25/3/2013

Thésaurus Covadis

Benoît DAVID – MEDDE / CGDD / DRI / MIG

Pourquoi un thésaurus ?

Le système actuel des Géobases, utilisé dans les DDT et les DRAAF est fondé sur la définition d'une arborescence utilisée à la fois comme hiérarchie de répertoires de stockage physique des fichiers dans chaque service et comme outil d'accès aux jeux de données au sein du patrimoine du service. La convergence des outils SIG au sein des ministères dans le cadre de Géo-IDE impose de prendre en compte les besoins d'autres services, et donc de faire évoluer cette arborescence, ce qui est particulièrement difficile car toute modification de l'arborescence impose une modification des stockages dans chacun des services. On voit donc qu'il est nécessaire de décorréler l'organisation physique des fichiers de l'outil de recherche des jeux de données qu'il est souhaitable de fonder sur un thésaurus, outil largement utilisé dans les sciences de l'information.

Outre ce besoin, la mise en œuvre de la directive Inspire impose d'affecter à chaque jeu de données un des thèmes des annexes de la directive. Pour aider les services à effectuer cette affectation, il est souhaitable de constituer la liste des types d'objets géographiques les plus fréquents pour définir à quel thème chacun doit être affecté.

Il n'existe pas de thésaurus couvrant l'ensemble du champ sémantique de Géo-IDE et il aurait été difficile d'utiliser dans Géo-IDE les différents thésaurus utilisés par les services concernés tels que Urbamet, Ecoplanète, Agrovoc, ...
Le choix est donc, pour Géo-IDE, de créer un nouveau thésaurus, appelé thésaurus Covadis, permettant d'indexer chaque jeu de données stocké par un ou plusieurs mots-clés de ce thésaurus. Les données stockées dans Géo-IDE étant gérées dans le cadre de la mise en œuvre des politiques publiques de l'Etat ou des Collectivités territoriales et de nombreuses données correspondent à des définitions réglementaires, afin de limiter au maximum le travail d'élaboration de ce nouveau thésaurus, il est largement constitué avec les termes juridiques correspondants aux données gérées.

Définition du thésaurus

Le thésaurus est constitué de deux sous-ensembles, appelés vocabulaires :

1. Le premier, appelé **vocabulaire des géo-concepts**, correspond aux types d'objets géographiques considérés comme bien définis. On y trouve en premier les types d'objets géographiques définis réglementairement (cad définis par un décret, une loi, un texte de l'Union européenne ou une traité international ratifié par la France). Par exemple, le terme « réserve naturelle » est défini dans la partie législative du code de l'environnement. On y trouve ensuite les objets géographiques statistiques définis par l'INSEE ou un service statistique d'un ministère, tel par exemple, le terme « petite région agricole ». On ajoute à ces deux premiers ensembles les types d'objets géographiques utilisés par les services et relevant de la directive Inspire. A chaque terme (cad mot-clé du vocabulaire) sont associés des documents de référence provenant si possible de Légifrance pour les types d'objets géographiques réglementaires.
2. Un certain nombre de jeux de données ne pourront pas être indexés par un géo-concept. Il est proposé de les indexer par un terme correspondant à une politique publique et donc de créer un deuxième **vocabulaire de politiques publiques**. Par exemple, le terme « chasse » est utilisé pour indexer diverses données utilisées pour mettre en oeuvre cette politique. L'intitulé de la politique est issu d'un texte juridique. Cette solution n'est utilisée que lorsque le premier vocabulaire ne peut pas l'être.

Organisation hiérarchique des termes des vocabulaires

Les différents termes (cad les géo-concepts et les politiques publiques) sont organisés hiérarchiquement, conformément aux pratiques des thésaurus, pour représenter les relations **partitives** et **génériques**. Un exemple de relation partitive est donné par le terme « parc national » ayant pour enfant le terme « cœur de parc national » pour exprimer que dans un parc national est défini un cœur qui est un type d'objet géographique particulier. Un exemple de relation générique est donné par le terme « servitude d'utilité publique » ayant pour enfant « réserve naturelle » exprimant le fait qu'une « réserve naturelle » est un cas particulier de « servitude d'utilité publique ». Cette organisation hiérarchique est très utile pour l'interrogation ; ainsi un jeu de données indexé avec le terme « réserve naturelle » apparaîtra dans les résultats d'une recherche sur le terme « servitude d'utilité publique ».

