Les référentiels : typologie et interopérabilité

Antoine Isaac

Les référentiels
Typologie et interopérabilité

Antoine Isaac

Coordinateur scientifique d'Europeana depuis 2009 et chercheur à la VU University Amsterdam depuis 2005, Antoine Isaac travaille, depuis son doctorat à la Sorbonne et l'Institut national de l'audiovisuel (INA), sur les technologies du web sémantique pour la représentation et l'interopérabilité des collections patrimoniales et de leurs référentiels. Il a aussi travaillé sur SKOS pour le groupe de travail Semantic Web Deployment du W3C et a co-dirigé l'incubateur Library Linked Data du W3C. http://www.few.vu.nl/~aisaac, aisaac@few.vu.nl

Dans ce chapitre, nous allons essayer de dégager comment certaines ressources publiées sur le web de données peuvent prétendre au statut de référentiel et, peut-être plus que d'autres, apporter une réponse aux enjeux présentés dans le chapitre précédent : interopérabilité, ouverture et réutilisation des données.

Une première observation s'impose : sur le web de données, il n'y a pas de définition stricte de ce qu'est un référentiel. De fait, d'un point de vue technique, tout artefact est susceptible de contenir des ressources dont la description et la publication ont fait l'objet de suffisamment de soins pour être réutilisés afin de structurer ou peupler des données produites ou échangées par d'autres que leur créateur initial.

Pourtant, des réalités moins techniques influencent évidemment la manière dont évolue ce web de données : on ne peut pas ignorer l'expertise accumulée par certains acteurs dans un domaine particulier, ni le haut niveau de qualité de certains jeux de données. Certains artefacts disponibles sont donc « plus égaux que d'autres »…

Tout d'abord, nous préciserons ce que la notion de référentiel peut recouvrir dans ce contexte. À savoir, des artefacts de type :
- éléments de métadonnées (ontologies formalisées dans des langages comme OWL) ;
- vocabulaires de valeurs et systèmes d'organisation des connaissances (thésaurus, fichiers d'autorités, etc.) ;
- autres jeux de données sur les « objets du monde réel ».

Pour chacune de ces catégories, nous discuterons ensuite les caractéristiques de référentiels typiques et ce que la technologie du web sémantique (en particulier le rôle crucial des URIs) permet de changer par rapport aux approches plus traditionnelles. Par exemple :
- pour les éléments de métadonnées, les possibilités de réutilisation, voire d'« édition distribuée » ;
- pour les thésaurus et autres systèmes d'organisation des connaissances, la transition d'une approche orientée termes (ou noms) à une approche orientée concepts, voire « entités du monde réel ».

Nous insisterons en particulier sur les aspects de réutilisation et d'interopérabilité des référentiels au niveau sémantique. Notamment ce à quoi la notion d'« alignement » peut faire référence, ainsi que différents scénarios de réutilisation, par exemple au travers du concept de « profil d'application » pour les éléments de métadonnées.

1 Différents types de référentiels

Dans l'environnement qui nous intéresse, plusieurs types d'artefacts peuvent prétendre au statut de référentiels. Nous reprenons dans ce chapitre la classification du groupe de travail du W3C *Library Linked Data* qui cherche à relier les notions du domaine des bibliothèques à celles de la communauté du web de données [14].

● *Ontologies ou éléments de description de métadonnées*. Les artefacts de cette catégorie fournissent les classes et propriétés qui servent de support à l'expression des descriptions. Ils fournissent également les définitions formelles (axiomes, règles de raisonnement) qui peuvent être employées par des moteurs d'inférence pour déduire de nouveaux faits ou détecter des incohérences dans les jeux de données. Dans la *semantic web stack* du W3C, cet ensemble recouvre surtout les ontologies exprimées à l'aide des langages RDFS et OWL – on parle souvent de « vocabulaires RDF ». FOAF[1], les éléments de Dublin Core[2], CIDOC-CRM[3], FRBR[4] ou le vocabulaire de schema.org peuvent être particulièrement utiles dans un contexte documentaire.

● *Vocabulaires de valeurs ou vocabulaires d'autorité*. Ces artefacts regroupent les conversions, dans le cadre du web de données, de thésaurus, systèmes de classification, listes d'autorité. Ils s'agit parfois de bases de connaissances très volumineuses, telles que la liste des vedettes matières de la Bibliothèque du Congrès

[1] http://xmlns.com/foaf/0.1

[2] http://dublincore.org/documents/dcmi-terms

[3] http://www.cidoc-crm.org

[4] http://iflastandards.info/ns/fr/frbr/frbrer

(LCSH)[5], le Répertoire d'autorité-matière encyclopédique et alphabétique unifié (RAMEAU)[6], AGROVOC[7], le Fichier d'autorité international virtuel (VIAF)[8], la Classification décimale Dewey[9] et GeoNames[10].

● *Jeux de données de référence* : ce sont de simples « graphes » de données, au sens RDF. Un jeu de données consiste en une série de descriptions de ressources pertinentes pour un domaine ou une application spécifique. D'un point de vue technique, sur le web de données, tout jeu peut être réutilisé, relié à d'autres ressources, faire l'objet d'une sélection arbitraire pour un usage particulier. Certains jeux, par leur portée, leur taille, leur provenance, font néanmoins de meilleurs candidats référentiels. DBpedia[11], Freebase[12] ou, dans le domaine des bibliothèques, crossref.org, data.bnf.fr ou Europeana[13], sont quelques exemples possibles.

Les première et troisième catégories suivent les distinctions déjà présentées dans le chapitre 3. La seconde correspond à une tradition préexistante de création de référentiels, en particulier dans le domaine documentaire.

Dans les trois sections suivantes nous allons discuter ces catégories, en présentant les caractéristiques de référentiels typiques et ce que le web de données change par rapport aux approches traditionnelles de conception et d'utilisation de ces artefacts.

Mais, au préalable, le lecteur doit être conscient que nous n'avons pas affaire ici à des catégories exclusives. D'abord, du point de vue technique de la représentation en RDF, rien ne distingue un vocabulaire de valeurs d'un autre jeu de données. Comme nous allons le voir, tous deux sont composés de descriptions d'« instances de classes » au sens RDFS/OWL. La différence, subtile et subjective, sera basée sur la façon dont leurs ressources (ou une partie d'entre elles) sont utilisées ou sur celle dont les descriptions se concentrent sur des éléments particuliers (« concepts » ou « libellés » pour les vocabulaires de valeurs). La perception d'un artefact par une communauté particulière et sa provenance jouent également un rôle important.

Ensuite, il est aussi communément accepté qu'une ontologie, en plus de définir des classes et des propriétés, introduise aussi un certain nombre d'« instances de référence » qui sont utiles à un domaine ou une application dans son ensemble, de la même façon que pourrait le faire un vocabulaire de valeurs ou un simple jeu de données. Par exemple, une classe « couleur » pourra se voir adjoindre l'énumération de ses instances admises (bleu, rouge, etc.).

