FROM RAW PUBLISHED DATA TO INTERLINKED SEMANTIC DATA

Authors	Bernard Vatant (Mondeca) Pierre-Yves Vandenbussche (Fujitsu)
Reviewer	Franck Cotton (INSEE) François Scharffe (INRIA)
Date	22/05/13
Reference	Datalift/2013/D2.2/v2
Version	2.0
Destination	public

PROJET DATALIFT
De la donnée brute publiée vers la donnée sémantique interconnectée

Appel ANR CONTINT 2010 ANR-10-CORD-009
RAPPORT DE RECHERCHE

DATALIFT D2.2
CATALOGUE DE VOCABULAIRES

Auteurs : Bernard Vatant, Pierre-Yves Vandenbussche
Relecteurs : Franck Cotton, François Scharffe

Table des matières

Introduction

Objet du livrable

Le processus d'élévation 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 tâche 2.1 a défini des méthodes pour identifier et sélectionner les vocabulaires intéressants à réutiliser lors du processus d'élévation des données.
- La tâche 2.2 utilise les méthodes d'identification de la tâche précédente pour constituer un catalogue de vocabulaires dont l'utilisation peut être intégrée à la plate-forme Datalift. **C'est la description de ce catalogue, nommé Linked Open Vocabulaires (ci-après LOV) qui fait l'objet du présent document.**
- L'intégration du catalogue à la plate-forme, ainsi les modules de sélection de vocabulaires pertinents pour un jeu de données et de conversion vers ces vocabulaires, sont définis par la tâche 2.4 et décrites dans les livrables correspondants.

Clarification terminologique

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 attachées 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 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 naturel 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 de description de 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 (lightweight 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 :

> *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.

Contenu et structure du catalogue

Entrepôt RDF et interfaces

Le catalogue de vocabulaires LOV se définit d'abord comme un entrepôt RDF contenant un ensemble de vocabulaires sélectionnés comme pertinents au regard des critères définis dans la tâche 2.1. Ces vocabulaires sont enrichis de métadonnées, en particulier les relations explicites entre vocabulaires et l'historique de leurs versions successives.
Le contenu RDF, ainsi que les pages HTML qui en sont automatiquement dérivées, est accessible sur les serveurs de l'OKFN [LOV] sous licence Creative Commons [CC BY 3.0].
Le contenu du catalogue est accessible via des interfaces variées présentées en détail plus bas : pages de navigation HTML enrichies de la représentation graphique des liens entre vocabulaires et de leur historique, SPARQL endpoint, recherche plein texte à facettes, interface de back-office dédiée et API permettant l'intégration du contenu du catalogue par des applications.
L'ensemble du code de ces interfaces est disponible sous Github [LOV-code]. Le contenu du catalogue est intégré dans la plate-forme Datalift via les modules de sélection, conversion et alignement qui sont décrits dans le livrable D2.4.
À sa création en mars 2011, le prototype de LOV a été développé et hébergé sur les serveurs de la R&D de Mondeca (Mondeca Labs). Devant le succès rapide et l'intérêt manifesté par l'ensemble de la communauté au-delà de Datalift, le catalogue est devenu en juillet 2012 un projet hébergé par l'Open Knowledge Foundation [OKFN]. Le sous-domaine lov.okn.org héberge depuis cette date l'entrepôt RDF de LOV et son SPARQL endpoint, la publication HTML du catalogue, ainsi que les interfaces de recherche et de gestion du back-office.
En mai 2013, le catalogue LOV décrit environ 350 vocabulaires, dont l'agrégation cumule plus de 25000 termes décrits et plus de 200000 triplets. Les libellés des éléments de

vocabulaires sont majoritairement en anglais (ou sans langue déclarée), mais au total 23 langues sont représentées.

