APPROCHES DOCUMENTAIRES : PRIORITÉ AUX CONTENUS

Sylvie Dalbin, Emmanuelle Bermès, Antoine Isaac, Romain Wenz, Yann Nicolas, Tayeb Merabti, Anila Angjeli, Thomas Francart, Lise Rozat, Pierre-Yves Vandenbussche, Bernard Vatant, Yves Raimond et Dominique Cotte

A.D.B.S. | « Documentaliste-Sciences de l'Information »

2011/4 Vol. 48 | pages 42 à 59

ISSN 0012-4508

Article disponible en ligne à l'adresse :

https://www.cairn.info/revue-documentaliste-sciences-de-l-information-2011-4-page-42.htm

Approches documentaires : PRIORITÉ AUX CONTENUS

[découvrir] Le modèle du Web, conçu comme un espace public, ne peut réellement s'appliquer à l'entreprise : son espace informationnel est un espace circonscrit qu'elle seule maîtrise. Toutefois, elle bénéficie de ce qui fait le Web depuis ses débuts : les principes fondateurs (universalité, simplicité et support technique) et les technologies. Il en va de même avec le web sémantique.

Le web sémantique en entreprise : quelques cas d'usage

Sans chercher l'exhaustivité, et en prenant appui sur des applications en projet, bien avancées ou déjà déployées sur des intranets, nous présentons ici quelques contextes où les principes et techniques du web sémantique sont mis en œuvre. Plutôt que d'exposer des éléments techniques, nous avons préféré montrer le contexte, les enjeux et limites de ce déploiement du web sémantique.

Les espaces numériques de travail

Les exigences du travail collaboratif ont conduit au développement de plateformes regroupant un ensemble de fonctions permettant la réalisation de tâches courantes (co-rédaction, échanges et traitements de ressources d'origines variées – profils issus d'annuaires, courriers internes/externes, dossiers d'affaires, ressources documentaires externes), dépassant ainsi les frontières des traditionnels silos applicatifs que les organisations ont créés au fil du temps. Pour fluidifier les échanges, l'usage des microformats[1], –on peut considérer ces derniers comme du « néo-web de données » – est ici renforcé et étendu. Lorsque l'écrit constitue le cœur des activités, ces plateformes intègrent des dispositifs de rédaction structurée et/ou de référentiels communs. Ces approches techniques sont progressives et donnent un nouvel élan aux projets de *groupware* et de gestion des connaissances.

La rédaction numérique

Le travail collaboratif a transformé certaines activités traditionnelles de rédaction en basculant sur des plateformes

de production numérique intégrant des outils de rédaction collective, de gestion de production (*workflow*), de diffusion et d'archivage. Par exemple, la (co)rédaction de comptes-rendus de réunion, rapports, bulletins, encyclopédies ou documents techniques sont produits aujourd'hui à partir de CMS, de wikis[2] ou d'outils éditoriaux plus spécialisés. Plus que de simples documents au format web, il s'agit le plus souvent de corpus numériques de données structurées et reliées les unes aux autres.

Ces dispositifs mettent en œuvre des schémas ou « feuilles de style » qui permettent d'une part de baliser des morceaux de texte avec des éléments communs et standardisés, voire normalisés, d'autre part d'établir des relations internes ou externes aux corpus. Ces éléments de données, balisés et normés, constituent des nœuds identifiables de façon unique qui renvoient vers d'autres unités, elles-mêmes identifiées, créant un réseau hypertextuel exploitable.

Pourtant, si le principe d'une rédaction HTML* est acté, certaines techniques du web de données comme la représentation en RDF ne sont pas toujours mises en œuvre, limitant la réutilisation de ces infrastructures au périmètre interne de l'entreprise. Mais la présence d'identifiants uniques et pérennes, sous forme d'URI, pour de nombreux objets (titres *minima*, chapitres, paragraphes, parties de contenu comme des présentations de cas, des entités nommées utiles pour le métier, etc.) et la normalisation de la structuration des contenus par le biais de vocabulaires reconnus (profil d'application à partir du Dublin Core ou du LOMFR[3] dans le secteur éducatif par exemple) préparent le terrain pour des développements futurs.

À l'échelle individuelle et collective, le véritable défi vient des changements à apporter sur le plan des pratiques de rédaction et d'organisation des connaissances. Et l'ergonomie pas toujours optimale des outils de production, maillon faible de ces dispositifs malgré d'indéniables progrès, reste un frein au déploiement de ces projets.

Un référentiel à usages multiples

Un autre type d'applications concerne la mise à disposition de référentiels de nature terminologique utilisables par plusieurs applications ou services, mais déconnectés de ceux-ci.

Un réservoir de concepts et de termes (thésaurus ou vedettes matière, nomenclature, classification, taxonomie) est encodé selon les règles du web de données : identifiant unique et pérenne pour chacun des concepts ou chacun des regroupements de concepts (champ sémantique), structure normalisée pour l'échange de ces concepts et termes (ici Skos[2]). La volonté d'optimiser l'exploitation de données métier sur des intranets, en proposant des accès fédérés et des filtres *ad hoc* (recherche à facettes), repose sur le partage des vocabulaires utilisés (schémas et terminologies).

Ce travail de portage aux couleurs du web de données est systématiquement l'occasion d'une révision des vocabulaires utilisés dans les différents systèmes et de leur reconfiguration en tenant compte de cette nouvelle orientation de mutualisation et de partage. Concepts et termes sont affinés et documentés et la structuration est explicitée par un enrichissement des types de liens. Chacun doit pouvoir y puiser le vocabulaire dont il a besoin selon ses caractéristiques métiers et ses exigences fonctionnelles.

Le déploiement de ces dispositifs est relativement simple sur le plan technique, car les applications informatiques ne sont en général pas directement reliées entre elles. Mais il reste tributaire de la volonté des équipes métier et informatiques, car ces vocabulaires peuvent concerner un nombre conséquent d'utilisateurs et d'applications en interne.

Notons enfin que certains de ces référentiels, sont mis à disposition sur le Web, comme le thésaurus des Archives de France, Eurovoc ou le TAG du Gouvernement québécois[5].

Sylvie Dalbin est consultante en organisation et ingénierie documentaires depuis 1989 au sein d'Assistance & Techniques Documentaires. Elle conduit auprès des entreprises et organismes des actions de conseil et d'accompagnement centrées sur l'architecture des ressources et des systèmes documentaires en vue de valoriser leurs usages et d'en faciliter l'exploitation par les utilisateurs finals.
sylvieatd@aol.com - http://claimid.com/sylviedalbin

Plusieurs applications exploitant un référentiel commun

Troisième cas, plus complexe : le développement d'un référentiel sémantique commun à plusieurs applications, dans le but d'unifier les accès et de rendre possible la création de services variés tout en préservant la spécificité des processus et applications métier.

Des environnements professionnels très différents sont concernés par ce type de projet : dans le secteur patrimonial, une mise en commun des ressources entre plusieurs services jusque-là très autonomes (vidéothèque, photothèque, *records management*, musée, bibliothèque et archives) ; dans le domaine de la santé animale, un portail reliant observatoire santé, veille scientifique, données réglementaires et documentation sur les animaux ; ou encore dans des services techniques, un système de prise en charge des pannes associant base d'appels clients, documentation technique des matériels et manuel technique de diagnostic et réparation.

Les différents systèmes préexistants ont leur raison d'être. Ils s'adressent à des publics privilégiés, et proposent une organisation, un fonctionnement et des outils métier adaptés qui doivent être souvent préservés. ///////

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

1 Par exemple hCal pour les évènements, hCard pour les contacts, hAtom pour... Atom.
2 http://fr.wikipedia.org/wiki/Wiki_sémantique
3 Learning Object Metadata, profil français.
4 http://www.w3.org/2004/02/skos/
5 *Thésaurus des Archives de France*, http://www.archivesdefrance.culture.gouv.fr/gerer/classement/normes-outils/thesaurus/ ; TAG, http://www.thesaurus.gouv.qc.ca/; Eurovoc, http://eurovoc.europa.eu/drupal/?q=fr.

La démarche repose ici d'une part sur la certitude qu'une infrastructure de nature sémantique commune peut soutenir les fonctions d'accès et de navigation dans ces différents corpus ; et d'autre part sur la pertinence des techniques du web de données pour assurer efficacité, économie et performance dans le développement des applications, actuelles mais surtout futures.

Ces applications ne se contentent pas de mettre en œuvre les modèles, règles ou outils propres aux systèmes techniques, fussent-ils du web sémantique. Elles sont aussi l'occasion de s'ouvrir aux pratiques des usagers, pratiques renouvelées par l'usage du Web, au cours d'un travail important de re-conception des systèmes métier existants et peuvent conduire à des systèmes s'appuyant sur des ontologies.

Des systèmes basés sur des ontologies

Chacun des cas exposés précédemment peut s'appuyer sur des ontologies informatiques. L'ontologie, en explicitant la connaissance portée traditionnellement par l'ensemble schéma de description et thésaurus, peut être utilisée pour effectuer des raisonnements logiques. Par exemple, dans le dispositif sur la santé animale évoqué ci-dessus, une question sur les maladies des volatiles dans telle région peut être étendue, sans intervention de l'interrogateur, en récupérant automatiquement de l'observatoire des éléments sur les animaux touchés et les catégories de maladie, ou en fournissant des données de veille réglementaire ou documentaire appropriées.

Ces techniques sont particulièrement pertinentes dès lors que les ressources à exploiter sont de nature différente et qu'elles proviennent de sources variées, internes ou externes à l'entreprise : elles ne peuvent être toutes connues des utilisateurs, fussent-ils spécialistes du domaine, et le soutien d'une base de connaissances organisée autour d'une ontologie informatique est dans ce cas très efficace.

Alors, le web sémantique en entreprise ?