Finalement, dans la tradition académique du web sémantique, la notion d'ontologie a souvent inclus des artefacts qui ne correspondent pas forcément, d'un point de vue fonctionnel et technique, aux jeux de classes et propriétés munies d'une sémantique formelle telle qu'assignée par RDFS et OWL. En particulier, thésaurus et systèmes de classification peuvent prétendre au statut de « système d'organisation de connaissances » qui permet d'articuler ensemble des concepts en utilisant des liens « sémantiques » semblables en première analyse à ceux utilisés dans une ontologie. Certains rapports font donc état d'un « spectre d'artefacts informationnels », comme dans la figure 1.

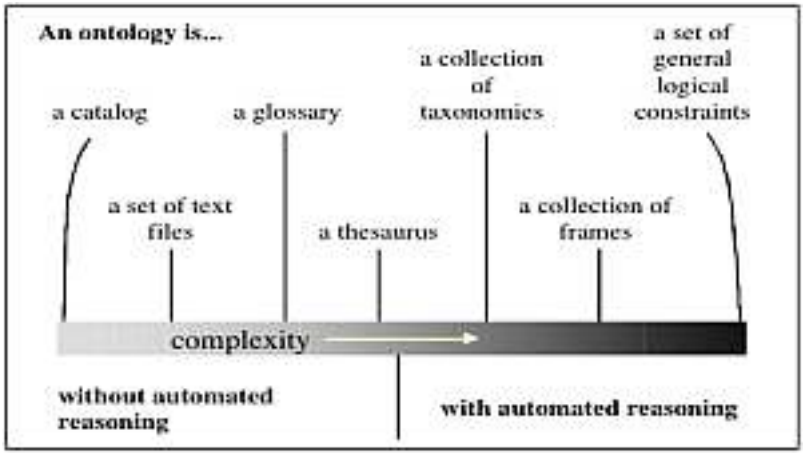

Figure 1 – Artefacts ayant été classés comme ontologies [12]

[5] http://id.loc.gov

[6] http://rameau.bnf.fr

[7] http://aims.fao.org/agrovoc/lod

[8] http://viaf.org

[9] http://dewey.info

[10] http://geonames.org

[11] http://dbpedia.org

[12] http://freebase.com

[13] http://europeana.eu, données liées à http://data.europeana.eu

La barrière du raisonnement permis par une sémantique formelle est souvent utilisée comme critère de distinction entre les deux familles. Il est cependant vrai que certaines ressources représentées en SKOS sont utilisées pour des formes d'inférence dans des systèmes documentaires…

2 Ontologies et éléments de métadonnées

Comme vu dans le chapitre précédent, les ontologies pour le web de données introduisent et définissent de façon formelle (utilisant les langages RDFS et OWL) les éléments nécessaires à l'expression de (méta)données. Par exemple, on pourra créer deux classes <mied:Personne>, <mied:Peintre> et définir la seconde comme sous-classe de la première, permettant ainsi à un moteur d'inférence de classifier toute instance explicite de peintre comme une instance de personne.

2.1 Une brève typologie

Il existe de nombreux travaux méthodologiques classifiant les ontologies existantes, et des approches permettant de construire ces artefacts.

On trouvera des ontologies « noyaux » qui regroupent les types d'entités de relations de haut niveau communs à plusieurs domaines. DOLCE[14] et CIDOC-CRM sont deux instances typiques de cette catégorie, qui proposent des classes telle que « événement », qui peut être réutilisée dans un grand nombre de domaines liés à la culture. OAI-ORE[15] en est une autre, qui définit des classes et des propriétés utiles à la représentation de ressources qui sont agrégées à partir d'autres entités. Finalement, les éléments du Dublin Core, incluant des propriétés telles que <dcterms:creator>, sont un exemple bien connu dans le contexte documentaire.

Les ontologies « de domaine » sont plus spécifiques, mais gardent un niveau suffisamment général pour être pertinentes à l'échelle d'un domaine. L'ontologie bibliographique Bibo[16], tout comme celle de FRBR ou l'ontologie du W3C pour les ressources médias[17], peuvent être considérées comme des ontologies de domaine.

Les ontologies applicatives sont, elles, beaucoup plus spécialisées, car elles répondent à des besoins applicatifs précis ou créés pour publier des jeux de données spécifiques. On pourra citer le vocabulaire défini par la Bibliothèque nationale allemande pour publier ses données d'autorité, en complément d'autres vocabulaires[18], ou bien l'ontologie C4O (Citation Counting and Context Characterization)[19].

Tout comme celle, plus générale, des référentiels de la section précédente, cette catégorisation n'est évidemment pas stricte – des ontologies comme le vocabulaire de schema.org, par exemple, peuvent dérouter par leur portée très générale associée à des choix de modélisation assez pragmatiques. D'autres classes d'artefacts ont été proposées. Notre objectif est ici surtout de montrer la diversité des ressources disponibles pour exprimer des données, ou pour relier ses propres vocabulaires (comme dans l'exemple du chapitre 3) à des vocabulaires déjà existants, augmentant par là l'interopérabilité d'un jeu de données.

Pour plus d'exemples, nous invitons le lecteur à consulter les répertoires et inventaires d'ontologies disponibles. Il en existe en effet un certain nombre, depuis les moteurs généraux de recherche d'ontologies sur le web de données entier (Sindice[20], Watson[21], Schemapedia[22]) jusqu'aux rapports dédiés à un domaine particulier [14], en passant par les répertoire utilisés par des communautés pour la conception et la publication de leurs propres ontologies (par exemple, le *Open Metadata Registry*[23]). Une ressource intéressante est le projet *Linked Open Vocabularies*[24], qui recense un grand nombre d'ontologies et de données précises sur celles-ci, en particulier sur la manière dont elles sont reliées [voir ci-dessous].

2.2 La réutilisation d'ontologies

À une organisation voulant publier ses données se pose un choix difficile entre créer une nouvelle ontologie *ex nihilo* ou réutiliser une ou des ontologies existantes. La seconde option est préférable du point de vue de l'interopérabilité sémantique : un jeu de données est immédiatement exploitable par les applications qui

[14] http://www.loa.istc.cnr.it/DOLCE.html

[15] http://www.openarchives.org/ore

[16] http://bibliontology.com

[17] http://www.w3.org/TR/mediaont-10

[18] https://wiki.d-nb.de/display/LDS

[19] http://purl.org/spar/c4o

[20] http://sindice.com

[21] http://watson.kmi.open.ac.uk

[22] http://schemapedia.com

[23] http://metadataregistry.org

[24] http://lov.okfn.org/dataset/lov

reconnaissent déjà les ontologies utilisées. Il est également plus facile d'interconnecter des jeux qui partagent le même modèle conceptuel. Cependant, une réutilisation intégrale se heurte au problème de l'incompatibilité des spécifications nécessaires à des applications ou des domaines différents. Une ontologie de domaine peut avoir une couverture trop large, tout en manquant des axiomes formels qui sont valables une application particulière. Inversement, de nombreuses ontologies ont tendance à être trop « engagées » (on parle en anglais de *semantic overcommitment*), leurs spécifications formelles contenant des contraintes qui les rendent inutilisables dans des contextes autres que celui de leur création.