Identification : URI, espaces de noms et préfixes

Chaque vocabulaire est identifié par une URI, et définit un espace de noms qui est lui-même une URI. Dans le meilleur des mondes sémantiques, ces deux URI (qui peuvent être distinctes ou non) sont déclarées explicitement et formellement par le vocabulaire lui-même, dans la description d'un élément owl:Ontology. Ces URI sont également en principe déréférençables et permettent d'accéder à l'ontologie formelle et à sa documentation via négociation de contenu. Cette même ontologie doit être accessible à partir de l'URI de tout élément du vocabulaire.
Mais la réalité des pratiques est beaucoup plus contrastée : l'élément owl:Ontology n'est pas toujours présent, ou bien son contenu est insuffisant ou incohérent avec l'URL à laquelle l'ontologie est effectivement disponible, ou encore la négociation de contenu n'est pas conforme aux bonnes pratiques, etc. Si on devait appliquer strictement les conditions d'admission dans le catalogue définies dans le D2.1, beaucoup de vocabulaires à contenu intéressant auraient été rejetés pour l'une quelconque des raisons précédentes. On a préféré en général inclure ces vocabulaires, et si possible contacter les éditeurs de vocabulaires (quand c'est possible) pour qu'ils remédient aux défauts constatés. Ce travail a permis d'améliorer la situation pour un nombre non négligeable de vocabulaires. Un travail complémentaire est en cours pour classifier les pratiques (bonnes ou moins bonnes) en ce domaine, et utiliser cette classification dans une évaluation globale de la qualité d'un vocabulaire. Ce critère pourra être ajouté à la description du vocabulaire dans LOV, ou utilisé par d'autres services d'évaluation des ontologies [OOPS].

Les préfixes représentent un élément d'identification « court » des vocabulaires. Bien que n'étant au départ en principe que de simples raccourcis techniques à l'intérieur d'un fichier, ils sont devenus pour les plus populaires d'entre eux des abréviations standard de fait, et dans la structure du catalogue, nous utilisons également ces préfixes comme des identifiants courts des vocabulaires, autrement dit des « codes » uniques. Ils sont utilisés dans la liste des vocabulaires présentée dans la page d'accueil du catalogue, dans les URL des pages html de chaque vocabulaire, dans les représentations graphiques, etc. Dans la mesure du possible, nous utilisons les préfixes recommandés par les vocabulaires eux-mêmes, s'ils existent et dans la mesure où ils n'entrent pas en conflit.
Nous avons également harmonisé dans la mesure du possible les préfixes utilisés dans LOV avec ceux qui sont soumis au service http://prefix.cc bien que ce dernier n'impose pas de contrainte de relation bi-univoque entre préfixe et URI. Ce travail d'harmonisation vise là encore à utiliser LOV pour susciter de bonnes pratiques collaboratives dans la publication et l'usage des vocabulaires.

Métadonnées et relations entre vocabulaires

L'objectif initial et majeur du catalogue reste de mettre à disposition pour chaque vocabulaire pertinent et bien identifié, non seulement son contenu brut, mais aussi une quantité et une qualité de métadonnées suffisante pour supporter efficacement les tâches de recherche et de sélection des producteurs de données. Ces métadonnées

peuvent être explicitement présentes dans les vocabulaires à la source, ou enrichies lors de l'insertion du vocabulaire dans LOV. Nous reviendrons dans la section « Gestion du catalogue » sur le processus d'enrichissement.
Lorsque nous avons trouvé dans les vocabulaires existants des propriétés correspondant aux métadonnées que nous souhaitions représenter, nous les avons bien sûr réutilisées. En particulier le catalogue utilise les vocabulaires suivants. Pour chaque vocabulaire le préfixe usuel est indiqué entre parenthèses.

- Dublin Core Terms (dcterms) est un vocabulaire pour la description de métadonnées sur des documents électroniques [Weibel1997]. Nous utilisons entre autres les propriétés dcterms:title pour décrire le titre; dcterms:modified pour indiquer la date de dernière mise à jour ou encore dcterms:creator pour préciser le ou les créateurs d'un vocabulaire;
- Vocabulary of Interlinked Datasets (void) est 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 [Alexander *et al.*, 2009];
- Vocabulary for ANNotating vocabulary (vann) est un vocabulaire pour la description de vocabulaires [Vann]. Nous utilisons ses propriétés vann:preferredNamespacePrefix et vann:preferredNamespaceUri pour représenter le préfixe à utiliser et l'espace de nom 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 [DArcus2010].

