuvent évoluer de manière indépendante sans que nous en soyons avertis, nous avons mis en place un programme automatique de veille. Ce programme a pour objectif de confronter les données contenues dans le LOV aux données des vocabulaires. Si nous prenons l'exemple du nombre de propriétés, le programme récupère dans un premier temps l'identifiant du vocabulaire (URI). Cette URI fait normalement référence au document descriptif du vocabulaire dans un format du web sémantique. Si tel est le cas, on exécute une requête qui retourne le nombre effectif de propriétés dans ce vocabulaire. Enfin si la valeur est différente de celle contenue dans les données du LOV, une annotation est ajoutée à ce vocabulaire dans le LOV qui est ensuite validée manuellement.

Résultats préliminaires

Les vocabulaires rassemblés dans LOV confirment une tendance déjà visible dans les jeux de données rassemblés dans CKAN. Un noyau de vocabulaires fortement interconnectés et fortement réutilisés s'organise autour de FOAF, Dublin Core et SKOS, et dans une moindre mesure les ontologies de l'espace et du temps Geo et Event. Ce noyau forme ce qu'on peut appeler des "Vocabulary Commons", c'est-à-dire un ensemble de classes et de propriétés d'usage courant dans tous les domaines de la représentation des connaissances. Ces éléments incontournables devraient être connus et réutilisés par tous les gestionnaires de vocabulaires et les producteurs de données, et ne plus être redéfinis et réinventés!

Travaux en cours

Une analyse plus fine de l'interconnexion entre vocabulaires autour des "Commons" est en cours pour préciser quels éléments (classes et propriétés) de ces vocabulaires sont les plus réutilisés. Il sera également intéressant de voir comment ces "Commons" vont s'interconnecter avec l'initiative [schema.org] lancée au printemps 2011 par les grands moteurs de recherche (Microsoft, Google et Yahoo!) et dont les classes et propriétés recouvrent largement le périmètre des "Commons". Bien que RDF ne constitue pas la cible principale de schema.org, qui recommande plutôt l'usage des formats "microdata" dans le HTML, RDFa est l'un des formats possibles du marquage, et une version OWL du schema est disponible [schema.org-OWL].

6. Conclusion

Dans l'état actuel de l'art et dans un avenir prévisible, le choix d'un vocabulaire reste une tâche qu'il semble illusoire de prétendre vouloir automatiser complètement. Mais nous avons montré un certain nombre de pistes allant vers une amélioration de la visibilité des vocabulaires existants, de leur qualification par des métadonnées convenables, et de leur

réutilisation potentielle.
Il s'agit là d'une conclusion provisoire dans la première année du projet. Ces pistes seront poursuivies, mises en pratique et consolidées tout au long du travail sur les données des acteurs de Datalift, dans le développement des ontologies spécifiques nécessaires. Le travail de catalogage et de liage des vocabulaires entrepris notamment avec le projet Linked Open Vocabularies sera poursuivi dans le cadre de la tâche T2.2.

7. Références

[Abadie 2010] Intégration des bases de données à partir de la formalisation de leurs spécifications. Thèse en cours.
[Alexander *et al.*, 2009] Alexander K., Cyganiak R., Hausenblas M. & Zhao, J. (2009). Describing linked datasets. Proceedings of the 2nd Workshop on Linked Data on the Web.
[BBC-PO] BBC Programmes Ontology http://purl.org/ontology/po/
[BD TOPO] BD TOPO® (RGE) database
http://professionnels.ign.fr/ficheProduitCMS.do?idDoc=5667214
[Berners-Lee, 2006] Tim Berners-Lee. Linked Data - Design issues
http://www.w3.org/DesignIssues/LinkedData.html
[Bizer *et al.,* 2009] Bizer C., Heath, T. & Berners-Lee T. (2009). Linked data-the story so far. Int. J. Semantic Web Inf. Syst., 5, p. 1-22. http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf
[Charlet, 2002] Charlet J. (2002). L'Ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales. Université Paris 6.
[CKAN] Comprehensive Knowledge Archive Network. http://www.ckan.net
[CKAN-LOD] Groupe dédié aux données liées http://www.ckan.net/group/lod
[CKAN-FOAF] Jeux de données utilisant FOAF : http://ckan.net/tag/format-foaf
[COG] Données géographiques de l'INSEE au format RDF http://rdf.insee.fr/geo/
[D'Aquin M *et al.*, 2007] D'Aquin, M.; Baldassarre, C.; Gridinoc, L.; Sabou, M.; Angeletou, S. & Motta, E. Watson: Supporting next generation semantic web applications**.**
[D'Arcus B *et al.*, 2010] D'Arcus B. & Giasson F. (2010). Bibliographic ontology specification. Retrieved October, 8.
[DC-2011] Dublin Core 2011 - Special Session on Vocabulary management and Alignment http://dcevents.dublincore.org/index.php/IntConf/index/pages/view/specialSessions-2011
[DCMI-FOAF agreement], 2011 http://dublincore.org/documents/dcmi-foaf/
[DL] Logique de Description http://fr.wikipedia.org/wiki/Logique_de_description
[FOAF-a-Matic] génération d'un profil FOAF à partir d'un formulaire
http://www.ldodds.com/foaf/foaf-a-matic
[Geonames] Ontologie de Geonames.org
http://www.geonames.org/ontology/ontology_v2.2.1.rdf
[GeoOnto] Projet GeoOnto - ANR-07-MDCO-005 (2008 - 2010) Constitution, alignement, comparaison et exploitation d'ontologies géographiques hétérogènes
[Gesbert 2005] Gesbert N., 2005, Etude de la formalisation des spécifications de bases de données géographiques en vue de leur intégration, *Thèse de doctorat en informatique, Université de Marne-La-Vallée*, soutenue le 2 décembre 2005.
[Gruber, 1993] Gruber, T. (1993). A translation approach to portable ontology specifications. Knowledge acquisition, Citeseer, 5, 199-199.