Les principes fondamentaux du web de données (identifiant pour toute ressource et liens entre entités) sont présents de façon assez systématique sur les intranets. Mais les techniques mêmes du web de données (URI, RDF ou SparQL), indispensables à l'économie générale ou à la performance sur le Web, y restent à ce jour moins fréquemment (URI), voire assez marginalement, mises en œuvre. Deux raisons principales peuvent être évoquées.

Tout d'abord l'entreprise n'a aucune raison de développer des projets « web sémantique ». Si elle a un projet, c'est celui d'améliorer des processus, de soutenir le travail des collaborateurs, ou de prendre en compte de nouvelles exigences dans son environnement. Le web sémantique est alors mis en concurrence avec des techniques plus traditionnelles, et de nombreuses contraintes propres aux environnements de travail et aux intranets peuvent jouer en sa défaveur.

De plus, alors que le web 2.0 laissait les utilisateurs libres de participer ou non et touchait somme toute assez peu les applications existantes, les projets web sémantique sont plus impliquants et en conséquence plus complexes à conduire. Leur nouveauté et leur impact sur l'existant supposent de constituer une équipe solide sur le long terme et de penser l'accompagnement au changement tout au long des différentes étapes, ce qui est loin d'être le cas pour des projets informatiques plus classiques.

Sans nous arrêter à une approche exclusivement technique du web sémantique, nous pouvons pour conclure faire quelques observations transversales aux différents cas énoncés :

Informatique

Dans la réalité, il est difficile d'envisager l'articulation de systèmes existants sans aucune intervention sur les systèmes eux-mêmes, tant les modèles de données métier, les systèmes techniques ainsi que les données elles-mêmes ont été conçus et produits en local, sans aucune prise en compte de l'environnement. La part des chantiers de reprise de l'existant dans l'économie du projet est de fait conséquente et impacte directement la durée et la gouvernance de ces projets. Une autre conséquence de ces nouvelles approches porte sur la place des systèmes de gestion de base de données relationnelles (SGBDR) et du langage SQL. Le repositionnement de ce dernier a démarré dans le cadre de la mise en œuvre des moteurs de recherche[6]. Le mouvement du web sémantique porte quant à lui le débat sur les bases de données elles-mêmes[7] , débat qui ne fait que commencer.

Usages

Pour parler de réutilisation dans les projets en entreprise, il est indispensable d'accorder une place plus centrale aux utilisateurs (et à leurs pratiques) et aux ressources (et à leur production en amont et leurs usages en aval) qu'aux fonctions de gestion. Autrement dit, les projets conduisent à dissocier les activités des utilisateurs des activités de gestion – ce qui inverse la représentation des systèmes d'information centrée sur le tockage et la gestion de données. Les modalités pratiques du travail en entreprise sont alors réévaluées à des niveaux de profondeur certes très variables mais qui impactent chacun d'entre nous sur des activités jugées jusque-là très personnelles : lire, écrire et coécrire, annoter, travailler ensemble, communiquer.
Au risque d'une approche exclusivement technique s'ajoute celui d'une approche de « prescription » qui porte son attention sur la structure et les normes. Il faut donc veiller à ne pas renforcer des logiques prescriptives déjà en germe dans l'intranet, de par son cadre largement structurant et normalisateur.

Coûts

Enfin, la question des coûts doit être abordée. Nous l'avons vu, la dimension de ces projets n'est pas exclusivement technique. Les aspects humains et ceux liés à l'organisation du travail sont au cœur de la plupart de ces projets dans les entreprises. De plus, le choix des techniques du web sémantique s'inscrit dans une économie de la réutilisation – c'est-à-dire de l'utilisation des résultats d'un projet pour d'autres à venir. Dans ce contexte, à qui imputer le coût d'un travail de modélisation et de qualification de métadonnées utilisables plus largement et pour des usages encore en devenir ? •

6 Citons les « accélérateurs de SGBD » (Database Offloading d'Exalead ou Database Accelerator Pertimm) lancés dans les années 2007/2008 dont l'objectif était de manipuler les données hors des SGBD.
7 Voir la notion de NoSQL (pour Not Only Sql), http://blog.neoxia.com/nosql-5-minutes-pour-comprendre

manue.fig@gmail.com

Diplômée de l'École nationale des Chartes et de l'Enssib, **Emmanuelle Bermès** est actuellement chef du service multimédia au Centre Pompidou, où elle pilote notamment le projet de Centre Pompidou Virtuel. Conservateur à la BNF entre 2003 et 2011, elle a travaillé sur Gallica, le projet de préservation numérique Spar et l'évolution des catalogues vers le web sémantique. Elle est également très active dans les réseaux internationaux (Ifla, Europeana et W3C).

[mutualiser] Sur le Web, la démarche de l'internaute n'est pas centrée sur les institutions mais sur les contenus. Pour faciliter cette transversalité, bibliothèques, archives et musées doivent dépasser leurs modèles historiques de formalisation des données… grâce aux standards du web de données.

Bibliothèques, archives et musées : l'enjeu de la convergence des données du patrimoine culturel

Le Web est avant tout un ensemble de standards qui permettent la dissémination de technologies partagées par tous, et indépendantes des environnements matériels et logiciels. Il constitue un espace global d'information, que l'on parcourt de lien en lien. Le principe de la navigation hypertexte et la généralisation de l'usage des moteurs de recherche ont profondément transformé l'accès à l'information. Sur le Web, la démarche de l'internaute n'est pas centrée sur les institutions (quelle bibliothèque est la plus à même de fournir l'information que je recherche ?) mais sur les contenus. La convergence entre institutions culturelles devient essentielle : le touriste qui prépare sa visite au musée cherche aussi bien des livres sur Picasso que les reproductions de ses œuvres ; le généalogiste qui trace l'histoire de ses arrière-grands-parents a besoin d'accéder aussi bien aux ouvrages biographiques qu'à l'état civil.

L'un des grands enjeux à l'heure actuelle réside donc dans le phénomène de convergence, sur le Web, entre les données des bibliothèques, des archives et des musées et d'interopérabilité avec d'autres données émanant de domaines métier différents, et notamment d'acteurs commerciaux. Or, les approches traditionnelles de l'interopérabilité ont montré leurs limites : la constitution de portails fournissant un accès fédéré à plusieurs bases de données présente le double inconvénient d'offrir une expérience de recherche appauvrie aux usagers, et de les obliger à se connecter sur le portail pour faire une recherche, ce qui suppose donc de connaître *a priori* son existence. De plus, le fonctionnement de ces portails repose sur l'utilisation de protocoles spécifiques comme le Z39.50 pour les bibliothèques ou, plus récemment et plus largement répandu, l'OAI-PMH[1]. Or, l'expérience a montré le peu d'intérêt des acteurs du Web, comme les moteurs de recherche, pour ces technologies trop complexes et trop spécifiques pour justifier l'investissement que représenterait pour eux leur implémentation.

Faciliter l'accès

Le web de données, en proposant une forme d'interopérabilité basée sur des standards du Web et sur des liens entre les ressources, semble à même de faciliter l'accès à des données structurées, stockées dans des bases telles que les catalogues de bibliothèques, les inventaires d'archives ou les bases culturelles des musées. Sur le Web, un utilisateur a la possibilité de naviguer d'un site à un autre sans avoir connaissance des moyens techniques utilisés pour publier les données, ni même avoir conscience des ruptures ou des frontières entre chacun des sites. De la même manière, sur le web de données, la navigation de lien en lien doit pouvoir se faire, d'un ensemble de données à un autre, sans nécessité de percevoir les limites des différentes bases de données ni leur format. L'objectif est de développer ///////

1 Open Archive Initiative Protocol for Metadata Harvesting, http://www.openarchives.org/OAI/openarchivesprotocol.html

////// une meilleure convergence entre données culturelles que celle qu'on a su élaborer avec les portails, et de les intégrer à l'écosystème du Web.

En effet, la première difficulté à laquelle on se heurte dans la quête de la convergence, c'est la diversité des modèles de données qui coexistent dans l'espace culturel.

Le modèle actuel des bibliothèques repose sur le format Marc[2], principalement conçu à l'origine pour favoriser les échanges de notices bibliographiques et éviter le catalogage multiple d'objets identiques. Les catalogues utilisant ce format sont structurés autour des concepts de notice bibliographique (pour décrire le document) et de notice d'autorité (pour décrire les personnes et les sujets). Le modèle FRBR[3] , validé depuis 1998, propose une modélisation plus souple, mais plus complexe.

Le modèle des archives met en avant la notion de contexte et de hiérarchie. Le format EAD[4], qui s'appuie sur le modèle de description de l'Isad(G)[5], permet de représenter les inventaires sous la forme d'une arborescence de composants qui favorise le respect des fonds.

Enfin, l'information des musées est déterminée par le fait qu'elle porte essentiellement sur des objets uniques, décrits en fonction des événements auxquels ils sont confrontés, de leur création à leur conservation en passant par les différentes opérations de restauration et d'exposition qui ont pu les affecter. Ce concept d'événement devient central dans le modèle, et c'est à travers lui que l'on relie les œuvres aux personnes, aux lieux, aux organisations. Ainsi, le modèle CRM du Cidoc (norme Iso 21127)[6] accorde une place structurante à l'événement.

Créer des liens

Ces différences de modèle au sein même des métiers du patrimoine culturel font de la convergence des données un véritable challenge. L'accès à ces données *via* des portails a généralement pour conséquence de réduire des données de bibliothèques, d'archives et de musées à un modèle commun, et donc de renoncer aux particularités de traitement et de conception de chacun de ces domaines. Sur le web de données au contraire, il est possible de créer des liens entre des ressources décrites suivant divers modèles, à partir du moment où la grammaire de base, commune à tous ces modèles, est le RDF*. L'emploi de divers vocabulaires ou ontologies permet ensuite d'exprimer la spécificité de chaque domaine. Les Dublin Core metadata terms[7] sont un bon exemple d'un tel vocabulaire : ils forment un cadre de référence pour exprimer les principales notions documentaires et peuvent être exprimés en RDF. D'autres vocabulaires, non spécifiques au domaine culturel, jouent un rôle similaire : Foaf[8] pour la description des personnes, Skos[9] pour la description de thésaurus... En outre se développent actuellement des vocabulaires spécifiques aux bibliothèques, archives et musées : l'Ifla[10], organisme chargé de la normalisation du catalogage, propose désormais une version RDFS du modèle FRBR et de certains vocabulaires développés traditionnellement par cet organisme. Le projet Locah[11], dont l'objectif est la publication de données archivistiques dans le web de données, a travaillé sur la création d'une version RDF de l'EAD. Il existe également une version en RDFS du CRM-Cidoc.