Les vocabulaires cités ci-dessus ne permettent cependant pas de décrire comment les vocabulaires dépendent les uns des autres. La réutilisation au sein d'un vocabulaire d'autres vocabulaires est un indicateur intéressant à plusieurs titres. Il peut indiquer une application du principe de modularité si les vocabulaires réutilisés sont dans le même espace de noms que le vocabulaire courant. Mais plus généralement il témoigne d'une bonne pratique, le souci de l'auteur à intégrer l'existant et à ne pas réinventer. Il s'agit autant d'un principe de simplicité et d'économie que d'une bonne pratique sociale.
L'approche utilisée pour la description de relations entre vocabulaires est similaire au travail du Linked Open Data pour les données. De même que le vocabulaire VoID sert à décrire les relations entre jeux de données, le vocabulaire Vocabulary of a Friend [VOAF], développé dans le cadre de Datalift, permet de décrire les relations entre vocabulaires, par l'intermédiaire de propriétés spécifiques.
Les classes de VOAF ont été créées spécifiquement pour représenter les vocabulaires et les groupes de vocabulaires. La classe voaf:Vocabulary, sous-classe de void:Dataset, représente un vocabulaire utilisé ou utilisable dans le Web de données. De façon récursive, une instance de voaf:Vocabulary s'appuie sur ou est utilisée par au moins une autre instance de voaf:Vocabulary. Un vocabulaire totalement isolé n'est donc pas a priori candidat à entrer dans cette classe.
En ce qui concerne les relations entre vocabulaires, VOAF définit huit types de relations entre vocabulaires qui spécialisent également le vocabulaire VoID :

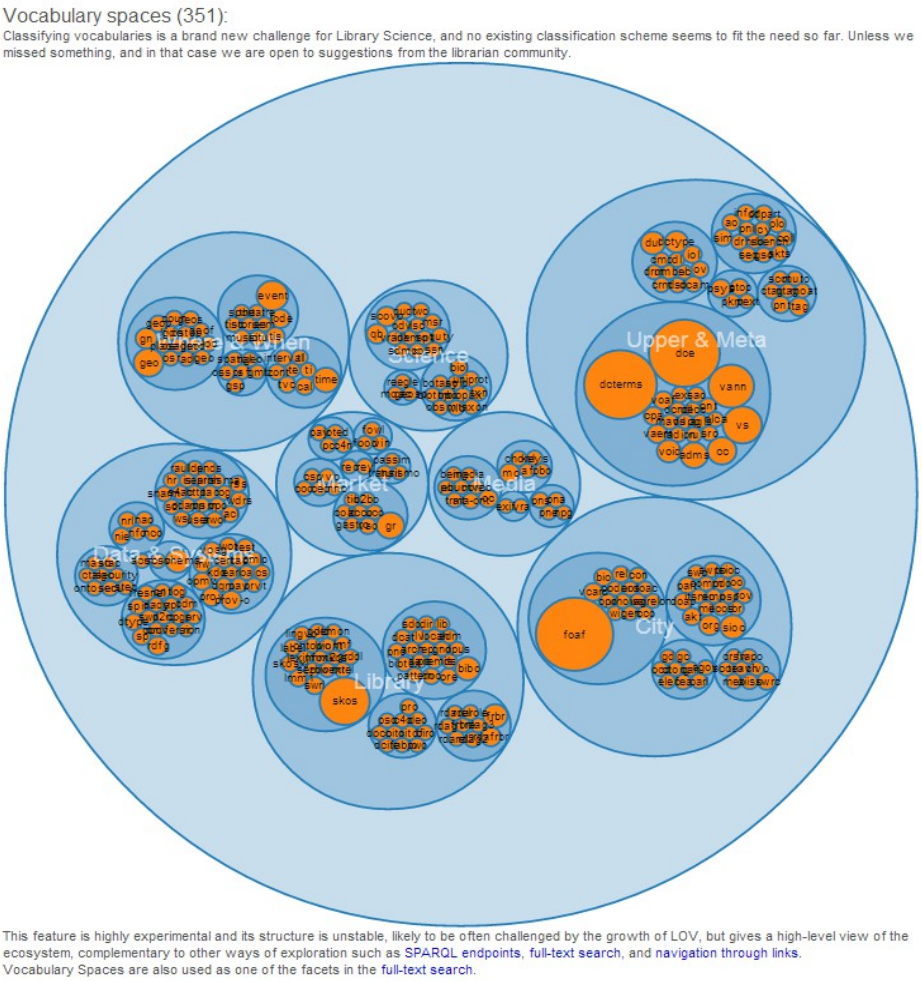
- voaf:reliesOn. Indique qu'un vocabulaire sujet utilise (sens large) des classes ou des propriétés d'un vocabulaire objet. Cette relation étend la propriété void:vocabulary et dcterms:references. C'est une propriété générique de lien entre deux vocabulaires.

