thomas.francart@mondeca.com, lise.rozat@mondeca.com
pierre-yves.vandenbussche@mondeca.com, bernard.vatant@mondeca.com

Thomas Francart, Lise Rozat, Pierre-Yves Vandenbussche et Bernard Vatant sont respectivement directeur technique, consultante en intégration de données et de vocabulaires, chercheur et architecte de données senior à Mondeca, éditeur de logiciel spécialisé dans les technologies sémantiques. Ils ont chacun participé et participent toujours à des projets de publication et de sémantisation de données publiques et de nombreux autres référentiels d'entreprise.

[**réaliser**] **Un nouveau type de patrimoine immatériel fait son apparition sous la forme de représentations sémantiques d'entités de référence de la vie publique : entités géographiques et administratives, services publics, vocabulaires et nomenclatures. Cette nouvelle forme de « bien public » devrait se concrétiser par des identifiants et adresses pérennes, des éléments de description standards, réutilisables et mis à jour par les organismes de référence. Illustrations.**

La sémantisation des données publiques : quelques premiers cas très parlants

Faciliter l'ouverture et la réutilisation des données publiques (*l'open data*) est au cœur du débat actuel sur le rôle de l'action publique en faveur de l'économie numérique [1]. Les données publiques ont certes une valeur économique – elles permettent de créer de nouveaux services à valeur ajoutée pour le citoyen – mais elles constituent également des données de référence susceptibles d'être réutilisées dans de nombreux contextes. Le travail de Mondeca s'inscrit dans ce deuxième axe : au-delà de la simple mise à disposition, les données doivent être « sémantisées » pour fonctionner comme des données de référence réutilisables. Nous illustrerons ici ce travail de sémantisation à travers les expériences de trois acteurs publics.

L'Insee ou la problématique de définition d'URI

Parmi les missions de l'Institut national de la statistique et des études économiques (Insee) figurent la définition et la mise à disposition du public d'un certain nombre de nomenclatures officielles décrivant des entités de référence dans le domaine de la statistique publique[1]. En particulier, les nomenclatures attribuent aux entités des identifiants (communément appelés *codes*) dont les plus connus – et les plus utilisés – sont ceux du Code officiel géographique (COG) qui définit les découpages administratifs et statistiques du territoire. On pourrait citer aussi les nomenclatures d'activités, de produits ou de services. Si ces codes accessibles au public constituent des références partagées, leurs nomenclatures sont publiées dans des formats variés (pages HTML, PDF, tableurs)*. En outre, un code n'a de signification que dans un contexte d'utilisation explicite. Ainsi « 05065 » identifie la commune de Guillestre uniquement dans un contexte où l'on sait que c'est la valeur d'un code commune. Pour identifier cette même commune sans ambigüité sur le web des données, il faut attribuer un URI à l'entité correspondant à ce code. La méthode la plus efficace est de construire des schémas d'URI incorporant les codes, concaténés à un espace de noms spécifique du type d'objet utilisant le code. On aura donc, pour identifier la commune en question, l'URI http://data.insee.fr/geo/Commune/05065. Un tel URI est construit et publié selon les bonnes pratiques du web sémantique[2]. Le domaine insee.fr est contrôlé par l'autorité de définition des codes, ce qui en garantit la qualité et la pérennité. Le contexte /geo/Commune/ sera identique pour tous les objets du même type. Ce type lui-même est défini dans l'ontologie du COG par la classe http://data.insee.fr/geo/def/Commune.

Sur cette base, la conversion en RDF des données du COG par l'Insee peut être facilement automatisée, et les applications tierces utilisant déjà des codes Insee peuvent également convertir leurs données et les lier aux entités de référence.

L'Asip ou la question du format de publication

L'Agence des systèmes d'information partagés de santé (Asip Santé) publie la terminologie Snomed 3.5 VF (Systematized Nomenclature of Medicine). Cette version française est disponible sous forme de fichiers dans un format tableur propriétaire avec une structure spécifique en colonnes[3]. En plus

** Les sigles des technologies relatives au web sémantique sont développés en page 29.*

////// de l'hétérogénéité, l'utilisation de ce format induit des erreurs. Par exemple, la structure hiérarchique de la Snomed 3.5 est présente de manière implicite dans le fichier publié. Comme le montre la figure suivante, ceci génère une ambiguïté d'interprétation où le même « TermCode » (identifiant unique d'un concept de la terminologie) est donné à plusieurs concepts. La documentation ne précise pas comment interpréter cette situation.

LES AMBIGUÏTÉS DU FORMAT TABLEUR : LE CAS SNOMED

	A	B	C	D	E
1	LIGN	TERMCODE	FMOD	FCLASS	FNOMEN
2	2	D0-00000			Chapitre 0 Maladies de la peau et des tissus sous-cutanés
3	3	D0-00000		-	Section 0-0 Maladies de la peau et des tissus sous-cutanés: termes généraux, types histologiques et infections
4	4	D0-00000		0	0-00 Maladies de la peau et des tissus sous-cutanés: termes généraux et types histologiques
5	5	D0-00000		00	0-000 Maladies de la peau et des tissus sous-cutanés: termes généraux
6	6	D0-00000		01	maladie de la peau et du tissu sous-cutané
7	7	D0-00004		01	maladie de la peau

En cas de doublon, par exemple « D0-0000 », il faut reconstruire l'arborescence selon l'ordre d'apparition des lignes du tableur et rectifier le code selon le niveau de la hiérarchie : « D0 » pour le premier niveau, « D0-0 pour le second », etc.