En complément, les référentiels sont appelés à jouer un rôle vital dans le web de données, en particulier lorsqu'il s'agit de construire l'interopérabilité entre des données issues de domaines différents. On parle de vocabulaire de valeurs ou

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

Zoom sur le Centre Pompidou Virtuel

Dans le cadre de sa stratégie numérique développée depuis 2007, le Centre Pompidou a créé une nouvelle plateforme de diffusion de contenus numériques culturels sur Internet : le Centre Pompidou Virtuel. Ce nouveau site renouvelle la stratégie de présence d'une grande institution culturelle sur le Web en partant d'une approche orientée vers les contenus. Il s'agit d'un facteur de différenciation fort, la plupart des institutions affichant une stratégie qui reste institutionnelle, et cible en priorité les visiteurs potentiels.

L'alimentation en ressources numériques repose sur la mise à disposition de l'ensemble des contenus produits à destination du public. Ces ressources renvoient au patrimoine du Centre (sa collection, son bâtiment), à sa programmation (notamment celle des conférences), et à sa production (notamment éditoriale et multimédia). Ils incluent aussi les bases de données de ses établissements associés, la Bpi et l'Ircam. Le parti-pris du site est l'absence d'éditorialisation, c'est à dire qu'il agrège des contenus existants, mais aucun n'est créé spécifiquement.

La mise en place du Centre Pompidou Virtuel a imposé un changement de paradigme, d'une numérisation de conservation vers une numérisation de diffusion. Le projet impliquait de réorganiser l'ensemble des processus de production, d'organisation, et de diffusion des contenus numériques afin de construire un centre de ressources de référence dans le domaine de la création moderne et contemporaine, et plus largement, du mouvement des idées contemporaines. Chaque contenu doit être identifié, indexé, traduit, rendu libre de droits, interopérable, et archivé de manière pérenne.

Pour cela, le Centre Pompidou virtuel innove en adoptant un modèle de données basé sur le Web sémantique. Les ressources ne sont pas organisées suivant une hiérarchie rigide, mais permettent à l'internaute de naviguer par le sens. Les contenus sont ainsi naturellement décloisonnés et traités de manière homogène, afin de permettre leur organisation en fonction des besoins de chaque utilisateur, et non en fonction d'une logique dictée par des usages ou des structures définis *a priori*. •

référentiel de valeurs pour désigner un ensemble de termes organisés en système de connaissance. On peut citer par exemple les LCSH (Library of Congress Subject Headings) ou encore le référentiel des codes de langues Iso 639-2, tous deux publiés en RDF sur le site maintenu par la Bibliothèque du Congrès[12]. Ces référentiels agissent comme une colonne vertébrale permettant de créer un point de contact entre des jeux de données différents. Dans le web de données, ce point de contact est suffisant pour naviguer sans contrainte d'un jeu à l'autre : le fait de parcourir ces liens permet alors de découvrir de nouvelles ressources de façon intuitive.

Pour les bibliothèques, le modèle des notices bibliographiques et d'autorité fonctionne déjà d'une manière similaire. Des concepts comme les personnes, les lieux, etc. peuvent être mutualisés avec d'autres domaines et ainsi contribuer à la convergence. C'est l'un des objectifs du projet Viaf[13], le fichier d'autorité international virtuel, qui réunit les fichiers d'autorités des personnes et collectivités d'une vingtaine de grandes bibliothèques nationales, pour les « ré-exposer » dans le web de données.

L'équivalent des notices d'autorité se développe aujourd'hui dans les archives avec l'EAC-CPF[14]. Dans les musées, il existe des référentiel de valeurs de type thésaurus et classifications (par exemple les différents thésaurus du Getty[15] pour les sujets, les lieux, les artistes, etc., ou encore le système de classification iconographique IconClass) qui permettent de rendre tangible le contenu des objets graphiques.

L'alignement des référentiels est un moyen de créer des passerelles entre les domaines. Par exemple, les Archives nationales de France utilisent un thésaurus généraliste nommé *Thésaurus W*, désormais publié dans le web de données[16] et relié à Rameau, le vocabulaire des vedettes matières de la Bibliothèque nationale de France. On peut ainsi relier entre elles une ressource des archives et une ressource d'une bibliothèque en utilisant ces deux thésaurus et leurs liens, même si elles sont décrites suivant des modèles différents.

Les questions qui demeurent

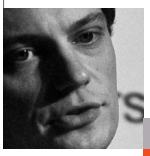
Créé en mai 2010 pour un an, le groupe de travail Library Linked Data du W3C (dit « LLD XG »)[17] avait pour objectif de faire le point sur ces premiers développements et d'identifier des pistes de travail pour faciliter l'adoption, par les bibliothèques et les autres acteurs du domaine éducatif et culturel, des standards du web de données dans les années à venir. Après avoir réuni plus de cinquante cas d'utilisation, le groupe a élaboré un recensement des vocabulaires pertinents et des données disponibles.

Dans son rapport final, le LLD XG identifie les bénéfices de l'adoption du web de données en bibliothèque, mais aussi les obstacles et les actions à entreprendre pour les surmonter. La première barrière porte sur la conversion des données pour permettre leur publication dans le web de données.

L'adoption du modèle RDF questionne les formats et modèles existants : pour les bibliothèques, par exemple, le modèle du web sémantique basé sur les liens suggère naturellement la compatibilité avec le modèle FRBR. Or, les données existantes, stockées en masses énormes dans les catalogues actuels, ne sont pas pleinement compatibles avec ce nouveau modèle. Le même type de question va se poser pour les données des archives, avec le modèle de description hiérarchique de l'EAD qu'il faut faire évoluer vers un modèle de graphe. Les standards propres au domaine culturel doivent maintenant se développer en synergie avec ceux du web sémantique, afin de garantir une meilleure interopérabilité.

D'autres questions se posent : la création et la maintenance des URI, la nécessaire montée en compétence des professionnels de la culture et des informaticiens, le coût de ces nouveaux services, les droits de réutilisation des données.., etc. Mais malgré ces obstacles, l'évolution des institutions culturelles vers le web de données a déjà commencé. Des bibliothèques, des musées, des archives publient leurs données sous cette nouvelle forme, qu'il s'agisse de référentiels ou de données d'autorité à la Bibliothèque nationale allemande[18], ou de la description des documents eux-mêmes dans le cas de la Bibliographie nationale de la British Library[19]. Pour l'instant, ces initiatives ne remettent pas en cause le modèle interne de leur système d'information, mais on commence à voir émerger des exemples d'institutions qui utilisent le RDF pour faciliter l'interopérabilité interne de leurs systèmes : c'est le cas, par exemple, du Centre Pompidou. Pour démontrer le réel potentiel de ces technologies, c'est le développement d'applications innovantes pour l'usager qu'il faudra, à terme, viser : sur ce terrain il reste encore du chemin à parcourir. •

2 MAchine-Readable Cataloging
3 Functional Requirements for Bibliographic Records. Version française accessible sur le site de la BNF, http://www.bnf.fr/fr/professionnels/modelisation_ontologies/a.modele_FRBR.html
4 Encoded Archival Description. Voir sur le site des Archives de France, http://www.archivesdefrance.culture.gouv.fr/gerer/classement/normes-outils/ead/
5 Isad(G): Norme générale et internationale de description archivistique. Voir sur le site du Conseil International des Archives : http://www.ica.org/7103/ressources-publiques/isadg-norme-gnrale-et-internationale-de-description-archivistique-deuxime-dition.html
6 Conceptual Reference Model. Voir http://www.cidoc-crm.org/
7 http://dublincore.org/documents/dcmi-terms/
8 Foaf : Friend Of A Friend. Voir http://www.foaf-project.org/
9 Skos : Simple Knowledge Organisation System. Il s'agit d'un standard du W3C : http://www.w3.org/TR/skos-reference/
10 Ifla : International Federation of Library Associations and Institutions. Voir http://www.ifla.org/
11 Voir http://blogs.ukoln.ac.uk/locah/
12 http://id.loc.gov
13 Viaf : Virtual International Authority File. Voir http://viaf.org/
14 Contexte archivistique encodé — Collectivités, personnes, familles. Voir : http://www.archivesdefrance.culture.gouv.fr/gerer/classement/normes-outils/eac/
15 http://www.getty.edu/research/tools/vocabularies/index.html
16 http://www.archivesdefrance.culture.gouv.fr/thesaurus/
17 L'information sur le groupe et le résultat de ses travaux sont accessibles sur http://www.w3.org/2005/Incubator/lld/
18 Voir http://www.d-nb.de/eng/hilfe/service/linked_data_service.htm
19 http://www.bl.uk/bibliographic/datafree.html



aisaac@few.vu.nl, http://www.few.vu.nl/~aisaac

Antoine Isaac est coordinateur scientifique d'Europeana et chercheur à la Vrije Universiteit Amsterdam. Depuis son doctorat à la Sorbonne et à l'Institut national de l'audiovisuel, il travaille sur les technologies du web sémantique pour la représentation et l'interopérabilité des collections patrimoniales et de leurs référentiels. Il a travaillé sur Skos pour le groupe W3C Semantic Web Deployment et a co-dirigé l'incubateur W3C Library Linked Data.

[réutiliser] La question de l'interopérabilité se pose entre ontologies et thésaurus. Les techniques du web sémantique lui apportent des réponses spécifiques, notamment à travers les méthodes d'alignement et le modèle de représentation Skos.