- voaf:usedBy. Indique qu'un vocabulaire sujet est utilisé (sens large) par un vocabulaire objet. Cette propriété est l'inverse de voaf:reliesOn.
- voaf:metadataVoc. Indique que le vocabulaire objet est utilisé pour décrire des métadonnées du vocabulaire sujet.
- voaf:extends. Indique que le vocabulaire sujet étend l'expressivité du vocabulaire objet en déclarant des relations de subsomptions, en limitant le domaine ou le co-domaine de propriétés, etc.
- voaf:specializes. Indique que le vocabulaire sujet définit une ou plusieurs sous-propriétés ou sous-classes du vocabulaire objet.
- voaf:generalizes. Indique que le vocabulaire sujet définit une ou plusieurs super-propriétés ou super-classes du vocabulaire objet.
- voaf:hasEquivalencesWith. Indique que le vocabulaire sujet déclare une ou plusieurs propriétés ou classes équivalentes à celles du vocabulaire objet.
- voaf:similar. Indique que deux vocabulaires ont un objectif et/ou une portée similaire. Cette propriété n'exclut pas la définition des autres relations ci dessus.

D'autres propriétés spécifiques sont définies par le vocabulaire VOAF, telles que l'indication du nombre de classes et propriétés d'un vocabulaire.

Classification des vocabulaires

Figure 1 : Les espaces de vocabulaires dans LOV



Le nombre de vocabulaires recensés atteignant maintenant plusieurs centaines, comme pour toute représentation de connaissances qui doit être intelligible par des humains (pour la maintenance et l'utilisation), il nous a semblé raisonnable d'organiser les vocabulaires en les groupant par une classification par domaines d'utilisation [Rothenberg1989].
Cependant aucune classification des vocabulaires n'existe dans l'état de l'art. Nous avons donc défini la classe voaf:VocabularySpace (groupe de vocabulaires). Un groupe de vocabulaires définit un ensemble pertinent de vocabulaires, regroupés selon un caractère commun : objectif commun, publication par la même source, etc. Un vocabulaire peut appartenir à un ou plusieurs espace(s) de vocabulaires. Les propriétés Dublin Core dcterms:isPartOf et dcterms:hasPart sont utilisées pour lier un vocabulaire à un espace de vocabulaires, et pour hiérarchiser les espaces entre eux.
Ce travail de classification, assez expérimental pour le moment, a été motivé par l'attente des modélisateurs qui veulent rapidement savoir quels vocabulaires existent pour décrire les données de leur domaine dans le cadre de leur cas d'utilisation. Par exemple le groupe "People" réunit les vocabulaires permettant de décrire des personnes comme [VCARD] ou encore [WHOIS].
La figure 1 ci-dessus présente l'interface graphique de ces espaces sur la page d'accueil de LOV.

Métriques d'utilisation des vocabulaires

Une métrique importante identifiée dans le livrable 2.1 est l'évaluation de la quantité de données utilisant un vocabulaire, indicateur qui pourrait apparaître dans les métadonnées du vocabulaire. Pour évaluer cette information nous avons commencé à utiliser les informations qualitatives renseignées sur [CKAN] ou utiliser des serveurs de description VOID [VOID-Browser], [VOID-KBExplorer]. Mais les informations fournies par ces sources ne sont ni complètes ni vraiment fiables.
Nous avons expérimenté également l'utilisation du « cache » du LOD Cloud fourni par OpenLink Software [LOD-cache] pour extraire des statistiques d'utilisation des différents vocabulaires dans le Web des données. Un développement prévu est d'utiliser les statistiques fournies par les descriptions VoID des jeux de données publiées par le projet LOD2 [LOD2Stat].
Cependant dans leur état actuel de telles métriques sont à utiliser avec précaution et restent indicatives, certains vocabulaires très spécialisés pouvant être de très grande qualité mais peu ou pas encore utilisés, certains types de données biaisant les statistiques en accumulant un très grand nombre de triplets, etc.

Interfaces du catalogue

Le catalogue peut être consulté à travers des interfaces de recherche, de navigation et de requêtage, et son contenu peut être intégré à des applications, en particulier bien sûr la plate-forme Datalift, via une API. Une interface Web de gestion back-office a également été développée à destination des curateurs du catalogue.