Les géo-concepts sont assez nombreux et il n'existe pas de structuration permettant un accès simple. La solution proposée consiste à organiser les politiques publiques hiérarchiquement et à rattacher hiérarchiquement chaque géo-concept à la ou les politiques publiques qu'il contribue à mettre en oeuvre. L'accès aux deux vocabulaires s'effectue naturellement au travers de la hiérarchie des politiques publiques fournie en annexe A.

Par convention, les politiques publiques de premier niveau sont très générales et ne doivent pas être utilisées pour indexer des ressources.

Le thésaurus accepte la **polyhiérarchie**, ce qui signifie qu'un terme peut avoir plusieurs parents. Par exemple une réserve naturelle peut à la fois avoir comme parents la politique publique « nature, paysage, biodiversité » et le géo-concept « servitude d'utilité publique ». Ainsi le jeu de données indexé avec le terme « réserve naturelle » apparaîtra dans les résultats d'une recherche sur le terme « nature, paysage, biodiversité ». En cas de polyhiérarchie, le terme apparaît à plusieurs endroits lors d'une présentation hiérarchique comme celle fournie en annexe A.

Affectation aux thèmes Inspire

Un objectif important du thésaurus est de faciliter l'affectation d'un thème Inspire à chaque jeu de données dans le cadre du catalogage. Ainsi à chaque géo-concept est affecté le thème Inspire correspondant ou le terme particulier « Hors Inspire » indiquant que le géo-concept ne relève pas d'Inspire. Cette affectation est fournie dans l'annexe A.
Ainsi, si un ADL associe un terme du thésaurus à un jeu de données lors du catalogage, il lui affecte automatiquement un thème Inspire.

Alignement avec divers thésaurus

Lors de la présentation à la COVADIS des principes de constitution du présent thésaurus, il a été demandé de prévoir un mécanisme permettant de relier les termes de ce nouveau thésaurus à des termes de thésaurus existants tels que AGROVOC de la FAO, EuroVoc de la Commission européenne, GEMET de l'Agence européenne de l'environnement, …
Pour répondre à cette demande, appelée alignement entre thésaurus, chacun des 2 vocabulaires définis précédemment est structuré conformément au standard SKOS (Système simple d'organisation des connaissances) du W3C. Le standard SKOS permet d'effectuer de tels alignements à condition que les termes des thésaurus à aligner correspondent chacun à un URI, ce qui est le cas pour les 3 thésaurus cités ci-dessus.
Pour une présentation détaillée du standard SKOS, se reporter au document d'introduction du W3C : http://www.w3.org/TR/skos-primer/

Première approche sur le périmètre de l'arborescence Covadis

Pour conforter la démarche entreprise et définir une première version du thésaurus, une correspondance entre les fiches nationales de l'arborescence Covadis et un terme du thésaurus Covadis a été établie. Cette correspondance permet de s'assurer qu'il est possible d'affecter un terme du thésaurus à chaque fichier correspondant à une fiche nationale de l'arborescence Covadis.
Afin de pouvoir utiliser efficacement cette correspondance dans la reprise des métadonnées du GéoRépertoire dans Géo-IDE Catalogue pour affecter un thème Inspire aux métadonnées, cette correspondance a été complétée par une indication pour chaque fiche nationale si le jeu de données est une copie d'un jeu existant en dehors de Géo-IDE.
Ce tableau de correspondance est fourni en annexe B.

Compatibilité ascendante du thésaurus avec l'arborescence Covadis

Grâce au tableau de correspondance défini ci-dessus, le thésaurus offre une compatibilité ascendante avec l'arborescence Covadis. Ainsi, un service disposant d'une Géobase constituée de jeux de données conformes aux fiches nationales, l'indexation avec le nouveau thésaurus de ces jeux est automatique et peut donc être utilisée en interrogation. En outre, ce service peut conserver sa Géobase structurée conformément à l'arborescence et utiliser le thésaurus uniquement pour les jeux de données ne correspondant à aucune fiche nationale dans l'arborescence.
D'un autre côté, un service disposant d'une arborescence de fichiers différente de l'arborescence Covadis peut décider de la conserver et d'indexer chaque jeu de données avec un terme du thésaurus.
Enfin, les patrimoines de ces deux services pourront être interrogés de manière uniforme en utilisant les termes du thésaurus sans avoir à connaître l'arborescence de répertoires utilisée par chacun des services.

Livrables

Les livrables fournis sont actuellement en version de travail.
Outre cette note de présentation, le thésaurus est fourni sous la forme de 3 types de livrables :

Documents annexés à la présente note

Les deux annexes à la présente note fournissent une bonne vision du contenu du thésaurus.