[Haase P *et al.*, 2008] Haase, P.; Lewen, H.; Studer, R.; Tran, D.; Erdmann, M.; d'Aquin, M. & Motta, E. (2008). The neon ontology engineering toolkit.
[INSEE-2011] Ontologie géographique de l'INSEE http://data.lirmm.fr/ontologies/geo-fr
[LIRMM-Onto] LIRMM ontologies publishing platform http://data.lirmm.fr/ontologies/
[LOD] Linked Open Data http://linkeddata.org
[LODE] Live OWL Documentation Environment http://lode.sourceforge.net/
[LOV] Linked Open Vocabularies http://labs.mondeca.com/dataset/lov
[LOV-Metadata] Linked Open Vocabularies - Metadata Vocabularies
http://labs.mondeca.com/dataset/lov/details/vocabularySpace_Metadata.html
[Lozano-Tello A *et al.*, 2004] Lozano-Tello, A. & Gómez-Pérez, A. Ontometric: A method to choose the appropriate ontology. Journal of Database Management (JDM), IGI Global, 15, 1-18
[MO] Music Ontology Specification Giasson, F, Raymond, Y. (2007)
http://motools.sourceforge.net/doc/musicontology.html
[Neologism] Une plate-forme pour le développement et la publication d'ontologies utilisant Drupal http://neologism.deri.ie/
[OAN] Ontologie de l'Assemblée Nationale http://data.lirmm.fr/ontologies/oan
[OntoEval] Ontology Evaluation, D. Vrandecic Ph. D Thesis, 2010
http://www.aifb.kit.edu/images/b/b5/OntologyEvaluation.pdf
[OntoSpec] OntoSpec, un module Prolog pour la génération de documentation XHTML des vocabulaires RDFS http://moustaki.org/ontospec/
[OWL] Web Ontology Language http://www.w3.org/2001/sw/wiki/OWL
[Parrot] a RIF and OWL documentation service http://ontorule-project.eu/parrot/parrot
[PBO] Play Back Ontology http://purl.org/ontology/pbo/core#
[Protégé] Ontology Editor and Knowledge Acquisition System http://protege.stanford.edu/
[RDF] Resource Description Framework http://www.w3.org/RDF/
[RDFS] RDF Schema http://www.w3.org/TR/rdf-schema/
[RDF Icons] http://www.w3.org/RDF/icons/
[schema.org] A shared marked up schema by Microsoft, Google and Yahoo! http://schema.org/
[schema.org-OWL] L'ontologie schema.org en OWL
http://schema.org/docs/schemaorg.owl
[Schemapedia] Schemapedia, the RDF schema compendium http://schemapedia.com/
[Sindice] The Semantic Web Index http://sindice.com/
[SKOS Subject] SKOS Issues : ISSUE-77 Subject Indexing
http://www.w3.org/2006/07/SWD/track/issues/77
[SW] Semantic Web Activity http://www.w3.org/2001/sw/
[SWBP] Best Practice Recipes for Publishing RDF Vocabularies, W3C Note, 2008
http://www.w3.org/TR/swbp-vocab-pub/
[Swoogle] Swoogle Semantic Web Search Engine http://swoogle.umbc.edu/
[TopBraid] TopBraid Composer
http://www.topquadrant.com/products/TB_Composer.html
[Uschold *et al.,* 1996] Uschold, M. & Gruninger, M. (1996) Ontologies: Principles, methods and application. The Knowledge Engineering Review, Cambridge Univ Press, 11, 93-136.
[VANN] A Vocabulary for annotating vocabulary descriptions. Ian Davis, 2005.

http://vocab.org/vann/.html
[VOAF] Vocabulary of a Friend http://labs.mondeca.com/vocab/voaf
[Vocab Status] An RDF vocabulary for relating SW vocabulary terms to their status. http://www.w3.org/2003/06/sw-vocab-status/ns
[VoID] Vocabulary of Interlinked Datasets http://rdfs.org/ns/void
[VOID-KBExplorer] http://void.rkbexplorer.com/
[W3C Voc] Ontologies - W3C http://www.w3.org/standards/semanticweb/ontology
[WATSON] Exploring the Semantic Web http://watson.kmi.open.ac.uk/WatsonWUI/
[Weibel S. 1997] Weibel, S. (1997). The Dublin Core: a simple content description model for electronic resources. Bulletin of the American Society for Information Science and Technology, Wiley Online Library, p. 9-11.
[WHOIS] whois http://www.kanzaki.com/ns/whois#