Interfaces de navigation et de recherche

Navigation

La page d'accueil du catalogue http://lov.okfn.org présente les vocabulaires sous forme de liste ordonnée par ordre alphabétique des préfixes, ainsi que la représentation graphique de la classification (groupes de vocabulaires) présentée plus haut.
La description de chaque vocabulaire est présentée dans une page HTML générée dynamiquement à partir de la description RDF du vocabulaire. Cette page présente :

- Les métadonnées du vocabulaire
- Une représentation graphique des relations avec d'autres vocabulaires
- Une « time line » des versions, à partir de laquelle celles-ci peuvent être téléchargées
- Un lien vers la description RDF correspondante

La figure 2 ci-dessous représente la page de publication de l'ontologie Geonames.

Recherche plein texte

Le LOV-search, accessible depuis la page d'accueil ou depuis la page dédiée http://lov.okfn.org/dataset/lov/search/ permet une recherche plein texte sur l'ensemble des vocabulaires agrégés dans le triple store. Un algorithme de classement est utilisé pour ordonner les résultats sur la base des statistiques de leur réutilisation par les autres vocabulaires et dans le Web de données. Les résultats peuvent être filtrés par différentes facettes, comme les types d'élément (classe ou propriété), ou les espaces de vocabulaire.

SPARQL endpoints

Deux SPARQL endpoints ont été mis en place pour requêter les données.

Le Linked Open Vocabularies Endpoint à http://lov.okfn.org/endpoint/lov interroge un triple store qui contient les vocabulaires, leurs métadonnées et leurs versions.

Le LOV Aggregator Endpoint http://lov.okfn.org/endpoint/lov_aggregator interroge un triple store qui agrège automatiquement les contenus des vocabulaires.

Pour simplifier les interfaces et la compréhension par les usagers, il est prévu de stocker toutes les données dans le même triple store et de ne conserver dans l'avenir qu'un seul endpoint. On pourra interroger ces données en distinguant par deux graphes distincts ce qui provient des vocabulaires eux-même et les métadonnées ajoutées dans la gestion de LOV . Par exemple un préfixe peut être déclaré dans le vocabulaire source, ou choisi par le curateur de LOV.

API

Une API a été développée qui permet d'intégrer dans des applications toutes les fonctionnalités du LOV-Search. Cette API est utilisée pour l'intégration du catalogue dans la plate-forme Datalift.