Heureusement, les possibilités techniques du web de données facilitent une approche extrêmement fine et flexible pour la création et la réutilisation d'ontologies. Les classes et propriétés des ontologies pour RDF sont en effet des ressources comme les autres. Munies d'URIs, il est possible de les réutiliser individuellement pour un jeu de données, mais aussi pour la conception de vocabulaires.

Le projet Europeana[25], par exemple, pour son futur modèle de données EDM[26], a réutilisé de nombreux éléments de vocabulaires existants lorsque ces éléments étaient pertinents pour ses propres besoins. En particulier, EDM inclut des classes et des propriétés de OAI-ORE (pour représenter des agrégations de contenus), Dublin Core (pour les métadonnées descriptives de base), SKOS [voir la section suivante], RDA et quelques autres pour la représentation d'« entités contextuelles » (concepts, personnes, lieux, etc.) qui gravitent autour des ressources culturelles auxquelles Europeana doit donner accès. Selon EDM, il est ainsi possible de créer des descriptions comme :

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix edm: <http://europeana.eu/schemas/edm> .
@prefix rdaGr2: <http://rdvocab.info/ElementsGr2/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .

<http://www.mied.fr/personne/Eugene_Delacroix>
        rdf:type edm:Agent ;
        foaf:name "Eugène Delacroix" ;
        skos:altLabel " Ferdinand Victor Eugène Delacroix" ;
        rdaGr2:dataOfBirth "26 avril 1798" .
```

Un agent consommant de telles données, s'il cherche une définition des éléments apparaissant dans une telle description, va pouvoir accéder à cette définition dans les représentations RDF (utilisant RDFS et OWL) qui sont publiées dans les espaces de noms des vocabulaires réutilisés. Encore une fois, le web de données fonctionne suivant le principe de l'hypertexte : un consommateur de données ne perçoit pas de différence entre un lien vers une ressource interne à un « espace d'information » (un graphe de données et une ontologie sur le même serveur) et un lien externe. De façon cruciale, cela vaut aussi pour les triplets RDF qui définissent les classes et propriétés d'une ontologie. Un axiome formel peut ainsi faire référence à une ressource d'une ontologie externe.

L'exemple suivant définit la classe <edm:Agent> comme regroupant des instances qui ne peuvent avoir plus d'une date de naissance, mais en utilisant pour cela une propriété d'un autre espace de noms :

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix edm: <http://europeana.eu/schemas/edm> .
@prefix rdaGr2: <http://rdvocab.info/ElementsGr2/> .

edm:Agent rdfs:subClassOf [
  rdf:type owl:Restriction ;
  owl:maxCardinality "1"^^xsd:nonNegativeInteger ;
  owl:onProperty rdaGr2:dateOfBirth
] .
```

[25] http://europeana.eu

[26] http://pro.europeana.eu/edm-documentation

Ceci permet des scénarios de réutilisation plus complexes que la simple combinaison de vocabulaires existants, suivant l'approche des « profils d'application » proposée par Dublin Core [11].

Une autre possibilité intéressante au regard de l'interopérabilité entre ontologies est l'extension de vocabulaires par spécialisation. EDM, par exemple, inclut une propriété <edm:year> utilisée pour positionner des objets sur une ligne temporelle dont le grain est l'année. Cette propriété peut être définie comme une sous-propriété de l'élément de Dublin Core <dcterms:temporal>. Comme nous l'avons vu dans le chapitre 3, une telle définition permettra de générer, pour chaque triplet utilisant <edm:year>, un triplet plus général reliant les mêmes ressources avec <dcterms:temporal>. Ceci permet de manipuler des données qui répondent aux besoins d'applications précises, tout en laissant ces données accessibles pour un niveau plus général de consommation d'information.

2.3 Aligner des ontologies

Pour de multiples raisons, il n'est parfois pas possible ni souhaitable de réutiliser directement une ontologie existante. Dans ce cas, on peut toujours réaliser un alignement *a posteriori* des éléments des ontologies qui sont sémantiquement comparables. C'est en particulier le rôle dévolu aux axiomes utilisant <owl:equivalentClass> et <owl:equivalentProperty> vus dans le chapitre précédent. Le premier permettra par exemple de déclarer que la classe « agent » d'EDM est sémantiquement équivalente à la classe <E39.Actor> dans l'ontologie CIDOC-CRM.

Se pose alors naturellement la question du ciblage des éléments de métadonnées pertinents pour la réutilisation ou l'alignement. De fait, il existe des efforts essayant de juger de la qualité de la conceptualisation portée par une ontologie. La méthode OntoClean, par exemple, utilise des critères formels (*i.e.* liées à la « forme » des catégories de connaissances auxquelles appartiennent les éléments d'une ontologie) permettant de détecter des « incohérences » dans une ontologie [3]. Dans notre exemple précédent, « Peintre » ne devrait pas être immédiatement déclaré comme sous-classe de « Personne », car la qualité d'être peintre peut être détachée d'un individu, alors qu'une personne demeure personne tout au long de son existence (on dit qu'elle est « rigide ») ; ce qui indique que les deux classes doivent relever de catégories différentes.

Cependant, beaucoup d'ontologies qui sont très souvent réutilisées sont critiquables d'un point de vue formel. De fait, d'autres critères dominent, tels que la simplicité des « patrons » mettant en relation les différentes classes et propriétés d'une ontologie. Ou le respect d'un « engagement sémantique » minimum, les éléments à réutiliser n'étant pas attachés à plus de contraintes formelles que strictement nécessaire, car celles-ci peuvent nuire à l'interopérabilité d'un vocabulaire qui réutilise ces éléments. On peut bien sûr mentionner aussi la disponibilité du vocabulaire, c'est-à-dire sa libre accessibilité en tant que données liées ou encore la présence d'une documentation adéquate. Un bon soutien au niveau organisationnel, par exemple via un éditeur et une communauté d'utilisateurs actifs (comme dans le cas de Dublin Core), est un argument décisif. Les vocabulaires à réutiliser doivent également être eux-mêmes bien connectés à leur écosystème, en réutilisant d'autres vocabulaires comparables ou en établissant si nécessaire des équivalences sémantiques avec ceux-ci.

À l'arrivée, dans un environnement idéal où les concepteurs d'ontologies prendraient soin de maximiser l'interopérabilité des données exprimées à l'aide de leurs artefacts, une ontologie se trouverait donc immergée dans un écosystème de vocabulaires interreliés et dont les définitions bénéficient les uns aux autres. Par exemple, le projet *Linked Open Vocabularies* permet de visualiser certaines des relations s'appliquant à EDM [figure 2].