Entre thésaurus et ontologies : une affaire d'interopérabilité et d'alignement

L'initiative du web sémantique s'articule autour de la publication et de la connexion de données, quelles que soient l'organisation productrice et les techniques utilisées. La promesse d'un gigantesque graphe de *connaissances* à l'échelle du Web motive beaucoup les efforts entrepris. De fait, les propositions techniques fondamentales sur lesquelles repose le web sémantique (RDF et URI)* rendent cette vision réalisable. Néanmoins, ces mécanismes ne résolvent pas tous les problèmes qui se posent lors de la publication, l'échange et l'exploitation des données. En particulier, le problème de l'interprétation, qu'elle soit mécanique (inférence, contrôle de cohérence) ou humaine (*via* les interfaces réalisées en exploitant ces données), nécessite la création d'artefacts permettant de conférer aux données le statut de réelles connaissances. Dans le contexte du web sémantique, ce rôle est dévolu principalement aux vocabulaires formalisés connus sous le nom d'ontologies. Là encore, les ontologies seules ne constituent pas en elles-mêmes la panacée.

Ontologies et interopérabilité

La question de l'interopérabilité se pose pour les ontologies au niveau sémantique : toute organisation constitue un contexte qui peut légitimer la conception d'une ontologie pour les données qu'elle manipule. Il en résulte une prolifération d'ontologies, contenant souvent des classes et propriétés similaires. Tant que leurs éléments ne sont pas connectés explicitement au niveau sémantique (par exemple, par un lien d'équivalence), les systèmes d'information ne peuvent pas avoir un accès unifié aux données dont ces ontologies régissent l'expression. Or, si Owl permet la représentation des liens formels entre éléments d'ontologies différentes, il ne permet pas leur découverte.

Des efforts peuvent être entrepris au niveau organisationnel pour promouvoir la réutilisation d'ontologies existantes. Cependant, une ontologie pertinente pour une application pourra ne pas convenir entièrement pour une autre. Il est dans ce cas possible de créer des *extensions* – des classes et propriétés d'une ontologie Owl peuvent être spécialisées pour répondre à des besoins applicatifs précis. Mais de nouveaux problèmes peuvent se poser : l'engagement sémantique propre à l'ontologie ainsi étendue complexifie les tâches de modélisation et peut conduire un moteur de raisonnement à effectuer des inférences moins pertinentes. De fait, le web sémantique n'a jamais cherché à restreindre la liberté des créateurs d'ontologies et, malgré des progrès réels, les catalogues d'ontologies orientés réutilisation[1] sont encore rares et sous-utilisés.

L'alignement d'ontologies paraît donc la seule solution dans bien des situations où des données doivent être fédérées. Des outils issus de la recherche existent déjà pour automatiser l'alignement par comparaison des libellés des éléments ou de la structure de l'ontologie[2]. Dans bien des cas, un alignement manuel sera tout aussi efficace, surtout pour des petites ontologies. L'alignement n'est cependant pas une science exacte : des approximations sont parfois nécessaires et la prise en compte des besoins applicatifs est cruciale pour juger de la validité d'un alignement.

Ontologies *versus* Kos

Le second problème des ontologies formalisées (c'est-à-dire représentées à l'aide des langages RDFS et Owl) est qu'elles ne capturent pas l'intégralité des référentiels utilisés en documentation (thésaurus, classifications, taxonomies, dénommés ici Kos - Knowledge Organization Systems).

Les thésaurus et autres Kos fournissent des référentiels contrôlés de termes, concepts ou classes pertinents pour un domaine. Ces éléments sont organisés par des relations terminologiques (synonymie) ou sémantiques (association, généralisation) pouvant être exploitées par un opérateur humain ou un système d'information approprié. Parce qu'ils sont souvent munis de relations sémantiques telles que la généralisation (« un concept A est plus général qu'un concept B »), les Kos sont souvent

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

comparés aux ontologies formalisées. Et un certain nombre de projets se sont attelés à la conversion de thésaurus en ressources Owl, par exemple. Cependant, ces deux types d'artefacts ont des natures et des fonctions bien différentes.

Tout d'abord, il y a peu de formalisation dans les Kos : les relations présentes dans un thésaurus font sens de manière locale et sont principalement dédiées à des utilisateurs humains ou à une exploitation dans des systèmes documentaires dont les fonctions d'« inférence » sont sans rapport avec les systèmes de raisonnement formels du web sémantique. Une relation « partie-tout » (par exemple, entre France et Europe) pourra être diversement représentée soit par un lien hiérarchique de généralisation, soit par une relation « instance-classe » (entre DocSI et Publications) ou encore sous-classe-classe (entre Chats et Mammifères). Cette incertitude rend tout travail d'« ontologisation » des Kos extrêmement difficile.

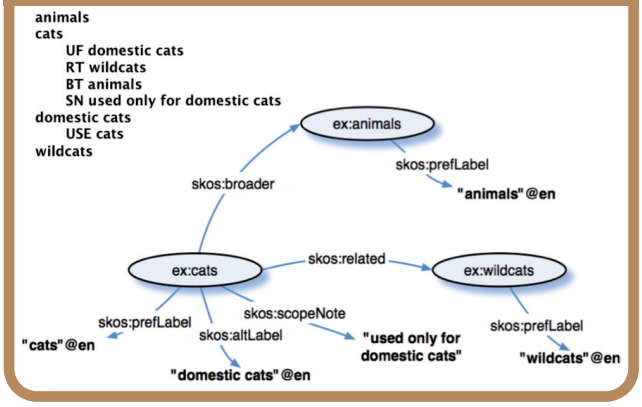
Ensuite, les fonctions sont différentes : les ontologies régentent la structure même des données. Ainsi Dublin Core pourra-t-il être utilisé comme vocabulaire organisant des données RDF créées à partir d'une notice bibliographique : un livre aura un créateur, un sujet, un lieu de publication, etc. Un thésaurus fournit, lui, un ensemble contrôlé de valeurs pour remplir ces champs[3].

Si la formalisation complète et détaillée des référentiels existants en base de connaissances de type ontologie constitue un idéal du point de vue du web sémantique, il est néanmoins toujours pertinent de chercher à porter directement les Kos existants dans l'environnement du web sémantique : ils contiennent souvent des données sur des centaines, voire des milliers, de concepts et peuvent permettre aux institutions qui possèdent ces référentiels d'expérimenter à moindre coût la technologie prometteuse du web sémantique.

Skos est une ontologie qui répond à ce besoin[4]. Son modèle, qui se veut à la fois simple et compatible avec une majorité d'approches existantes (thésaurus, classifications, etc.) permet la représentation de concepts et vocabulaires (ConceptScheme), de données terminologiques attachées à un concept (libellé préféré ou alternatif), potentiellement multilingue, de liens sémantiques entre concepts (relations générique ou associative) et de notes (d'application, définitions).

Le graphique ci-dessous présente un exemple simple où les éléments d'un thésaurus fictif (partie gauche) sont représentés en RDF en utilisant Skos.

UN GRAPHE SKOS



L'objectif de Skos n'est pas de remplacer les référentiels documentaires dans leur contexte d'utilisation initial, mais de faciliter la publication, l'échange et l'interconnexion de ces vocabulaires dans le contexte nouveau du web sémantique. L'un des principaux scénarios d'utilisation de Skos est la « mise en réseau » des référentiels conceptuels : Skos permet en effet l'assertion de relations sémantiques (en particulier des liens d'équivalence) entre concepts provenant de vocabulaires différents aussi facilement que s'ils provenaient d'un même référentiel. En particulier, on peut représenter des liens (ou y accéder) entre deux vocabulaires créés indépendamment, ou bien créer un nouveau référentiel en tant qu'extension d'un référentiel préexistant.

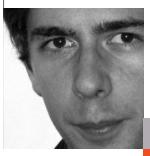
Ces possibilités ont encouragé l'adoption de Skos par un nombre significatif de producteurs de référentiels : Library of Congress, Deutsche Nationalbibliothek, *New York Times*, Nasa, DBpedia, OCLC, Office des publications de l'Union européenne, BNF, Abes, ministère de la Culture…[5]

L'application de Skos au thésaurus Agrovoc[6], édité par l'Organisation des Nations Unies pour la nourriture et l'agriculture (FAO), en est une illustration. En plus de publier Agrovoc en tant que *Linked Data*, la FAO a en effet connecté son thésaurus aux concepts de six autres référentiels publiés en Skos : LCSH, Rameau, Eurovoc, NALT, STW – thesaurus for economics et Gemet.

S'il permet de représenter des « alignements sémantiques » entre référentiels, Skos ne résout pas le problème de la découverte de ces liens, pas plus que Owl ne le permet au niveau des ontologies formelles. De façon générale, les travaux 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. Ceux-ci sont difficiles à manipuler par des outils dédiés en priorité à la mise en relation d'ontologies formalisées plus petites et sémantiquement plus riches.

Certains projets, tel Macs[7], font donc toujours le choix d'un alignement entièrement manuel, gage de précision. D'autres, tels qu'Agrovoc, essaient de combiner l'utilisation de programmes de détection automatique de correspondances à une phase d'évaluation manuelle des résultats, qui requiert toujours l'implication d'experts du domaine concerné. •

1 Par exemple, http://metadataregistry.org
2 http://ontologymatching.org
3 Les travaux récents de l'Incubator Group du W3C « Library Linked Data » ont mis en relief cette complémentarité entre « *metadata element sets* » et « *value vocabularies* ».
4 http://www.w3.org/2004/02/skos. Pour une introduction : http://www.w3.org/TR/skos-primer
5 Voir le wiki du W3C dédié à Skos : http://www.w3.org/2001/sw/wiki/SKOS/Datasets
6 http://www.fao.org/agrovoc/
7 Multilingual Access to Subject Headings, http://macs.cenl.org



romain.wenz@bnf.fr

Conservateur des bibliothèques, archiviste-paléographe, Romain Wenz a rejoint en 2009 le département de l'Information bibliographique et numérique de la BNF comme expert métadonnées. Il travaille à l'élaboration de l'outil « data.bnf.fr » qui vise notamment à fournir des pages synthétiques sur les auteurs et les œuvres.