L'API est documentée à l'adresse http://lov.okfn.org/dataset/lov/apidoc/

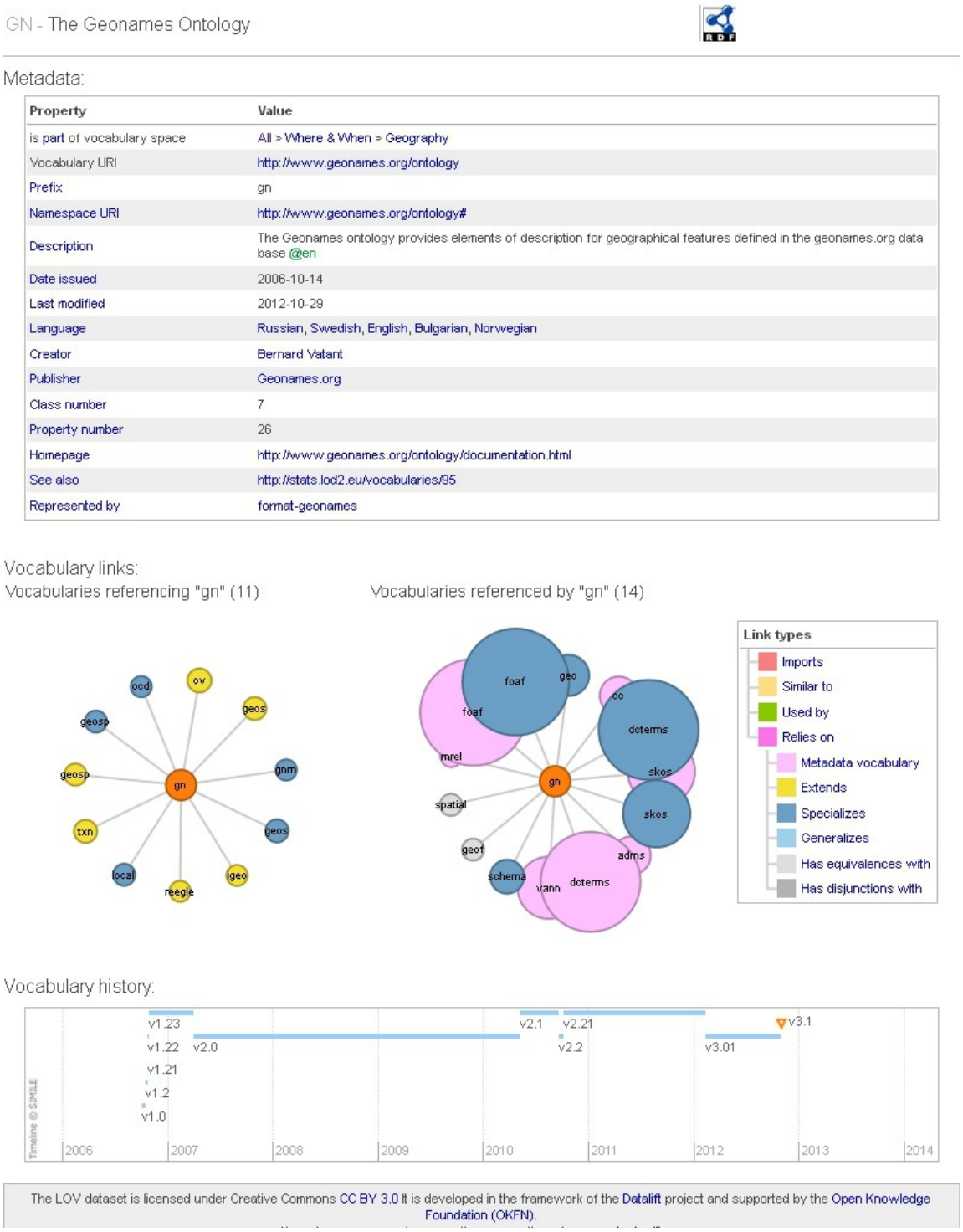


Figure 2 : Page de publication de l'ontologie Geonames

Interfaces de gestion

Les interfaces de gestion permettent aux curateurs de LOV d'intégrer de nouveaux vocabulaires dans le catalogue, de mettre à jour leurs métadonnées, de gérer la classification et les versions. Le travail initial de construction du catalogue LOV des vocabulaires liés a été manuel, sur la base des vocabulaires les plus populaires dans le Web des données : FOAF, SKOS, et Dublin Core. Sur cette base restreinte un mode

d'extension naturel et récursif du catalogue est d'utiliser les liens entre vocabulaires. Si un vocabulaire V1 est intégré dans LOV et qu'il s'appuie sur des éléments d'un vocabulaire V2, ce dernier est a priori candidat à être intégré dans LOV. V2 peut lui-même s'appuyer sur V3, etc.

Cette première phase a permis de réunir une première base de plusieurs dizaines de vocabulaires. Les extensions suivantes se sont faites par un travail de veille et d'exploration pragmatique du Web des données pour commencer, et assez rapidement par les suggestions des éditeurs de vocabulaires eux-mêmes, le catalogue publié étant vite devenu populaire dans la communauté Linked Data.

Sélection et mise à jour

Pour assister le travail de mise à jour et compte tenu du fait que les vocabulaires peuvent évoluer de manière indépendante sans que l'on en soit averti, a été mis en place un programme automatique de veille appelé "LOV-Bot". Ce programme a pour objectif de confronter les données contenues dans le LOV aux données des vocabulaires. En cas de différences constatées, une annotation est ajoutée à ce vocabulaire dans le back-office, et si les modifications détectées sont validées, une nouvelle version est créée et stockée dans le catalogue.
Le LOV-Bot permet également la détection automatique des relations entre vocabulaires. A chaque type de relation définie dans VOAF est associée une requête SPARQL. Toutes les requêtes utilisées sont explicitées dans la documentation du vocabulaire VOAF. Par exemple la relation voaf:specializes est automatiquement détectée par le LOV-Bot en utilisant la requête SPARQL suivante.

```
SELECT distinct ?generProp ?generClass
WHERE{
    {
       ?spec rdfs:subPropertyOf ?generProp.
       FILTER(!isBLank(?generProp))
    }UNION{
       ?spec rdfs:subClassOf ?generClass.
       FILTER(!isBLank(?generClass))
    }
}
```