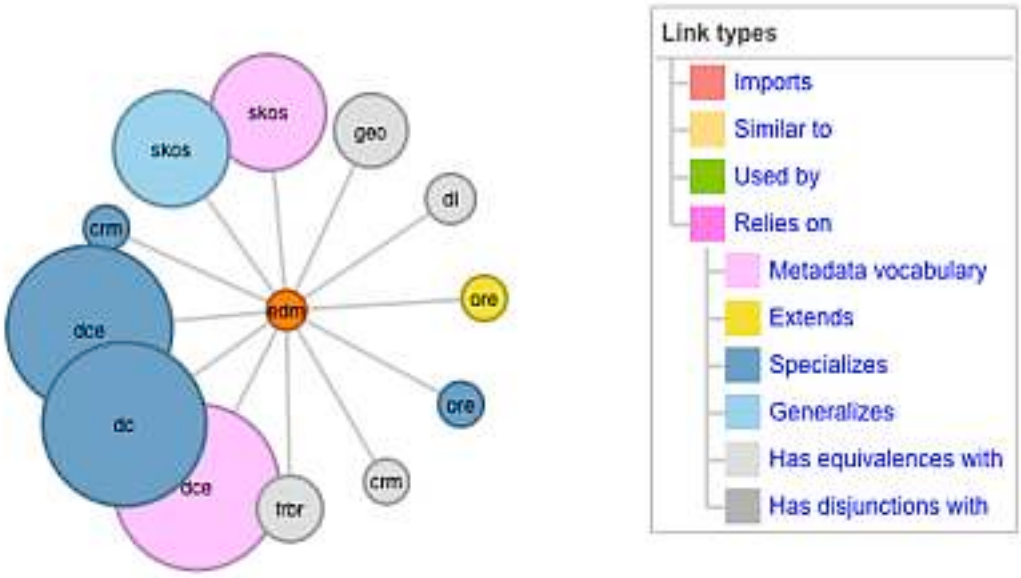

Figure 2 – Relations entre EDM et autres ontologies
(*Linked Open Vocabularies Project*)

Dans les faits, maintenir une liste de liens pertinents s'avère difficile, surtout dans les communautés où l'absence de communication entre créateurs d'ontologies – ou leur manque d'intérêt – aboutit à la prolifération de vocabulaires à la portée similaire. Pour résoudre ce problème, un courant de recherche s'intéresse à l'alignement automatique d'ontologies [2]. Comme dans beaucoup de domaines où existent des outils automatiques, les résultats peuvent être bons, surtout dans le cas d'ontologies de petite taille, riches en axiomes formels et pourvues de libellés en anglais. Et même en cas de performances moins bonnes, ces outils peuvent toujours aider un opérateur humain à orienter des efforts d'alignement manuels, quand les possibilités de départ pour une mise en correspondance sont souvent très nombreuses et l'exploration de ces possibilités, coûteuse en temps.

3 Vocabulaires de valeurs et systèmes d'organisation de connaissances

La seconde catégorie de référentiels, bien établie dans la tradition documentaire, est celle qui comprend listes de vedettes matières (comme RAMEAU), thésaurus (GEMET[27]), systèmes de classification (Classification décimale universelle[28]), listes d'autorité (*Name Authority File* de la Library of Congress[29]), index géographiques (GeoNames) et autres systèmes d'organisation de connaissances – ou KOS pour *knowledge organization systems*. Ces vocabulaires sont souvent le résultat d'efforts patients de collecte et de mise en forme de connaissances par des personnels spécialisés, notamment dans les bibliothèques. Ils rassemblent des termes et concepts pertinents pour un domaine, et sont organisés par des relations terminologiques (synonymie) ou sémantiques (association, généralisation) pouvant être exploitées par un opérateur humain ou un système de recherche documentaire.

3.1 KOS et ontologies

Parce que ces liens, tels que celui de généralisation (« A est plus général que B »), rappellent ceux qui organisent les éléments des ontologies formalisées, les KOS leur sont souvent comparés. Sur le web sémantique, ces vocabulaires jouent cependant un rôle technique différent de celui des ontologies. Il ne s'agit pas de modèles de données ; en particulier, ils ne prescrivent pas la structure de celles-ci. Ils servent en revanche à peupler des données régies par d'autres référentiels (les ontologies, donc), en fournissant de vastes réservoirs de ressources que des créateurs de données peuvent réutiliser en fonction de leurs besoins.

Les éléments de Dublin Core, mentionnés dans la section précédente, peuvent être utilisés pour créer des triplets RDF à partir d'une notice bibliographique : un livre aura un créateur, un sujet, une date de publication (représentés respectivement à l'aide de <dcterms:creator>, <dcterms:subject> et <dcterms:issued>), etc. Un thésaurus fournit, lui, un ensemble contrôlé de valeurs qui apparaissent dans les assertions utilisant certaines de ces propriétés – par exemple, en tant qu'objet d'un triplet utilisant <dcterms:subject> comme prédicat.

Ensuite, les systèmes d'organisation des connaissances hérités du domaine documentaire reposent sur une sémantique différente, très souvent dénuée de toute interprétation formelle. Les relations entre concepts d'un thésaurus sont principalement destinées à des utilisateurs humains ou sont exploitées par des systèmes documentaires plus simples que les moteurs d'inférence dédiés aux raisonnements formels du web sémantique.

Ceci mène parfois à une représentation variable d'un même type de relation. Un lien « partie/tout » pourra être représenté par une relation associative (de type *related* ou *see also*) ou une relation de généralisation (de type *broader*). Cette dernière pouvant aussi servir à représenter des liens hiérarchiques génériques (par exemple entre « oncologie » et « médecine »), ou des liens d'occurrence « instance/classe » (« Le Monde » et « journaux ») ou encore des liens « sous-classe/classe » (« chats » et « mammifères »)…

Il existe un certain nombre de projets, par exemple FinnONTO [4], qui ont construit des ontologies formalisées à partir de KOS en restructurant les informations contenues par celles-ci pour obtenir des liens compatibles avec la sémantique des axiomes RDFS et OWL. Il s'agit cependant d'un travail difficile et de longue haleine si le référentiel de départ est de taille importante – de nombreux KOS contiennent des centaines, voire des milliers d'éléments.

De fait, beaucoup d'applications ne nécessitent pas une formalisation complète de ces référentiels. Des fonctions comme la recherche documentaire, l'appariement de préférences d'utilisateurs, la navigation thématique par facettes, l'aide à la saisie dans un champ texte ou la traduction de requêtes ne nécessitent pas de sémantique « lourde ».

Il est donc toujours pertinent de chercher à publier directement les KOS sur le web de données : ils contiennent souvent des données sur une quantité énorme de ressources et peuvent permettre aux utilisateurs qui accèdent à ces référentiels d'expérimenter à moindre coût avec les technologies « sémantiques ».

---

[27] http://www.eionet.europa.eu/gemet

[28] http://udcdata.info

[29] http://id.loc.gov/authorities/names

3.2 SKOS

SKOS (Simple Knowledge Organization System) est une ontologie qui répond à ce besoin [10]. Son modèle de données, qui se veut simple et compatible avec une majorité d'approches KOS existantes (thésaurus, classifications, etc.), permet de représenter des concepts avec les données qui leur sont le plus couramment rattachées :
- informations terminologiques : libellés préféré ou alternatif, traductions ;
- liens sémantiques entre concepts : relations générique ou associative ;
- notes : notes d'application, définitions, exemples.

La figure 3 présente un exemple extrait de la conversion en SKOS de RAMEAU.