[ouvrir] Pour les bibliothèques, l'ouverture au web de données est l'opportunité de passer outre la diversité des formats et des outils en usage actuellement autour de leurs catalogues. Pour certaines, la démarche est déjà lancée, comme en témoigne le cas de la Bibliothèque nationale de France.

Data.bnf.fr : au-delà des silos

La Bibliothèque nationale de France expose sur le Web des données structurées[1], c'est-à-dire utilisables dans des programmes informatiques. Le projet data.bnf.fr[2] regroupe des données issues des catalogues BNF[3] et de la bibliothèque numérique Gallica[4].

Identifier, exposer et relier

Qu'il s'agisse de références de catalogues ou de documents numériques, les ressources doivent disposer d'identifiants[5] (ARK[6]) pour pouvoir être citées. Il s'agit ensuite de rendre accessible l'ensemble des données. Des entrepôts ouverts (OAI-PMH)[7] existent déjà pour le Catalogue et pour la bibliothèque numérique[8]. Pour autant, ces entrepôts restent des silos séparés, sans passerelle entre eux, et la BNF souhaite aller plus loin en exposant dans data.bnf.fr des données plus complètes et reliées entre elles dans des formats structurés. Dans la continuité du projet européen TELplus[9] en 2008, avec la traduction de Rameau en Skos, les standards du web sémantique ont été adoptés.

Ainsi, le projet data.bnf.fr s'efforce de créer des liens entre les différents catalogues et la bibliothèque numérique, et d'utiliser ces données de façon nouvelle.

Pages ou données brutes

Mis en ligne en juillet 2011, data.bnf.fr présente deux volets : la publication de pages web simples pour le grand public, et l'ouverture des données en RDF* pour les programmes informatiques. Les premières réunissent les différentes informations existantes sur les auteurs et les œuvres, en allant les chercher dans les différents catalogues. Le grand public y trouve des informations rassemblées : éditions, manuscrits et numérisations d'un livre. Pourtant, le principe de fonctionnement est différent de celui d'un catalogue, puisque l'alignement sémantique permet de fournir des pages HTML simples.

Les données brutes de data.bnf.fr sont accessibles selon le standard RDF. On peut les récupérer avec un programme d'extraction[10] ou par téléchargement complet[11], avec déjà plusieurs centaines de milliers de ressources signalées. Ces données sont juridiquement lacées sous la « licence ouverte » de l'État[12], ce qui répond à la principale demande des ré-utilisateurs lors du lancement.

En outre, ces pages intègrent des données embarquées de nature descriptive (issues d'autres réservoirs, structurés au format RDFa ou schema.org) et fournissent des liens directs vers les documents numériques, pour que les moteurs puissent facilement les identifier.

Ce projet évolue avec souplesse et croît en se liant aux autres bibliothèques[13], et plus largement au secteur culturel – bibliothèques[14], archives[15] et musées[16] –, qui s'ouvre peu à peu au web sémantique, avec l'échange de données et la création de liens. Par la suite, l'emploi de formats structurés et d'identifiants devrait fédérer des ressources de bases multiples au type de contenu différent, comme les œuvres musicales par exemple. •

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

1 http://www.bnf.fr/fr/professionnels/web_semantique_donnees/s.web_semantique_intro.html
2 http://data.bnf.fr
3 Catalogue général http://catalogue.bnf.fr et base BnF Archives et Manuscrits http://archivesetmanuscrits.bnf.fr
4 http://gallica.bnf.fr
5 Emmanuelle Bermès. « Des identifiants pérennes pour les ressources numériques », 2006, http://bibnum.bnf.fr/identifiants/identifiants-200605.pdf
6 Archival Resource Key : https://confluence.ucop.edu/display/Curation/ARK
7 DublinCore, selon le protocole OAI-PMH, http://bibnum.bnf.fr/oai/
8 Elles sont utilisées notamment par The European Library et Europeana.
9 http://www.theeuropeanlibrary.org/telplus
10 Par « négociation de contenus ». Voir http://data.bnf.fr/semanticweb
11 Accès aux données décrit dans http://data.bnf.fr/semanticweb
12 http://data.bnf.fr/licence
13 Aussi bien la bibliothèque du Congrès (http://id.loc.gov) que le Sudoc (http://www.idref.fr/autorites/autorites.html).
14 Exemple du « Linked data summit » de Stanford, http://www.clir.org/pubs/reports/pub152/LinkedDataWorkshop.pdf
15 Exemple du Thésaurus W : http://www.archivesdefrance.culture.gouv.fr/gerer/classement/normes-outils/thesaurus/
16 Exemple : Linked Open Data in Libraries Archives and Museums Summit, http://lod-lam.net/summit/

nicolas@abes.fr

Depuis 2008, Yann Nicolas est responsable du département Études & Projets à l'Agence bibliographique de l'enseignement supérieur (Abes) où il supervise différents projets comme theses.fr, la version 2 de Star (gestion des thèses numériques), Step (gestion des thèses en cours), IdRef (application web et web service du Sudoc pour interroger, consulter et créer des données d'autorité), Self (services à la demande de l'Abes), etc. De 2006 à 2010, il a été chef de projet Calames (Catalogue en ligne des archives et manuscrits de l'enseignement supérieur).

[**exposer**] Depuis 2008, l'Abes est engagée dans un vaste mouvement d'exposition de ses données. Ses différentes applications adoptent progressivement les technologies du web de données afin de permettre à d'autres métiers d'exploiter ses données dans de nouveaux contextes.

L'Abes engage ses forces et ses données

L'Agence bibliographique de l'enseignement supérieur (Abes) a trempé un orteil dans le web de données dès 2008, quand Calames (CAtalogue en ligne des archives et des manuscrits de l'enseignement supérieur) s'est mis à exposer ses métadonnées de manuscrits et d'archives en RDFa*. Depuis, le corps entier des données Abes y est passé. En octobre 2010, les autorités Sudoc (Système universitaire de documentation) ont été exposées en *linked data*, à travers la nouvelle application dédiée IdRef. En juillet 2011, ce fut le tour des notices bibliographiques du même Sudoc. Quelques jours après ouvrait theses.fr : côté face, un moteur de recherche et côté pile, des métadonnées brutes en RDF. Cet effort systématique est aujourd'hui un exercice imposé pour tout grand catalogue, avec deux buts à l'esprit : contribuer à la constitution d'un réservoir bibliographique global et libre ; exposer les données bibliographiques de façon à ce qu'elles soient compréhensibles et réutilisables dans d'autres contextes.

Principes directeurs

Des URL simples. Si une notice Sudoc a pour URL http://www.sudoc.fr/154673293, il suffit d'ajouter l'extension .rdf pour obtenir les métadonnées RDF : http://www.sudoc.fr/154673293.rdf.

Des données liées… entre elles. Depuis 2006-2007, l'Abes a développé de nouveaux catalogues, mais en veillant à les arrimer au pivot des autorités Sudoc, renommées en 2010 « IdRef ». En effet, les données de Calames et de theses.fr sont connectées à celles d'IdRef et, par ce biais, désormais aux données du Sudoc. Comme s'il y avait un « web de données liées Abes », intégré comme un seul homme au *web of data*, comme on peut le constater sur le Lod Cloud[1] (nuage de données liées).

Des données Sudoc liées … à d'autres. Les données Abes ne vivent pas en autarcie. Celles du Sudoc, notamment, sont liées à d'autres corpus de métadonnées RDF comme Dewey, MeSH, Lexvo ou Geonames. Mais c'est trop peu. Nous voulons continuer.

Par ailleurs, à mesure qu'IdRef deviendra un référentiel de plus en plus utilisé, on peut imaginer que ces applications exposeront leurs propres données en RDF et les lieront à celles d'IdRef et donc du Sudoc – et réciproquement.

Elles sont fraîches mes données. Le Sudoc en RDF n'est pas un export de données Marc qui serait converti en RDF et mis à jour une fois par mois. C'est le Sudoc, en RDF, en temps réel.

Tout. Là encore, le Sudoc en RDF, c'est tout le Sudoc. Pas un extrait. Chaque notice Marc a son pendant RDF, mais chaque sous-zone Marc n'est pas convertie en RDF. Le Marc reste plus riche.

Pas de vocabulaire maison. Plutôt que de forger de nouvelles propriétés ou classes, nous n'avons utilisé que des vocabulaires du marché : Dublin Core, Foaf, Marc Relators surtout, mais aussi les vocabulaires RDF des ISBD et de RDA. Si nous avions forgé un vocabulaire Sudoc, nous aurions pu en dire plus, mais à qui ?

Pas de retraitement des données. Pour l'instant, nous nous sommes interdit toute restructuration des données Marc.

Oui aux redondances. Parfois, il peut être judicieux d'exprimer une seule information Marc sous deux formes en RDF pour s'adresser à des communautés métiers différentes.

Demain

Plus de données. Il nous faudra presser plus fort les données Marc pour en extraire plus de triplets RDF. Il faudra les triturer, mettre de la structure là où on trouve aujourd'hui des blocs de texte, faire émerger l'architecture FRBR implicite. Les vocabulaires RDA ou ISBD prendront de plus en plus de place, mais il faudra aussi parler à tous et pas seulement au monde des bibliothèques, en utilisant des vocabulaires plus généralistes.

Plus de liens. C'est surtout à travers les données d'autorité IdRef que celles du Sudoc vont se connecter à d'autres bases. IdRef sera lié à d'autres référentiels, comme Viaf, qui fédère des dizaines de fichiers d'autorité du monde entier – et vice versa.