Le site de LOV permet aux utilisateurs de suggérer un vocabulaire. Dans ce cas l'URI est soumise à l'analyse du LOV-Bot et les résultats transmis au curateur.
Dans tous les cas de figure, les suggestions du LOV-Bot sont validées manuellement par les administrateurs du catalogue, via l'interface de back-office dédiée. Cette interface permet de :

- Éditer les métadonnées de chaque vocabulaire, leurs liens et leurs versions
- Gérer la classification des vocabulaires
- Gérer la base des agents (auteurs, contributeurs et éditeurs de vocabulaires)
- Soumettre au LOV-Bot un seul ou la totalité des vocabulaires
- Régénérer sur demande l'ensemble de la publication HTML du catalogue après un ensemble de modifications.

Représentation et stockage des versions

Depuis juillet 2012 nous avons commencé à stocker les versions successives des

vocabulaires du catalogue, que celles-ci soient explicitement déclarées par la publication (ce qui n'est malheureusement pas encore la pratique la plus répandue), ou constatées par des modifications détectées par le LOV-Bot. Ce stockage permet d'une part de pallier l'absence de politique explicite des versions du côté des producteurs de vocabulaires eux-mêmes, et par ailleurs de visualiser l'historique dans les interfaces de navigation. Il permet aussi bien sûr d'avoir accès à des vocabulaires qui ne sont plus (provisoirement ou définitivement) disponibles en ligne.

Pour la représentation des versions nous avons choisi le modèle FRBR [FRBR].

- Chaque vocabulaire, identifié par son URI pérenne, est défini comme une instance de frbr:Work.
- Chaque version est une instance de frbr:Expression, reliée à l'œuvre par la propriété frbr:realization.
- Chaque fichier spécifique ou format représentant la version (Turtle, RDF-XML…) est une instance de frbr:Manifestation, reliée à la version par la propriété frbr:embodiment.

La pertinence d'appliquer le modèle FRBR à des vocabulaires a pu paraître douteuse a priori, en particulier dans le monde bibliothécaire qui considère les vocabulaires comme des outils destinés à classer et organiser les ressources bibliothécaires classiques (livres et autres documents), et non comme des ressources à cataloguer. Mais dans LOV nous cataloguons bien les vocabulaires eux-mêmes, et nous les avons donc considérés comme des œuvres au même titre que toute création de l'esprit humain. La grille FRBR semblait alors tout à fait pertinente.
Cependant nous étudions actuellement la possibilité de décrire les vocabulaires et leurs versions en utilisant plutôt le vocabulaire ADMS, en cours de développement dans le cadre du projet Joinup de la Communauté Européenne [ADMS] [Joinup]. Ce vocabulaire définit la notion générale de « Semantic Asset », dont chaque version d'un vocabulaire est une instance particulière. Ce vocabulaire n'est pas tout à fait compatible avec FRBR cependant, donc la question reste en discussion à ce jour.

Perspectives

Travaux en cours

La dernière année du projet doit voir la réalisation d'un certain nombre de tâches en cours.

- Amélioration des outils de gestion du catalogue, pour aboutir à une véritable plate-forme collaborative ouverte à tous les acteurs intéressés par la vie et la bonne santé de l'écosystème des vocabulaires.
- Intégration des statistiques de l'utilisation des vocabulaires dans le Web des données, en s'appuyant sur les résultats de LOD2.
- Intégration des critères d'évaluation de la qualité des vocabulaires issus du travail en cours en collaboration avec le projet OOPS.
- Poursuite de l'intégration dans Datalift via l'API, en particulier alimentation avec le contenu du catalogue des outils d'alignements et de mesure de similarité. Ces outils apporteront une valeur ajoutée au contenu du catalogue en facilitant encore les tâches de sélection et d'évaluation des vocabulaires.

- Intégrer dans la gestion de LOV d'autres services en ligne liés aux vocabulaires, comme [prefix.cc] déjà cité, ou [TripleChecker] qui permet de rechercher des erreurs ou typos dans des vocabulaires.