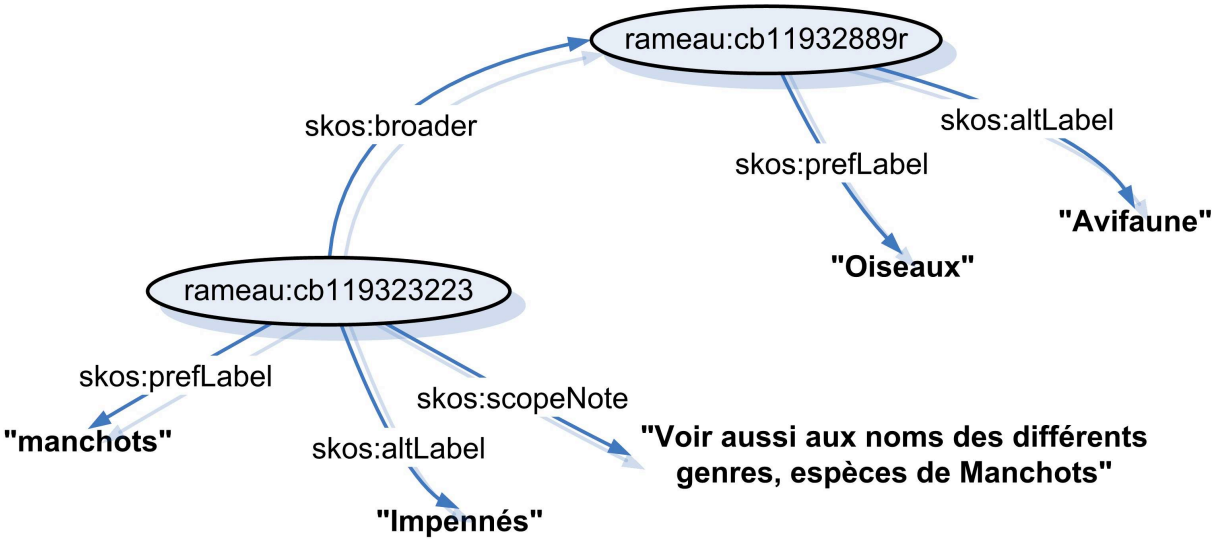

Figure 3 – Représentation partielle de deux vedettes-matières de RAMEAU en SKOS [5]

SKOS est inspiré par des guides de bonnes pratiques comme la norme ISO 2788 pour les thésaurus [7]. Mais son objectif n'est pas de remplacer ces guides, pas plus que les données converties en SKOS n'ont vocation à supplanter les référentiels documentaires dans leur contexte d'utilisation originel. Il s'agit davantage de faciliter l'échange de données « sémantiques » – même si cette sémantique n'est pas formelle – dans le contexte du web de données et, partant, de permettre l'émergence de nouveaux usages pour ces référentiels.

Là encore, la technologie du web de données change les choses. Les mécanismes de publication simplifient l'accès aux données, et leur mise à disposition suivant un vocabulaire standard et ouvert (SKOS) facilite leur consommation, trop souvent entravée par des systèmes et formats propriétaires.

Par exemple, Europeana a commencé à enrichir ses objets de façon automatique, en les liants (entre autres) au thésaurus GEMET. Cela permet de retrouver, pour une requête comme « cheval », des objets russes qui n'ont bien sûr jamais été catalogués en utilisant un terme français : dans ce cas, ce sont les données accessibles pour le concept <http://www.eionet.europa.eu/gemet/concept/3995> qui permettent de faire le lien…

Tout comme pour les ontologies, la mise à disposition d'URIs et de données pour les ressources des KOS permet de les employer dans des descriptions externes. Il est aussi facile, pour une application particulière, de créer un nouveau référentiel en tant qu'extension d'un référentiel existant, comme dans l'exemple suivant extrait de l'introduction à SKOS [6], où <ex1:> et <ex2:> font référence aux espaces de noms de référentiels (fictifs) différents :

```
ex1:referenceAnimalScheme rdf:type skos:ConceptScheme;
  dct:title "Reference list of animals"@en.
ex1:cats rdf:type skos:Concept;
  skos:prefLabel "cats"@en;
  skos:inScheme ex1:referenceAnimalScheme.

ex2:catScheme rdf:type skos:ConceptScheme;
  dct:title "The Complete Cat Thesaurus"@en.
```

```
ex1:cats skos:inScheme ex2:catScheme.

ex2:abyssinian rdf:type skos:Concept;
  skos:prefLabel "Abyssinian Cats"@en;
  skos:broader ex1:cats;
  skos:inScheme ex2:catScheme.

ex2:siamese rdf:type skos:Concept;
  skos:prefLabel "Siamese Cats"@en;
  skos:broader ex1:cats;
  skos:inScheme ex2:catScheme.
```

D'un point de vue technique, il n'y a pas de différence entre utiliser la propriété <skos:broader> pour deux concepts d'un même thésaurus et l'utiliser pour deux concepts provenant de KOS différents. SKOS fournit néanmoins des relations dédiées à l'interconnexion de vocabulaires différents, lorsque celle-ci se fait *a posteriori* et selon des procédés qui, comme l'alignement automatique, peuvent différer de ceux employés pour la conception « traditionnelle » des KOS. En particulier, on peut représenter des liens d'équivalence conceptuelle exacte ou approximative (<skos:exactMatch> ou <skos:closeMatch>) qui serviront de passerelle entre des systèmes utilisant des vocabulaires conçus indépendamment, provenant de contextes applicatifs ou d'environnements linguistiques différents. On passe par exemple sans problème de <http://data.bnf.fr/ark:/12148/cb11931913j> (« Eau » dans RAMEAU) à <http://id.loc.gov/authorities/subjects/sh85145447> (« Water » dans LCSH) et à <http://d-nb.info/gnd/4064689-0> (« Wasser » dans la liste de vedettes-matières SWD de la Deutsche Nationalbibliothek).

Ces possibilités ont encouragé la publication de données SKOS par un nombre significatif d'institutions qui utilisent des vocabulaires d'autorité, notamment pour leurs systèmes documentaires. Outre ceux déjà mentionnés, on citera le *New York Times*, la Nasa, OCLC, l'Office des publications de l'Union européenne, l'Abes, le Ministère de la Culture, etc.

Pour en savoir plus le lecteur consultera les listes présentées sur le wiki de SKOS[30] ou bien interrogera le répertoire *The Data Hub*[31]. Cette adoption est allée de pair avec la création d'un certain nombre d'outils dédiés, tels qu'éditeurs, validateurs, répertoires[32], etc. À noter, étant donné les rapprochements possibles entre ontologies formalisées et KOS, que certains répertoires de domaines (comme metadataregistry.org) contiennent les deux types de référentiels.

Évidemment, SKOS n'est pas le seul jeu d'élément de métadonnées disponible pour représenter les vocabulaires de valeurs. MADS/RDF[33], FRAD[34], FRSAD[35] et les vocabulaires RDA[36] proposent des éléments qui peuvent compléter ceux de SKOS. Par exemple, en tant que modèle générique pour représenter des KOS, SKOS manque d'éléments dédiés à la représentation des notions liées à la coordination de vedettes-matières : subdivisions, vedettes construites, etc. Une extension comme MADS/RDF, aisément utilisable à côté des éléments « standard » de SKOS, permet de résoudre certains de ces problèmes.