Licences. L'Abes souhaite enfin associer aux données Sudoc (et aux autres bases qu'elle gère) une licence juridique la plus libérale possible. Ce dossier est en cours d'instruction. •

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

1 Version à jour du « nuage de données liées » : http://richard.cyganiak.de/2007/10/lod

tayeb.merabti@chu-rouen.fr

Tayeb Merabti est postdoc en informatique médicale au sein de l'équipe Cismef. En juin 2010, il a soutenu sa thèse intitulée « *Méthodes pour la mise en relation des terminologies médicales : contribution à l'interopérabilité sémantique inter et intra terminologique* » à l'université de Rouen. Il travaille principalement sur les méthodes et les outils pour mettre en relation les différentes terminologies médicales intégrées dans le système d'information de Cismef.

[fédérer] Le Cismef intègre désormais plusieurs terminologies au service de l'indexation et de l'interrogation de ses ressources. Un chantier mené à bien grâce à l'emploi de standards du web sémantique tels que RDF et Owl.

Le Cismef agrège ses terminologies pour une meilleure recherche dans ses fonds

Créé en 1995, le site Cismef (Catalogue et index des sites médicaux francophones)[1] recense et décrit plus de 23 000 sites et documents dans le domaine de la santé présents sur la Toile. Ce catalogue se caractérise par sa ligne éditoriale et par des descriptions documentaires précises. Depuis 2005, Cismef est passé d'un univers mono-terminologique fondé essentiellement sur le Mesh (Medical Subject Headings) vers un univers multi-terminologique incluant l'indexation automatique, la recherche d'information et l'intégration de plusieurs terminologies médicales.

Trente-deux terminologies intégrées

Pour intégrer toutes les terminologies médicales au sein d'une même structure globale, un modèle générique a été conçu pour le système d'information de Cismef (Cismef_SI). Ce modèle est fondé sur la notion de « descripteur » qui représente le concept principal de la terminologie « mot-clé ». Chaque descripteur peut être en relation avec d'autres descripteurs suivant des relations de type « exacte » ou « hiérarchique », par exemple. Il était aussi nécessaire de stocker les terminologies dans un format standard : la syntaxe RDF* a été choisie avec le format Owl. Le portail terminologique de santé (PTS) est un « portail terminologique » connecté au Cismef_SI pour faciliter la recherche des termes dans toutes les terminologies médicales qui y sont incluses. En septembre 2011, 32 terminologies étaient ainsi disponibles dans Cismef_SI et par conséquent dans le PTS. De plus, toutes ces terminologies sont accessibles gratuitement ou *via* un accès restreint[2].

Depuis 2005, l'équipe Cismef a développé, en collaboration avec la société Vidal, l'outil FMTI (French Multi-Terminology Indexer) : un algorithme d'indexation automatique multi-terminologique qui utilise plusieurs techniques de TALN (Traitement automatique du langage naturel). Outre le Mesh, d'autres terminologies ont été incluses dans FMTI. Ce dernier a permis l'indexation automatique de 33 951 ressources dans le catalogue Cismef et une indexation semi-automatique (supervisée) de 12 440 ressources à partir de leur titre. Pour adresser ce nouvel univers multi-terminologique, l'algorithme de recherche d'informations a été lui aussi modifié. Cette stratégie de recherche est ainsi fondée non seulement sur une expansion de requêtes s'appuyant sur l'enrichissement par synonymie et hiérarchisation, mais aussi par la mise en relation des différentes terminologies présentes.

L'interopérabilité inter-terminologique

Plusieurs améliorations ont également été effectuées pour permettre une interopérabilité inter et intra terminologique efficace. Des méthodes et algorithmes permettant la mise en relation des termes de chaque terminologie intégrée ont été implémentés. Ces améliorations ont commencé avant même le passage vers l'univers multi-terminologique avec la création de plusieurs synonymes Mesh (devenus synonymes Cismef) et la traduction de plusieurs termes Mesh (concepts chimiques), exclusivement de l'anglais vers le français. Ces ajouts ont permis une amélioration considérable de la recherche d'informations. Les premières méthodes, développées avec l'équipe Lertim de Marseille[3], utilisaient principalement le métathésaurus UMLS (Unified Medical Language System)[4] pour mettre en relation les termes des terminologies en français incluses dans ce métathésaurus. Par la suite, d'autres méthodes et algorithmes, fondés sur des techniques de TALN, ont été développés. Ils ont aussi permis la traduction automatique d'un nombre considérable de termes de l'anglais vers le français. •

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

1 http://www.cismef.org
2 http://pts.chu-rouen.fr
3 Laboratoire d'enseignement et de recherche sur le traitement de l'information médicale, http://cybertim.timone.univ-mrs.fr
4 Développé par la National Library of Medicine (NLM), http://www.nlm.nih.gov/research/umls

focus

Expert en modélisation documentaire à la BNF, **Anila Angjeli** anime dans des instances internationales de normalisation et à l'Afnor des groupes de travail sur la structuration des données d'autorité pour les bibliothèques et les archives. Elle copréside notamment le groupe de travail de la Société des archivistes américains sur le schéma EAC-CPF. Elle est également active dans le domaine des identifiants et représente la BnF dans le Comité des directeurs de l'Agence Internationale ISNI.

anila.angjeli@bnf.fr

[**repérer**] **Les standards du web sémantique concernent également le monde des bibliothèques, qui possède par ailleurs ses propres normes métier. Au-delà des transpositions et nécessaires adaptations, c'est la pensée normalisatrice qui s'en voit renouvelée.**

La normalisation en bibliothèque à l'heure du web sémantique

Le web sémantique offre enfin aux bibliothèques une possibilité sans précédent de tirer bénéfice des données structurées de leurs catalogues mais aussi de sortir de leur isolement, dû à la conception de systèmes d'information en silos, aux technologies vieillissantes des formats Marc non adaptés aux technologies du Web. Ce dernier a besoin aujourd'hui d'une information typée, structurée, normée, qui tient compte de la nature des données. Relier ces dernières par le sens, tel est le postulat du web sémantique. Dans la grande partition du Web intelligent, si le W3C bâtit le socle normatif technique, les normes métier ont à jouer leur partie au niveau du « sémantique », car elles fixent le sens en traduisant des concepts métier.

UN EXISTANT – DÉJÀ COMPLEXE

Notre propos se limitera ici aux normes concernant les bibliothèques qui rendent possible la production, la diffusion, l'échange de l'information produite et l'accès à cette information. En premier lieu sont concernées les normes, règles et recommandations appliquées dans la réalisation des catalogues – celles relatives à la description bibliographique et aux dispositifs associés facilitant l'accès aux catalogues. En font partie le bon vieux ISBD*, les normes Afnor de catalogage[1] et autres codes, mais aussi les normes de description archivistique, principalement la norme ISAD(G).

Les normes jouent aussi un rôle pour les identifiants de produits éditoriaux (ISBN, ISSN, ISMN), de contenus (Isan, ISWC, ISTC) ou d'« agents » (Isni). N'oublions pas les normes génériques pour le codage des caractères, des écritures, des noms de langues et de pays, ou les normes de translittération. S'ajoutent à cet arsenal les standards des formats de structuration des données bibliographiques et d'autorité (Marc et Iso 2709, puis MarcXchange, et Dublin Core), et les protocoles d'échange afférents (Z39.50, SRU, OAI-PMH).

Mais où s'arrêter ? L'adaptation à un environnement en perpétuelle mutation amène les bibliothèques à s'impliquer dans des travaux de normalisation qui auparavant étaient hors de leur périmètre d'action : standards de métadonnées relatifs à la gestion des objets numériques, comme METS, d'encodage du contenu des documents comme la TEI, et la liste est longue. La création en 2010 du groupe d'incubation du W3C « Bibliothèques et web de données » en est la preuve la plus récente.

Dans ce contexte mouvant, le paysage des acteurs reste complexe. Des organismes de normalisation internationaux tiennent le palmarès, comme l'ISO - TC 46, où les professionnels participent *via* les organismes nationaux de normalisation tels que l'Afnor pour la France, Ansi et Niso pour les États Unis, BSI pour le Royaume Uni, ou DIN pour l'Allemagne, qui par ailleurs produisent des normes nationales. Néanmoins, de nombreux autres organismes, experts dans leur domaine, rallient les professionnels autour de travaux de normalisation spécifiques. Mentionnons ici des associations professionnelles comme l'Ifla, la SAA, l'ICA, des organisations interprofessionnelles comme la DCMI, mais également des institutions isolées comme la Library of Congress, ou des organismes dédiés à la réalisation d'un projet comme le JSC pour l'élaboration de RDA, etc. Ces travaux de normalisation aboutissent non seulement à des normes à proprement parler – statut réservé à ce qui est produit par des organismes de normalisation dédiés – mais aussi à des standards, codes, bonnes pratiques, recommandations, tous aussi valables dès lors qu'ils sont adoptés largement par les communautés intéressées.

Interopérabilité, partage, échange – principes historiquement fondamentaux du processus de normalisation - resurgissent sous un nouvel angle. Bon nombre de normes ont été élaborées bien avant le web sémantique et, de ce fait, hors de ses problématiques. Pour autant, leur valeur n'est pas remise en cause par les développements récents.

TRANSPOSER SUR LE WEB DE DONNÉES

Typiquement, des normes génériques existantes ont pu facilement être portées sur le web sémantique. Citons ici le cas des normes Iso 639-2 et -5 des codes de langues. La Library of Congress, auto-

rité d'enregistrement Iso de ces normes, a attribué aux codes de celles-ci des URIs pérennes, permettant leur transposition sur le web des données.