Vers une gouvernance des vocabulaires

L'approche novatrice apportée par le catalogue LOV, la vue d'ensemble sans précédent qu'il offre sur l'écosystème des vocabulaires, la réflexion qu'il a suscitée parmi les producteurs de vocabulaires et la communauté Linked Data en général ont dépassé largement le cadre contractuel d'un simple composant de la plate-forme Datalift. Plus qu'un simple catalogue, LOV se positionne donc comme un outil de gouvernance, et il est d'ores et déjà au cœur d'initiatives importantes visant à une gestion durable en ce domaine, en collaboration avec des acteurs majeurs comme le W3C, le DCMI, la Communauté Européenne à travers le projet Joinup, les grandes bibliothèques regroupées au sein de l'OCLC, et bien sûr la fondation OKFN qui a accueilli le projet avec enthousiasme et le supporte techniquement avec efficacité.
Par exemple la conférence Dublin Core 2013 [DC 2013] comprendra une session spéciale consacrée au sujet de la conservation et gouvernance à long terme des vocabulaires [VocPres]. Les créateurs du LOV ont été naturellement invités à animer cette session.
A travers cette dynamique nous espérons voir se développer les utilisations de LOV au-delà de Datalift, et trouver les moyens de pérenniser sa maintenance et son évolution. Ce dernier point est le plus crucial. Il serait dommage d'avoir à figer ce catalogue dans un livrable de fin du projet, faute de moyens pour continuer à le faire vivre.

Références

[ADMS] http://www.w3.org/ns/adms

[Alexander2009] Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets. In: Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009). (2009)

[Bizer2009] Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. Int. J. Semantic Web Inf. Syst. 5 (2009) 122.

[CC BY 3.0] Licence Creative Commons 3.0 http://creativecommons.org/licenses/by/3.0/

[CKAN] http://ckan.net/

[DArcus2010] D'Arcus, B., Giasson, F.: Bibliographic ontology specication. Retrieved October 8 (2010)

[DC 2013] DC-2013, Lisbon, Portugal. "Linking to the Future" http://dcevents.dublincore.org/index.php/IntConf/dc-2013/schedConf/

[Heath2011] Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space: Theory and Technology. Volume 1. Morgan & Claypool Publishers (2011)

[Klyne2004] Klyne, G., Carroll, J.: Resource description framework (rdf): Concepts and abstract syntax. w3c recommendation, 10 february 2004. World Wide Web Consortium (2004)

[FRBR] Expression of Core FRBR Concepts in RDF http://purl.org/vocab/frbr/core

[Joinup] http://joinup.ec.europa.eu/

[LOD] Linked Open Data http://linkeddata.org

[LOD-cache] http://datahub.io/dataset/openlink-lod-cache

[LOD2Stat] http://stats.lod2.eu/

[LOV] Linked Open Vocabularies http://lov.okfn.org/dataset/lov/index.html

[LOV-code] Code source LOV https://github.com/pyvandenbussche

[LOV-Endpoint] Linked Open Vocabularies SPARQL Endpoint http://lov.okfn.org/endpoint/lov

[LOV-RDF] Linked Open Vocabularies RDF dump http://lov.okfn.org/dataset/lov/lov.rdf

[OKFN] Open Knowledge Foundation http://okfn.org

[OOPS] OOPS! (OntOlogy Pitfall Scanner!) http://oeg-lia3.dia.fi.upm.es/oops/index-content.jsp

[prefix.cc] http://prefix.cc

[Rothenberg1989] Rothenberg, J., Agency, U.S.D.A.R.P.: The nature of modeling. Citeseer (1989)

[SchemaC] Schema-Cache http://schemacache.test.talis.com/

[SchemaP] Schemapedia http://schemapedia.com/

[TripleChecker] http://graphite.ecs.soton.ac.uk/checker/

[Tummarello2007] Tummarello, G., Delbru, R., Oren, E.: Sindice. com: Weaving the open linked data. In: Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference, Springer-Verlag (2007) 552-565.

[Vann] A vocabulary for annotating vocabulary descriptions http://purl.org/vocab/vann/

[VCARD] http://www.w3.org/2006/vcard/ns#

[VocPres] DC 2013 Speciel Session : Long-term Preservation and Governance of RDF Vocabularies http://dcevents.dublincore.org/IntConf/index/pages/view/vocPres

[VOID-Browser] http://kwijibo.talis.com/voiD/

[VOID-KBExplorer] http://void.rkbexplorer.com/

[Weibel1997] Weibel, S.: The dublin core: a simple content description model for electronic resources. Bulletin of the American Society for Information Science and Technology 24 (1997) 911

[WHOIS] whois http://www.kanzaki.com/ns/whois#