Dans le cadre du projet de recherche InterSTIS[4], nous avons expérimenté le passage d'un format de représentation type tableur à un format du web sémantique, Skos [2], dédié à la représentation de terminologie. Ce travail [3] lève toute ambiguïté d'interprétation de cette terminologie grâce à la représentation formelle des entités, propriétés et relations entre les éléments de la terminologie.

Cet exemple illustre l'importance du format de publication pour éviter toute ambiguïté d'interprétation lors de la réutilisation de données publiques.

Service-public.fr ou les défis de la diffusion de données sémantisées

La Direction de l'information légale et administrative (Dila) publie l'*Annuaire de l'administration française* sur le portail Service-public.fr. Le premier défi relevé par la Dila a été de centraliser et de structurer l'annuaire sur un modèle de connaissances commun (services, fonctions, personnes). Les entrées de l'annuaire sont décrites selon un schéma cohérent assurant la qualité des données diffusées. La Dila assurera bientôt la gestion et la diffusion des informations concernant les services locaux ; les préfectures, les mairies et de nombreux partenaires seront amenés à enrichir et utiliser ces données pour alimenter leurs applications.

Ainsi, l'enjeu actuel auquel se confronte la Dila est la définition d'un identifiant unique et pérenne pour chaque ressource publiée. Comme pour l'Insee, l'utilisation d'URI dans un espace de nom propre à la Dila accentuera la crédibilité des données, facilitera la communication avec les partenaires locaux et l'association des ressources à d'autres.

Se pose ensuite la question de la publication des données de l'annuaire. Quel format adopter ? Quelles données publier pour le grand public ? Pour les partenaires ? Actuellement, l'organisme s'appuie sur un flux RDF structuré qui est transformé en page HTML pour l'annuaire en ligne ou en flux XML pour les partenaires. La maintenance et la mise à jour de ces trois flux pourraient être unifiées par le mécanisme de *négociation de contenu* sur Service-public.fr. Ce mécanisme donne une représentation différente d'un même URI en fonction du mode de consultation adopté : la description RDF serait transmise aux applications supportant les langages du web de données, la page HTML serait envoyée aux navigateurs web pour la consultation par le grand public et enfin le format XML actuellement utilisé serait fourni aux partenaires locaux.

Enfin, les questions concernant les évolutions dans le temps de l'annuaire surviennent dès lors que des organismes tiers utilisent les données de la Dila. Comment publier les évolutions de l'annuaire ? Quelle granularité adopter pour représenter ces évolutions ? La publication d'un différentiel entre la version publiée et la version n-1 est nécessaire[5] ; si aucun standard ne s'est encore imposé sur le web de données à ce sujet, la communauté du web sémantique y travaille . En effet, la mise à jour des données inter-connectées et pérennes est un défi commun à tous les fournisseurs de données ouvertes.

Ces trois études de cas montrent l'importance, pour les organismes publics, d'une prise en charge coordonnée de ce nouveau type de patrimoine immatériel que constitue la représentation sémantique des entités de référence de la vie publique : entités géographiques et administratives, services publics, vocabulaires et nomenclatures. C'est une nouvelle forme de « bien public » à mettre en place et à gérer, concrétisée par des identifiants et adresses pérennes, des éléments de description standards, réutilisables et mis à jour par les organismes de référence.

Le portail data.gouv.fr, développé par la mission Etalab[6], constitue un point d'accès unique aux données ouvertes de l'État (administration nationale et collectivités territoriales). Dans cet annuaire des données publiques, chaque jeu de données est décrit sémantiquement par un ensemble de métadonnées : format, couverture territoriale, autorité responsable, période couverte, etc. Cette initiative est la concrétisation de l'effort d'ouverture des données publiques en France et d'un premier travail de sémantisation et d'harmonisation des données publiques, qui s'appuiera entre autres sur les référentiels de l'Insee et de la Dila.

Nous avons vu que les technologies sémantiques apportent des réponses fiables à ces problématiques récurrentes : URI pour l'identification, RDF pour le format, négociation de contenu pour la diffusion. L'évolution dans le temps et le versionnement des données publiées sont des problématiques qui n'ont pas de solution définitive, mais auxquelles tente de répondre actuellement le groupe de travail sur la provenance du W3C[7]. Quoi qu'il en soit, l'*open data* ne se fera pas sans sémantique. •

1 http://insee.fr/fr/methodes
2 La publication de ces URI est en cours à la date de publication de l'article.
3 Accessible à l'adresse http://esante.gouv.fr/snomed/snomed
4 Projet Interopérabilité sémantique de terminologies de santé francophones, ANR-07-TecSan-10
5 http://www.w3.org/wiki/DatasetDynamics
6 http://data.gouv.fr/
7 WC Provenance Working Group : http://www.w3.org/2011/prov