3.3 Le problème de l'alignement

Avant de conclure cette section, il est important de revenir sur un problème crucial pour l'interopérabilité des référentiels au niveau de la sémantique des éléments qu'ils contiennent : l'alignement des KOS. Beaucoup de ces référentiels contiennent des informations qui se réfèrent aux mêmes concepts, personnes, lieux, etc.

S'il permet de représenter des alignements entre KOS, SKOS ne résout pas le problème de la découverte de ces liens. Les travaux, mentionnés dans la section précédente, de la communauté du web sémantique en matière d'alignement automatique d'ontologies ne se sont que peu intéressés au cas des référentiels de type thésaurus, qui sont souvent beaucoup plus volumineux que les ontologies formalisées, moins bien structurés, et dont les libellés sont d'une très grande hétérogénéité d'un référentiel à un autre. Le problème de l'alignement multilingue, par exemple, est très difficile à résoudre.

Certains projets, comme MACS [8], font donc toujours le choix d'un alignement manuel. Des organisations telles que la FAO des Nations Unies, pour la publication de son thésaurus AGROVOC, essaient de combiner

---

[30] http://www.w3.org/2001/sw/wiki/SKOS/Datasets

[31] http://ckan.net/dataset?q=format-skos

[32] http://www.w3.org/2001/sw/wiki/SKOS

[33] http:// www.loc.gov/standards/mads/rdf

[34] http://www.ifla.org/publications/functional-requirements-for-authority-data

[35] http://www.ifla.org/en/node/1297

[36] http://www.rdatoolkit.org

l'utilisation de systèmes de mise en correspondance automatique avec une évaluation manuelle intensive des résultats par des experts de domaine.

Dans une telle situation, un éditeur sur le web de données doit s'appliquer à choisir les référentiels les plus pertinents pour un alignement de KOS ou une utilisation dans le cadre d'un jeu de données. De façon assez comparable à ce qui se passe dans le cas des ontologies formalisées, l'adéquation à l'application visée est cruciale. Dans un cadre documentaire, cela pourra impliquer l'accord avec le thème de collections spécifiques ou avec un contexte institutionnel et/ou disciplinaire précis. Le nombre, la granularité et le type des entités présentes, par exemple, joueront un rôle important. La qualité et l'exhaustivité des liens sémantiques sont aussi des critère fondamentaux. Dans le cas de vocabulaires de valeurs, une attention toute particulière devra en plus être portée aux informations lexicales (libellés, définitions) attachées aux ressources, surtout dans des contextes d'application multilingues. Finalement, la question des conditions d'utilisation du vocabulaire (licence ouverte ou non) ne peut être ignorée.

Le choix dépend aussi de la topologie du réseau d'alignement – à la fois des correspondances existantes et de celles que l'on veut construire, le cas échéant. Comme il a été indiqué dans le chapitre 3, les référentiels de type KOS sont volontiers associés au modèle *hub and spoke* : ils agissent comme un pivot entre des jeux de données différents. C'est le cas du projet VIAF, qui consolide en un point central les données d'autorité sur les personnes de dizaines de bibliothèques. Il y a cependant des stratégies alternatives qui vont jusqu'à l'alignement d'un référentiel avec tous ses « voisins » (structures de « paires ») [1]. La figure 4 présente ces deux modèles structurels d'interopérabilité.

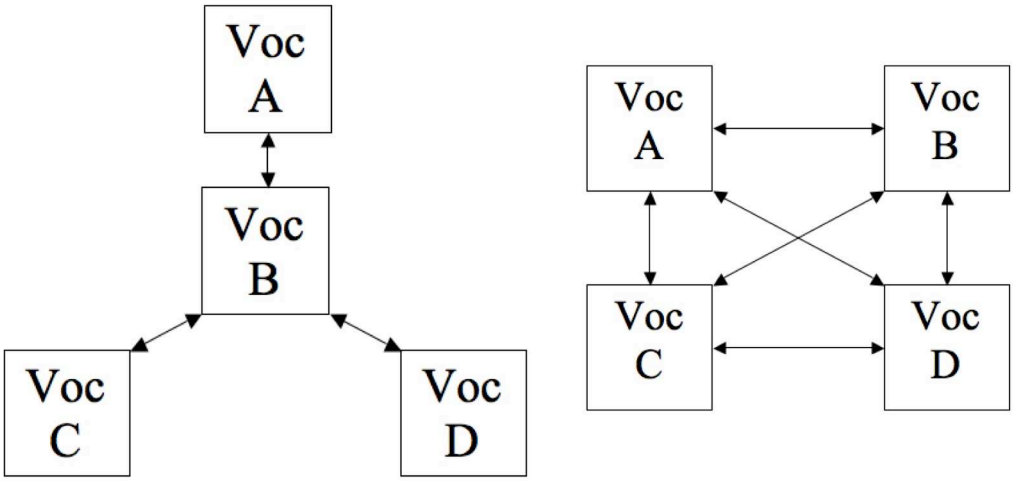

Figure 4 – Modèles structurels d'interopérabilité : pivot (à gauche) et paires (à droite) [1]

Les deux modèles ont bien sûr leurs avantages et leurs inconvénients. Le pivot est le plus économique en efforts, mais il peut déboucher sur une perte d'information, notamment lorsqu'on s'intéresse à l'échange de données utilisant des référentiels satellites et que le pivot ne couvre pas les éléments présents dans ceux-ci. Dans beaucoup de cas, surtout ceux qui couvrent plusieurs domaines, plusieurs contextes linguistiques, ou qui ont déjà un historique de mise en relation de référentiels suivant des logiques différentes, il est difficile de déterminer un pivot unique. On pourra alors se tourner, par exemple, vers une approche pivot plus « souple » avec plusieurs pivots relativement généraux et importants pour une communauté, qui servent de points d'ancrage à des vocabulaires plus précis et petits. Si ces pivots sont eux-mêmes relativement bien connectés, pour les éléments qu'ils partagent, alors le niveau d'interopérabilité peut se révéler suffisant. Un éditeur de données aura alors plus de flexibilité pour réutiliser un vocabulaire, pivot très général ou référentiel plus spécialisé.

4 Jeux de données de référence

Comme mentionné au début de ce chapitre, le dernier type de référentiel est constitué des jeux de données RDF de tous types qui, par leur contexte de production et de publication, peuvent bénéficier d'une audience importante et faire l'objet de réutilisation ou d'alignement de la part des éditeurs d'autres jeux. Certains jeux intéressants pour le domaine documentaire ont été recensés par l'incubateur *Library Linked Data to W3C* [14], et certains d'entre eux ont été rappelés en fin du chapitre précédent et en ouverture de celui-ci.

En particulier, des jeux de données bibliographiques peuvent tenir un rôle important pour le domaine documentaire. OCLC publie depuis peu des données liées sur Worldcat. Des données sur de nombreux articles scientifiques possédant un DOI sont également publiées par crossref.org. À un niveau plus local, des

bibliographies nationales telles que la British National Bibliography ou data.bnf.fr peuvent servir de point d'ancrage à de nombreux jeux de données.