Les standards traditionnels métier font ainsi alliance avec les standards du web sémantique. Des initiatives se multiplient pour exprimer ces standards en logique formelle ontologique et en RDF avec, en parallèle, la déclaration des espaces de nom, l'attribution d'URI et la mise à disposition sur des registres ouverts, en vue de leur réutilisation aisée sur le web des données, par des machines ou des humains. À titre d'exemple, mentionnons l'expression en RDF et la déclaration des éléments de Dublin Core. Le mouvement se poursuit avec l'expression, sous forme d'ontologies Owl, d'autres standards existants comme Mads, ou le schéma EAC-CPF. Il se prolonge avec des projets de publication sur le web de données d'éléments des vocabulaires associés aux modèles et normes de l'Ifla, comme c'est le cas des modèles FRBR, FRAD et de l'ISBD.

Mais ne banalisons pas le propos, car l'esprit du web sémantique et ce qui gravite autour infiltrent la pensée normalisatrice, avec pour corollaire un changement de posture dans les travaux. La matière à normaliser est pensée différemment. Dès le processus d'élaboration des normes et standards, corrélations et synergies entre les travaux de normalisation et d'autres initiatives sont recherchées. Les travaux adoptent une dynamique itérative, celle inhérente au mouvement du web sémantique.

ÉLABORER EN VUE DU WEB SÉMANTIQUE

Une illustration exemplaire est l'élaboration du nouveau code international de catalogage RDA, dont l'objectif est la modernisation de la production des données catalographiques et la structure des catalogues. Dès l'origine, il est fondé sur un modèle conceptuel métier (FRBR) et vise à concevoir l'information bibliographique comme un réseau d'informations interconnectées. Ses éléments et valeurs sont, au fur et à mesure, déclarés et publiés dans un registre de métadonnées, exprimés en RDF et identifiés par des URI.

Mais, au-delà du code lui-même, c'est le processus d'élaboration qui illustre cet esprit. S'y entrecroisent des collaborations avec d'autres initiatives dans un souci d'interopérabilité et de synergie. Ainsi le groupe RDA/ONIX, qui travaille à faciliter le transfert et l'utilisation de la description des ressources dans les deux communautés des bibliothèques et des éditeurs, a rendu son premier travail : un cadre ontologique distinguant enfin le contenu du contenant. Parallèlement, d'autres initiatives continuent dans ce sens et d'autres se mettent en place, comme l'étude des passerelles entre DCMI et RDA, l'harmonisation entre RDA et l'ISBD, etc. En France, au sein de la commission CG46 Documentation de l'Afnor, un groupe technique et un groupe stratégique étudient l'opportunité d'adoption du code et ses implications au niveau national[2].

Dernier-né de la famille des identifiants ISO, l'Isni est un identifiant unique, international, pour toute partie impliquée dans les maillons de la chaîne, allant de la création des contenus intellectuels et artistiques à la gestion des droits, qu'il s'agisse d'une personne ou d'une collectivité. Bien que les travaux de normalisation aient démarré en dehors de la problématique du web sémantique, ils ont eu pour toile de fond l'économie du numérique et les besoins et usages qu'elle génère. Émanant d'une collaboration entre acteurs des secteurs public et privé, Isni se positionne en identifiant passerelle. Il est destiné à relier entre eux des systèmes d'information, à la fois ceux contenant des données sensibles et ceux destinés à une diffusion large de données. La connexion se fera *via* la « couche » publique des données, susceptible d'être exposée sur le Web. Les identifiants Isni et les métadonnées publiques qui les accompagnent (notamment des URI vers des réservoirs contenant des informations d'autorité sur l'identité en question) seront exposés sur le web des données en RDF, de même que les liens entre Isni et Viaf.

Un identifiant unique pour les « identités publiques » des parties, ne tenant compte ni de leur appartenance nationale, ni de leur secteur d'activité, de portée mondiale et dont le succès réside dans l'adoption large par tous les secteurs concernés ! Avec ces caractéristiques, Isni s'inscrit pleinement dans la logique d'interopérabilité par l'identification des objets informationnels à mettre en relation. D'autres collaborations sont en cours : l'Orcid pour les acteurs du domaine scientifique et académique et le Niso I[2] pour les institutions.

On le voit : le mouvement du web sémantique ne fait que confirmer les motivations et principes fondamentaux de la normalisation : consensus des parties prenantes, optimisation des processus de travail, retombées économiques, interopérabilité. La tension permanente entre normalisation, au sens de réduction des variétés, et besoin d'expression des diversités et des spécificités s'atténue par la mise en avant du *sens*. •

Les sigles en jeu

Ansi : American National Standards Institute | **BSI** : British Standards institution | **DCMI** : Dublin Core Metadata Initiative | **Din** : Deutsches Institut für Normung e.V. | **EAC-CPF** : Encoded Archival Context – Corporate bodies, Persons and Families | **FRAD** : Functional Requirements for Authority Data | **FRBR** : Functionnal Requirements for Bibliographic Records | **ICA** : Conseil international des archives | **Isad(G)** : International Standard Archival Description-General | **Isan** : International Standard Audiovisual Number | **ISBD** : International standard bibliographic description | **ISMN** : International Standard Music Number | **Isni** : International Standard Name Identifier | **ISTC** : International Standard Text Code | **ISWC** : International Standard musical Work Code | **JSC – RDA** : Joint Steering Committee for Development of Ressource Description and Access | **Mads** : Metadata Authority Description Schema in RDF | **Marc** : Machine-Read Cataloging | **Mets** : Metadata Encoding & Transmission Standard | **Niso** : National Information Standards Organization | **OAI-PMH** : Open Archives Initiative – Protocole for Metadata Harvesting | **Onix** : Online Information Exchange | **Orcid** : Open Researcher & Contributor ID | **SAA** : Society of American Archivists | **SRU** : Search/Retrieval via URL | **TEI** : Text Encoding Initiative | **Viaf** : Virtual International Authority File •

** Les sigles des normes métier sont développés dans l'encadré ci-dessus et ceux relatifs au web sémantique sont développés en page 29.*

1 http://www.bnf.fr/fr/professionnels/normes_francaises/s.cat_normes_françaises.html
2 Ressource Description and Access (RDA) : en France http://www.bnf.professionnels/rda/s.rda_en_france.html

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

yves.raimond@gmail.com

Senior Research Engineer à la BBC, au sein du département Recherche & dévelopement, Yves Raimond a travaillé sur BBC Programmes et depuis mi-2011 sur un projet d'extraction de données d'archives. Il a obtenu son doctorat (Queen Mary, University of London) dans le domaine du web sémantique et de l'informatique musicale en 2008 et a étudié à Télécom Paristech de 2002 à 2005.

[faciliter] L'offre de la BBC est connue pour être très riche. L'opérateur britannique tire un double bénéfice de son usage des techniques du web de données : il pérennise ses contenus en leur attribuant une adresse persistante et en facilite l'accès en les agrégeant, quelle que soit leur origine.

Les programmes de la BBC tirent avantage du web de données

Garantir une présence en ligne de qualité pour les 1 000 à 1 500 programmes diffusés chaque jour était jusqu'à récemment une tâche quasi-impossible. Seuls quelques programmes, parmi les plus populaires, pouvaient s'offrir une présence sur le Web, sous la forme d'une multitude de sites indépendants, peu souvent liés entre eux et sans cohérence au plan de l'expérience utilisateur. Tous ces sites, de par leur hétérogénéité, sont difficilement maintenables et fréquemment abandonnés après quelques années. Ce dernier point pose un vrai problème : les URI* pour nos programmes doivent être stables et persistants, afin de permettre aux utilisateurs de créer des liens vers eux, et de constituer un point d'ancrage pour des discussions autour de nos programmes. Ce modèle implique un coût, en terme de développement, mais aussi en terme d'opportunités : en effet, le temps investi à maintenir un ensemble de sites hétérogènes est autant de temps perdu pour innover.

BBC Programmes[1] est une première étape vers le support automatisé de programmes sur le Web. Le service agrège l'information provenant de plusieurs sources (données de production, archives, données éditoriales, etc.), et crée automatiquement un URI persistant pour chaque programme de la BBC, suivant les principes de données liées[2] énoncés par Tim Berners-Lee. Pour chacun de ces URI servant de point d'ancrage pour tout type de données concernant le programme correspondant, nous offrons différentes représentations par négociation de contenu : mobile, desktop, grand écran, mais aussi sous forme de données structurées disponibles en RDF, JSON et XML. Ainsi, notre site web est aussi notre API. Tout ces flux sont basés sur la BBC Programmes Ontology[3], qui décrit un certain nombre de concepts tels qu'une série, un épisode, une transmission, un segment, ou une fenêtre de disponibilité.

BBC Music[4] et BBC Wildlife Finder[5] suivent les mêmes principes : créer une page par entité, dans deux autres domaines, comme point d'ancrage à des données agrégées les concernant. Ces deux services ont comme particularité d'utiliser le Web comme système de gestion de contenus. Nos éditeurs contribuent à Musicbrainz[6] et Wikipédia[7], et les modifications qui y sont apporté sont automatiquement répercutées sur les sites de la BBC. Cela nous permet d'être très réactifs en cas d'évènements imprévus, mais aussi de contribuer à la qualité des informations sur ces plates-formes collaboratives. BBC Music et BBC Wildlife Finder mettent aussi à disposition leurs données en se basant respectivement sur la Music Ontology1[8] et la BBC Wildlife Ontology2[9].