- L'annexe A présente hiérarchiquement les politiques publiques et les géo-concepts. Le document est structuré au travers de la hiérarchie des politiques publiques et les différents géo-concepts du thésaurus sont listés rattachés aux politiques publiques auxquelles ils sont associés. Pour chaque géo-concept, le thème Inspire correspondant est fourni.
  Ce document est aussi disponible sur http://geocat.fr/?sortie=ppgc&themeInspire=1
- L'annexe B fourni le tableau des fiches du Géorépertoire avec notamment pour chaque fiche le terme correspondant du thésaurus et le thème Inspire associé ainsi qu'une information indiquant si le jeu de données est une copie d'un jeu existant en dehors de Géo-IDE.
  Ce document est aussi disponible sur http://geocat.fr/?sortie=tabGRppgcThi&ppgc=1&themeInspire=1

Accès interactif aux informations au travers d'un site Internet

Le site de démonstration accessible à l'adresse http://geocat.fr/ permet de naviguer au sein des différents termes et d'accéder à la documentation correspondante. Une présentation est disponible à l'adresse : http://geocat.fr/presentation.html

Fichiers SKOS Turtle et RDF

Les deux vocabulaires formalisés en SKOS sont disponibles pour une réutilisation dans les applications informatiques aux adresses suivantes :

- http://geovoc.fr/covadis/vocabulaire/geo-concept/turtle pour le vocabulaire des géo-concepts en format Turtle et http://geovoc.fr/covadis/vocabulaire/geo-concept/xml en format RDF
- http://geovoc.fr/covadis/vocabulaire/politique_publique/turtle pour le vocabulaire des politiques publiques en format Turtle http://geovoc.fr/covadis/vocabulaire/geo-concept/xml en format RDF

Le standard SKOS imposant l'utilisation d'URI, dans cette démarche le choix a été fait d'utiliser les préfixes http://geovoc.fr/covadis/vocabulaire/geo-concept et http://geovoc.fr/covadis/vocabulaire/politique_publique.
La question se pose d'utiliser un nom de domaine en gouv.fr. Cela nécessiterait alors de mettre en oeuvre un projet informatique plus formalisé.

Suite de la démarche

Les différents livrables ne sont pas définitifs et doivent être considérés comme une version de travail. L'objectif à ce stade est de fournir une compréhension suffisante des principes exposés dans cette note pour qu'ils puissent être **validés** par la Covadis. Cette validation est indispensable avant que les services se les approprient et les mettent en oeuvre.

Ces principes ont été présentés aux représentants des Dreal lors des journées de décembre 2012 et aux CMSIG fin 2012 ; des contacts ont été pris avec les réseaux de documentalistes du MEDDE pour les informer de la démarche et mettre en place une coopération pour la maintenance du thésaurus.

Après validation, la démarche pourra être mise en oeuvre de manière progressive principalement dans les services qui ne disposent pas actuellement d'une Géobase et qui auront à migrer leur patrimoine vers Géo-IDE.
Pour les services qui disposent d'une Géobase, la mise en oeuvre sera plus légère grâce au mécanisme de compatibilité ascendante présenté ci-dessus.

L'organisation hiérarchique du thésaurus devrait faciliter la saisie des termes lors du catalogage et surtout étendre les possibilités d'interrogation des jeux de données à partir de termes parents ; elle ne sera cependant pas utilisable dans un premier temps dans les applications Géo-IDE Catalogue et Géosource qui ne permettent que d'exploiter des listes de termes non hiérarchisées. Le thésaurus devra donc être utilisé dans un mode simplifié dans un premier temps ce qui est suffisant pour saisir les termes lors du catalogage. Une démarche a été lancée avec le BRGM pour étendre Géosource afin qu'il exploite un thésaurus hiérarchisé tel que celui présenté dans cette note.

Conclusion

Cette note présente les raisons et la démarche mise en oeuvre pour constituer un nouveau thésaurus dans le cadre de Géo-IDE, appelé thésaurus Covadis. Ce thésaurus est nécessaire pour le déploiement de Géo-IDE dans les services qui n'utilisent pas actuellement une Géobase. Il offre une compatibilité ascendante aux services qui en utilisent une et permettra une interrogation homogène des différents patrimoines.

A ce stade, il est proposé à la COVADIS de valider les principes exposés dans cette note avant d'entamer une démarche d'accompagnement et de concertation avec les services.