Mais, une fois encore, rien ne distingue techniquement ces jeux des référentiels discutés dans la section précédente. Nous avons donc affaire à un ensemble continu de référentiels issus du domaine documentaire, une partie se concentrant sur la description des objets documentaires eux-mêmes et l'autre proposant des données sur les entités « de contexte ». Ce qui a permit à l'incubateur *Library Linked Data* de les inventorier au sein d'une même liste sur le Data Hub[37] et d'en proposer une visualisation commune [figure 5].

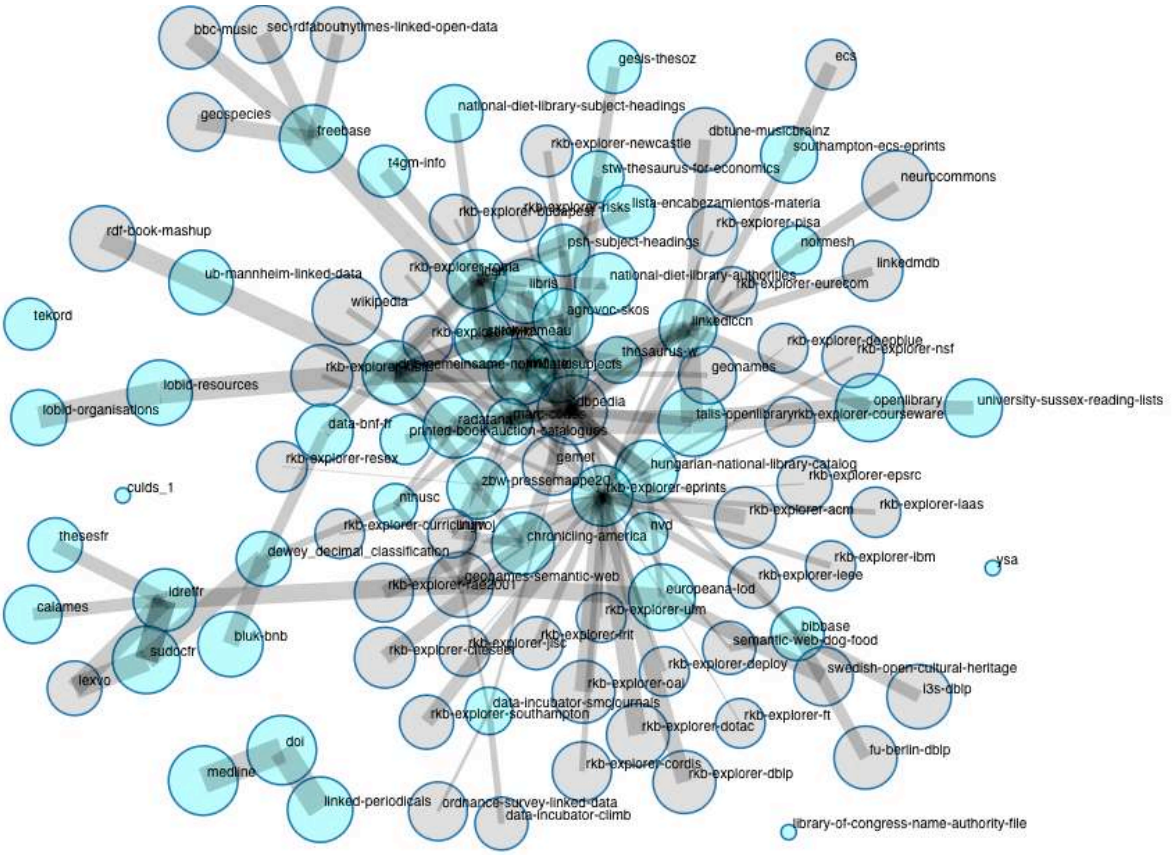

Figure 5 – Données bibliographiques sur le web de données (en bleu ciel), octobre 2011 [14]

Comme le montre ce graphe, les liens entre jeux de données ne s'arrêtent pas à ce qui est issu du monde documentaire « traditionnel » (et plus précisément, ici, de la communauté des bibliothèques). Tous les types de données mentionnés à la fin du chapitre 3 (web social, données gouvernementales, ressources géographiques et touristiques, ressources médicales et biologiques, etc.) sont susceptibles de fournir des référentiels pertinents.

Il est à noter que quelques-unes des ontologies discutées au chapitre précédent pour représenter les vocabulaires de valeurs incluent des éléments permettant de lier systèmes conceptuels et ressources décrivant directement des entités du monde « réel » – par exemple, une autorité de nom de personne et une instance de la classe <foaf:Person>. SKOS encourageait déjà une approche orientée concepts, ou les ressources dénotent des notions externes au système, en contraste avec les approches orientées termes où les chaînes de caractères jouent le premier rôle. Ceci rend plus naturel, par exemple, la gestion de systèmes multilingues ou l'appariement de concepts de référentiels différents.

Ces nouveaux éléments matérialisent un peu plus une « transition douce » entre systèmes conceptuels et bases de connaissance formalisées sur les entités du monde réel. Ce faisant, ils augurent des liens fructueux entre référentiels des systèmes documentaires traditionnels et autres systèmes à base de connaissances dont les données sont publiées sur le web. Le mouvement est déjà amorcé, puisque des jeux de données comme DBpedia et GeoNames sont déjà réutilisés dans certains projets venant d'un contexte purement documentaire.

Il faut néanmoins reconnaître les défis que pose une vision très souple de la publication et de l'alignement de jeux de données, inspirée de ce qui se passe sur le web traditionnel. La technologie du web de données permet

[37] http://thedatahub.org/dataset?groups=lld

l'expression précise de liens entre données et facilite l'accès à ces liens. La communauté des chercheurs ainsi que certains industriels fournissent des outils permettant de détecter (semi-)automatiquement de tels liens, tels que Silk [13] ou Google Refine[38]. Des catalogues centralisés de liens sémantiques déjà établis entre jeux de données, comme sameAs.org, sont également disponibles.

Mais le problème de l'interopérabilité sémantique d'un ou plusieurs jeux de données n'est pas résolu tant que les producteurs et/ou consommateurs de données n'ont pas fait l'effort d'utiliser de tels outils ou de créer par eux-mêmes les correspondances qui sont pertinentes pour les applications qui consomment les données. De fait, l'aspect régulier du *linked data cloud* (voir le chapitre 3) cache une structure plutôt éclatée ; des outils d'analyse de réseaux font en particulier apparaître un certain nombre de regroupements assez bien interconnectés, mais reliés par un très petit nombre de jeux de données pivots [figure 6].