Le fondement de ces trois projets est la notion de « page par entité », qui permet de créer des expériences utilisateurs riches, avec la navigation d'entité en entité et de domaine en domaine. Lier nos données (entre elles, et *via* d'autres sources de données comme DBpedia[10] ou Freebase[11]) explicite les liens qui existent entre ces différents domaines, et permet à nos équipes aussi bien qu'à des développeurs externes de déployer des applications utilisant ces liens. Ces applications externes basées sur nos données génèrent du trafic vers le site de la BBC, et nous donnent des idées de nouvelles sources de données à intégrer, ou de nouvelles façons de naviguer dans nos contenus. Wildlife Finder, exemple d'utilisation des liens au sein du site de la BBC, propose sur la page d'une espèce particulière (par exemple le lion) une liste de programmes traitant de cette espèce. •

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

1 http://www.bbc.co.uk/programmes
2 http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html
3 http://www.bbc.co.uk/ontologies/programmes/2009-09-07.shtml
4 http://www.bbc.co.uk/music
5 http://www.bbc.co.uk/nature/wildlife
6 http://musicbrainz.org/
7 http://www.wikipedia.org/
8 http://musicontology.com/
9 http://www.bbc.co.uk/ontologies/wildlife/2010-11-04.shtml
10 http://dbpedia.org/About
11 http://www.freebase.com/
12 http://www.bbc.co.uk/nature/life/Lion
13 Selon Alexa : http://www.alexa.com/siteinfo/bbc.co.uk

dominique.cotte@ourouk.fr

Dominique Cotte est professeur à Lille (laboratoire Geriico). Il travaille également comme consultant pour les sociétés Aphania et Ourouk sur l'ingéniérie du document numérique et les systèmes d'information. Son dernier ouvrage *Émergences et transformation des formes médiatiques* est paru en 2011 aux éditions Hermès.

[**positionner**] Le mouvement général du web sémantique mobilise des compétences parfaitement intégrées par les documentalistes. Mais plus qu'une simple adaptation, cette évolution suppose de remettre en cause ses propres conceptions et comportements en matière de traitement de l'information.

Les nouvelles perspectives du web sémantique pour les professionnels de l'information

L'appellation de « web sémantique » pourrait faire croire aux professionnels de l'information qu'ils sont moins concernés que les informaticiens ou les linguistes par ce nouveau domaine. Ce serait une erreur, liée notamment à la compréhension de ce qui relève du sémantique dans l'expression *web sémantique*, Tim Berners-Lee lui-même ayant entrepris de critiquer rétroactivement cette notion (voir à ce sujet l'excellent article d'Alexandre Monnin sur la définition du web sémantique[1] ainsi que le papier de James Hendler[2]).

De ce point de vue, parler de « web de données » ou de *linked data* (les deux concepts ne se recouvrant pas) permet de mieux comprendre comment le mouvement général du web sémantique mobilise des compétences parfaitement intégrées par les documentalistes, et ceci de longue date. Nous examinerons dans cet article la façon dont les différentes tâches nécessaires à une pratique de gestion de l'information dans la logique et les standards du web sémantique sont en parfaite continuité avec les savoir-faire documentaires. Néanmoins, il ne faut pas négliger le fait que ce transfert de compétences sur un terrain nouveau est plus qu'un simple déplacement. Il implique également un changement de paradigme et impose d'acquérir de nouveaux comportements.

Travailler dans et avec le web de données suppose de se placer dans les perspectives suivantes : rompre avec la logique de clôture du « fonds documentaire » ; rompre avec la logique du travail sur « ses » données et s'impliquer dans une perspective d'ouverture ; se confronter à une technicité plus importante que celle de l'usage de l'informatique documentaire classique ; dépasser le rôle de relais d'une information documentaire déjà élaborée pour réaliser (ou participer à l'élaboration) de supports d'information. Tous ces éléments posent bien sûr des questions d'identité professionnelle et interrogent les frontières avec les métiers existants, eux-mêmes déstabilisés par l'explosion du numérique. Comme dans tout grand changement technologique, les frontières entre les occupations, les tâches et les métiers ont tendance à être redéfinies. Autrement dit, dans le grand mouvement de « documentarisation » du Web qu'incarne le web sémantique, une place pour les tâches documentaires existe, mais il appartient seulement aux professionnels de l'information d'en prendre conscience et de s'approprier cette part de l'activité, sans toutefois chercher à aller trop loin.

À notre sens, on ne peut aujourd'hui éditer de l'information sur le Web sans faire appel à des pratiques documentaires, mais cela ne veut pas dire que les documentalistes doivent devenir des éditeurs et s'emparer de toute la chaîne de production. Au fondement des métiers de la gestion de l'information se retrouvent les besoins suivants, quels que soient les outils et les environnements techniques : structurer, normaliser, qualifier l'information. Le web sémantique, quelle que soit finalement l'interprétation plus ou moins extensive que l'on peut en donner, non seulement n'échappe pas à cette règle, mais il vient au contraire remettre sur le devant de la scène ce type de compétences que les évolutions du Web tout court (et les illusions informatiques en général) avaient un temps semblé reléguer à l'arrière-plan.

En effet, quelles sont les principales tâches de traitement de l'information dans le web sémantique ? Comme en

documentation, il s'agira de sélectionner, structurer, normaliser, qualifier et lier entre elles des ressources.

Sélectionner des sources et des ressources

Le mouvement du *linking of data* (tout comme par ailleurs l'ouverture des données publiques) implique la forte structuration d'un existant qu'il n'est, par définition, pas nécessaire de recréer en d'autres endroits. De ce point de vue, il existe bien une rupture avec le paradigme documentaire basé sur la localisation ou la détention *in situ* de ressources pour pouvoir les communiquer à un public plus ou moins captif. Dans cette perspective, le documentaliste emprunte au veilleur les techniques de *sourcing*, d'identification de ressources pertinentes et, surtout, de qualification en fonction d'un projet et d'un contexte donnés. Créer, à des fins de publication sur un site web, d'alimentation de flux RSS ou de pages Netvibes, des agrégats de contenu suppose d'avoir défini, à titre de filtre intellectuel et humain, les critères de pertinence qui répondront au mieux aux besoins des utilisateurs.

La même chose vaut pour les ontologies. Comme le souligne Jeffrey Heflin[3], il convient de limiter la prolifération des ontologies sur un même sujet et d'essayer au maximum de réutiliser l'existant, en prenant évidemment en compte les spécificités locales, qu'elles soient nationales ou censées refléter le point de vue d'une organisation.

Une compétence documentaire sollicitée par le web de données consiste en la qualification des ressources. Les documentalistes n'auront plus (ou plus que partiellement) à constituer des « fonds documentaires » qui leur seraient spécifiques.

Structurer de l'information sur des objets

Le travail documentaire portait le plus souvent, comme son nom l'indique, sur ces objets achevés que sont les documents. On sait qu'une première brèche a été ouverte dès lors que le document, devenu numérique, évoluait vers une structure ouverte, granulaire et recomposable[4]. L'idée du web de données est bien de repérer et de structurer des grains d'information à l'intérieur des structures documentaires pour les exploiter en les reliant entre elles de manière logique et standardisée. Les standards du web sémantique comme RDF créent du sens en typant des relations entre objets, qui permettent ensuite de relier des ressources de façon pérenne.

Annoter les documents

Dans le vocabulaire du web sémantique, on parle souvent d'annotations, ce qui peut être source de confusions. Classiquement, l'annotation est un ajout de type commentaire, effectué pour soi-même ou pour autrui sur un document déjà existant. Elle vise dans ce cas à éclairer le texte et pas à le décrire. Dans le monde du web sémantique, *annoter* c'est taguer, autrement dit indexer en langage documentaire. L'activité traditionnelle consistait à indexer des documents « du dehors » alors qu'ici, on indexe des données à l'intérieur d'un document ou d'une micro unité d'information.

Structurer de l'information pour sa publication

Afin de communiquer correctement avec d'autres ressources, l'information, dans le web sémantique (mais ceci ne fait que poursuivre une des exigences du Web, et plus généralement, de l'informatique), doit être correctement structurée. Traditionnellement, les documentalistes sont habitués à indexer des documents à partir de formulaires de bases de données qui organisent leur description dans des champs. Il s'agit donc en quelque sorte d'une description externe, qui vient préciser du dehors des éléments sur le document, la fiche ainsi produite ne se substituant pas au document, et le document original (dit « primaire ») ne contenant pas tous les éléments de la fiche (par exemple son indexation ou les vedettes matières). La logique du document numérique en général et plus particulièrement sur le Web est différente puisque les données de description (ce que l'on appelle les métadonnées) sont directement intégrées dans le document. Cette logique d'encapsulation est poussée encore plus loin dans le web sémantique puisqu'elle se fait sous un format de données qui permettra d'enrichir le document original avec une information complémentaire. Typiquement, si un document mentionne un nom de personne et qu'il existe quelque part une ressource identifiable qui me permet de connaître des éléments biographiques de cette personne, je peux enrichir ma première ressource par l'ajout d'un lien vers ce deuxième élément.

Le champ de l'entreprise

Dans l'expression web sémantique, le terme web ne désigne pas seulement la Toile ; il s'agit des technologies du Web, lesquelles ne se déploient pas seulement dans l'espace de l'Internet mais renouvellent également l'informatique d'entreprise. Le développement des intranets « 2.0 », des architectures Soa[5], la mise en place de services web[6] sont autant de phénomènes qui favorisent et exigent la mise en place des standards du web sémantique. Au lieu de décrire *n* fois des données dans différentes applications, ou d'essayer de les concentrer et de les normaliser dans des ERP dont la mise en place est extrêmement lourde, le web sémantique permet d'organiser une circulation plus fluide entre les différents gisements d'information. La cartographie des données, la réalisation d'ontologies propres au micro domaine de l'entreprise sont autant d'exemples pour lesquelles des compétences documentaires sont requises. Quant aux entreprises publiques ou semi-publiques, aux services de l'État et des collectivités, leur présence sur le web de données et l'ouverture de leurs fonds au public exigeront un investissement croissant des professionnels de l'information. •

1 http://cblog.culture.fr/2011/09/07/web-semantique-iri-opendat
2 http://blogs.nature.com/jhendler/2009/06/16/what-is-the-semantic-web-really-all-about
3 http://www.cse.lehigh.edu/~heflin/pubs/heflin-thesis-orig.pdf
4 Roger T. Pédauque, Le document à la lumière du numérique, C&F éditions, 2006
5 Service Oriented Architecture : architecture orientée services.
6 Programme informatique permettant la communication et l'échange de données entre applications et systèmes hétérogènes dans des environnements distribués (source : Wikipédia).
7 Enterprise Resource Planning : progiciel de gestion intégré (PGI).