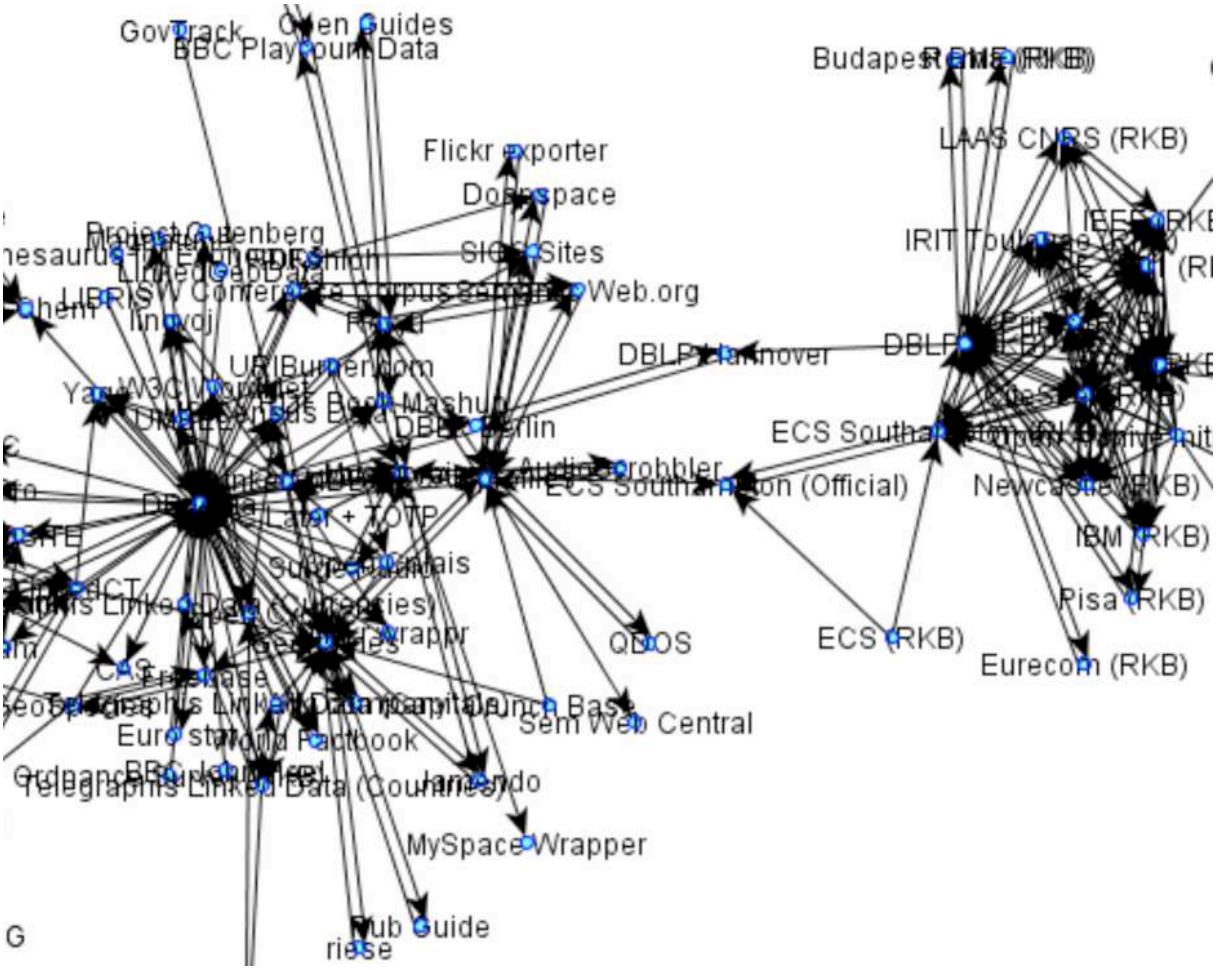

Figure 6 – Une visualisation alternative du *linked data cloud*, mars 2010 – détail [9]

Un énorme travail d'alignement de ressources (*mesh-up*) reste à faire. Comme le recommande le rapport final de l'incubateur *Library Linked Data*, les acteurs du monde documentaire (mais pas seulement eux) ont encore un rôle important à jouer en ce qui concerne la création et l'alignement de jeux de données de référence sur le web de données, à l'intérieur de leur propre domaine et en coopération avec d'autres.

Références

[1] BRITISH STANDARDS INSTITUTION. *BS 8723-4:2007 – Structured vocabularies for information retrieval. Guide. Part 4: Interoperability between vocabularies*. London : The British Standards Institution, 2007

[2] Jérôme EUZENAT, Pavel SHVAIKO. *Ontology Matching*. Berlin : Heidelberg : Springer-Verlag, 2007

[38] http://code.google.com/p/google-refine

[3] Nicola GUARINO, Chris WELTY. « Evaluating Ontological Decisions with OntoClean ». *Communications of the ACM*, 2002, vol. 45, n° 2, p. 61-65

[4] Eero HYVÖNEN, Kim VILJANEN, Jouni TUOMINEN, Katri SEPPÄLÄ. « Building a National Semantic Web Ontology and Ontology Service Infrastructure – The FinnONTO Approach ». *In* : *Proceedings of the 5th European Semantic Web Conference, ESWC 2008, Tenerife, Spain, June 1-5, 2008*. Berlin : Heidelberg : Springer-Verlag, 2008

[5] Antoine ISAAC, Thierry BOUCHET. « Rameau et SKOS ». *Arabesques*, mai-juin 2009, n° 54, p. 13-14. http://www.abes.fr/Arabesques/Arabesques-n-54

[6] Antoine ISAAC, Ed SUMMERS (eds.). *SKOS Primer*. W3C Group Note, 2009. http://www.w3.org/TR/skos-primer

[7] ISO TC 46/SC 9. *ISO 2788:1986 – Documentation – Guidelines for the establishment and development of monolingual thesauri*. Second ed. Genève : International Organization for Standardization, 1986

[8] Patrice LANDRY. « Multilingualism and subject heading languages: how the MACS project is providing multilingual subject access in Europe ». *Catalogue & Index*, 2009, issue 157. http://www.cilip.org.uk/get-involved/special-interest-groups/cataloguing-indexing/Documents/CandII157.pdf

[9] *LOD cloud shows surprisingly lumpy structure*. Larkc project blog, 2010. http://blog.larkc.eu/?p=1941

[10] Alistair MILES, Sean BECHHOFER (eds.). *SKOS Reference*. W3C Recommendation, 2009. http://www.w3.org/TR/skos-reference

[11] Mikael NILSSON, Thomas BAKER, Pete JOHNSTON. *The Singapore Framework for Dublin Core Application Profiles*. DCMI Recommended Resource, 2008. http://dublincore.org/documents/singapore-framework

[12] Barry SMITH, Chris WELTY. « Ontology: Towards a new synthesis ». *In* : Nicola Guarino (ed.). *Formal Ontology in Information Systems*. New York : ACM Press, 2001. P. 3-9

[13] Julius VOLZ, Christian BIZER, Martin GAEDKE, Georgi KOBILAROV. « Discovering and Maintaining Links on the Web of Data ». *8th International Semantic Web Conference, ISWC2009, Westfields, USA, October 25-29, 2009*. Berlin : Heidelberg : Springer-Verlag, 2009. http://dx.doi.org/10.1007/978-3-642-04930-9_41

[14] W3C LIBRARY LINKED DATA INCUBATOR GROUP. Antoine ISAAC, William WAITES, Jeff YOUNG, Marcia ZENG. *Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets*. Octobre 2011. http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset