RÉPUBLIQUE FRANÇAISE

Ministère de la culture et de la communication

Direction générale des patrimoines

Service interministériel des Archives de France

Note d'information DGP/SIAF/2012/015 en date du 12 novembre 2012

relative aux évolutions de l'outil de consultation du *Thésaurus pour la description et l'indexation des archives locales*

Le directeur chargé des archives de France

à

Mesdames et Messieurs les directeurs des services centraux nationaux des archives nationales
Mesdames et Messieurs les directeurs des services départementaux d'archives
sous couvert de Mesdames et Messieurs les préfets de région
et de Mesdames et Messieurs les préfets de département

Depuis mars 2011, le *Thésaurus pour la description et l'indexation des archives locales anciennes, modernes et contemporaines* est accessible en ligne via un outil de consultation[1], afin d'en faciliter l'utilisation tant par les archivistes ou les internautes que par les éditeurs de logiciels documentaires. Cet outil de consultation du thésaurus a récemment fait l'objet d'évolutions majeures (transformation des identifiants de concepts en identifiants pérennes de type ARK, ajout d'une fonction de recherche, aménagement en vue de l'intégration prochaine d'autres vocabulaires contrôlés du ministère de la Culture et de la Communication - notamment des services du patrimoine et de l'architecture). La présente note d'information fait état de ces évolutions.

1. Le Thésaurus pour l'indexation des archives locales et le projet d'harmonisation des données culturelles (HADOC)

Les évolutions de la plate-forme de consultation du *Thésaurus pour la description et l'indexation des archives locales* s'inscrivent dans le cadre plus général du projet d'Harmonisation des données culturelles (HADOC).

[1] Pour en savoir plus, voir la note d'information du 14 mars 2011 relative à la mise en ligne de l'application de consultation : <http://www.archivesdefrance.culture.gouv.fr/gerer/classement/normes-outils/thesaurus/>

Inscrit dans l'axe stratégique « Documentaire » du schéma directeur des systèmes d'information du Ministère de la Culture et de la Communication (MCC) et piloté, depuis 2010, par le Département des systèmes d'information patrimoniaux au sein de la Direction générale des patrimoines (DGPat), le projet HADOC réunit au sein de différents groupes de travail des acteurs de tous les services métiers de la DGPat, dont les Archives de France. L'enjeu majeur assigné au projet est d'**accroître la qualité des données et d'améliorer les processus de production** en agissant sur l'harmonisation des données, l'harmonisation des processus liés à la production de ces données et la généralisation des normes et standards de production.

Une réflexion est à présent engagée sur la mise en œuvre d'un nouvel environnement de gestion des vocabulaires scientifiques et techniques qui doivent constituer à terme un élément fort de mise en cohérence de l'ensemble de la production documentaire du ministère. **L'objectif est de créer un référentiel terminologique unifié permettant d'offrir aux usagers un accès unique et cohérent aux ressources terminologiques produites par le MCC** et d'en démultiplier les usages. Sont concernés de 70 à 100 vocabulaires scientifiques et techniques tels que les vocabulaires de la désignation, les vocabulaires relatifs à la description (matérielle et conceptuelle) des biens culturels, les vocabulaires relatifs à l'histoire des œuvres (auteurs, activités, chronologie, géographie historique etc.). Ces vocabulaires étant gérés dans diverses applications, le MCC souhaite disposer d'un outil spécifique pour concevoir, maintenir et diffuser ces différents vocabulaires. Un prestataire devrait être prochainement être retenu pour le développement de cet outil. Pour le moment, le MCC a souhaité permettre à l'application de consultation du *Thésaurus pour la description et l'indexation des archives locales* de devenir une brique de son système d'information pour tout ce qui concerne la diffusion de référentiels terminologiques.

2. Principales évolutions de l'outil de consultation

Jusqu'à présent, n'étaient présentés dans la plate-forme de consultation que le *Thésaurus pour la description et l'indexation des archives locales* et les trois listes d'autorité non hiérarchisées l'accompagnant (actions administratives, contexte historique et typologies documentaires). L'application permettait d'accéder aux concepts principaux du thésaurus (intitulés des onze chapitres entre lesquels sont répartis les descripteurs et non descripteurs), puis de naviguer à l'intérieur en suivant les relations « générique/spécifique », « termes associés » et éventuellement « termes équivalents » dans d'autres référentiels (Dbpedia, Rameau). Il était possible de récupérer le thésaurus et les listes d'autorité (tout ou partie) au format SKOS[2], l'outil affichant les propriétés SKOS du référentiel pour chaque concept. Enfin, le thésaurus pouvait être interrogé par l'intermédiaire d'un service Web (lien « SPARQL endpoint ») avec le protocole et langage de requête SPARQL [3].

Une première évolution devrait concerner le **nombre de vocabulaires consultables dans l'interface**, avec la mise à disposition prochaine du *Thésaurus de la désignation*, qui permet l'indexation des notices des bases Mérimée et Palissy relatives aux œuvres architecturales et mobilières[4]. En conséquence, **le titre de l'application sur la page d'accueil a été transformé en « Les vocabulaires du Ministère de la Culture et de la Communication ».** Pour le moment, il a été décidé de conserver la page d'accueil actuelle avec la liste des référentiels triée par ordre alphabétique.

---

[2] SKOS (« Simple Knowledge Organization System » ou « Système simple d'organisation des connaissances ») est un langage de représentation des thésaurus, classifications ou tout autre vocabulaire contrôlé et structuré. SKOS est depuis le 18 août 2009 une recommandation du World Wide Web Consortium ou W3C.

[3] Voir note d'information DGP/SIAF/2011/006 du 14 mars 2011 <http://www.archivesdefrance.culture.gouv.fr/static/4698>

[4] Pour en savoir plus sur le *Thésaurus de la désignation*, voir: <http://www.culture.gouv.fr/culture/inventai/patrimoine/>

Toutefois, dans le futur, le nombre de référentiels croissant, cet affichage devrait évoluer. L'objectif d'HADOC étant d'essayer de partager le plus possible de référentiels entre domaines métier**, les vocabulaires pourraient être organisés par facette ou par domaine fonctionnel** correspondant aux concepts structurants du projet (par exemple, les vocabulaires de la désignation, les vocabulaires de la description physique, les vocabulaires de la description iconographique etc.).

Une autre modification concerne l'**affichage d'informations spécifiques sur la page de présentation de chaque référentiel**. On y trouvera désormais :

- des informations statiques (ne variant pas d'un référentiel à l'autre) : lien vers la page d'accueil de l'application, lien « SPARQL endpoint » vers le service Web d'interrogation et lien « Téléchargement » permettant le téléchargement du référentiel au format RDF/XML[5] ;
- des informations dépendant du référentiel : lien vers la représentation au format RDF/XML de l'information affichée dans la page (ensemble du référentiel, ou concept sélectionné), lien vers le « producteur » du référentiel, lien « En savoir plus ».

Les pages d'affichage des concepts ont évolué. Jusqu'à présent, seuls étaient affichés le terme utilisé pour représenter le concept (skos:prefLabel), le concept générique (skos:broader), éventuellement le(les) non descripteur (s) (skos:altLabel), le(les) concepts spécifique(s) (skos:narrower), le(les) concept(s) associé(s) (skos:related), le vocabulaire source, et au besoin une note d'application (skos:scopeNote). **Apparaissent désormais des propriétés SKOS non gérées jusqu'à présent**, par exemple, pour les notes de mise à jour (skos:changeNote), l'auteur de la note, la date, voire des liens vers d'autres ressources. **Les descripteurs utilisés pour représenter tel ou tel concept sont affichés dans toutes les langues présentes dans le référentiel** (descripteurs en français en premier, les autres par ordre alphabétique du code de langue, puis par ordre alphabétique du descripteur). Cette dernière fonctionnalité a été introduite pour le *Thésaurus de la désignation,* qui comprend, pour certains termes, des équivalents linguistiques, comme par exemple « abbey » pour « abbaye », « aqueduct » pour « aqueduc », « house » pour « maison », etc. Les non descripteurs peuvent également être affichés dans différentes langues.

L'affichage des alignements avec d'autres vocabulaires a également été modifié. Les concepts du *Thésaurus pour l'indexation des archives locales* peuvent en effet être reliés à des concepts d'autres thésaurus externes (dbpedia et RAMEAU), à travers des relations de type skos:exactMatch (« terme équivalent »), skos:closeMatch (« terme approchant ») ou foaf:focus (« ressource équivalente »). A la publication, l'affichage de ces concepts par l'interface précédente faisait apparaître :

- pour les identifiants RAMEAU, l'étiquette du concept équivalent ou approchant ;
- pour les identifiants dbpedia, l'étiquette du concept équivalent ou approchant ainsi qu'un commentaire et la page équivalente (en français de préférence).

Pour assurer cet affichage, l'ensemble des triplets RDF[6] correspondants étaient extraits des référentiels externes et alimentaient la base de données RDF (triple-store) du MCC de manière statique. On observait donc au bout d'un certain temps des décalages entre ce que qui avait été chargé dans le triple-store du MCC et ce qui existait dans les triple-stores des référentiels externes (dbpedia et RAMEAU). D'autre part, quand on ajoutait dans le *Thésaurus pour la description et l'indexation des archives locales* de nouveaux liens vers des concepts de RAMEAU ou de dbpedia qui n'étaient pas dans la dernière extraction, l'interface ne les affichait pas (puisqu'elle ne trouvait aucune information dans le

[5] Développé par le W3C, RDF (Resource Description Framework) est le langage de base du Web sémantique. RDF est un modèle de description des ressources Web et de leurs métadonnées sous forme de graphes, de façon à permettre le traitement automatique de telles descriptions.

[6] RDF consiste en déclarations simples appelées triplets, dont la structure est invariablement de la forme de sujet-prédicat-objet, par exemple « Jean a trois enfants », « Jean est marié à Marie ». Un triplestore (ou triple store) est une base de données destinée au stockage des données du web de données : les triplets.

triple-store). Il fallait donc soit avoir une interrogation dynamique des référentiels externes lors de leur affichage, soit mettre régulièrement à jour les extractions que l'on en faisait.

L'alignement de thésaurus, même dans le cas d'une simple mise à jour, **ne pouvant pas se faire entièrement de manière automatique** et demandant une relecture par un expert métier, **il a été décidé d'afficher l'URI des concepts alignés, et non leur étiquette**. Apparaissent donc :

- le type d'alignement (« « Concept(s) équivalent(s) dans d'autres vocabulaires » ou « Concept(s) approchant(s) dans d'autres vocabulaires » ou « Ressource(s) associée(s) ») ;
- l'ensemble des URI.

L'utilisateur déduira l'identité du vocabulaire en fonction de l'URI. Par exemple, on peut déduire de l'URI http://stitch.cs.vu.nl/vocabularies/rameau/ark:/12148/cb11934328n qu'il s'agit du vocabulaire RAMEAU.

Par ailleurs, les normes et standards actuellement en vigueur pour la production des vocabulaires et pour leur mise à disposition en ligne insistent sur la **nécessité de recourir à un système d'identification unique et permanente de chaque concept**. Cette contrainte normative rencontre la demande des utilisateurs des vocabulaires scientifiques et techniques du MCC, qui veulent pouvoir disposer de véritables référentiels terminologiques. La mise en place d'un système d'identification permanente constitue un pré-requis indispensable à la production et à la publication de vocabulaires de référence dans le cadre du projet HADOC. En l'absence d'un guide général de nommage des ressources du MCC, il avait été décidé dans un premier temps de construire les identifiants du *Thésaurus pour la description et l'indexation des archives locales* sur le nom de domaine archivesdefrance, par exemple http://www.archivesdefrance.culture.gouv.fr/thesaurus/resource/T1-246 pour le concept « agriculture », la deuxième partie de l'URI correspondant à l'identifiant du vocabulaire et à celui du concept au sein du vocabulaire. Le MCC a souhaité profiter de l'extension de l'outil de consultation aux autres vocabulaires scientifiques pour **mettre en place des directives de nommage communes** afin de faciliter la citabilité et le référencement des ressources du MCC. **Le système d'identification pérenne choisi est l'Archival Resource Key (ARK).** La syntaxe des URI des vocabulaires et des concepts gérés dans la plate-forme est désormais de la forme suivante : {autorité d'adressage}{identifiant}. L'autorité d'adressage a pour valeur par défaut « http://data.culture.gouv.fr/thesaurus/resource/ ». L'identifiant est de la forme : ark:/67717/idvoc-idconcept (« 67717 » correspond à l'identifiant du ministère de la culture, « idvoc » correspond à l'identifiant du vocabulaire et « idconcept » à l'identifiant du concept dans le vocabulaire). On aura donc : http://data.culture.fr/thesaurus/resource/ark:/67717/T1-246 pour le concept « agriculture ». La mise en place de ce système d'identification pérenne a entraîné la création d'un espace de nom « data.culture.gouv.fr » qui devrait à terme servir à diffuser toutes les ressources du Web sémantique du MCC[7].

Enfin, outre la navigation au sein des vocabulaires et l'interrogation avec le langage SPARQL, **l'outil de consultation offre des fonctionnalités de recherche simple**. Une boîte de recherche a été ajoutée sur la page d'accueil, **permettant de rechercher en plein texte, dans tous les référentiels, sans restriction de langue, tous les termes représentant un concept** : descripteurs (skos:prefLabel) et non descripteurs (skos:altLabel), mais aussi formes lexicales destinées à rester cachées dans une interface de visualisation tout en demeurant disponibles pour les recherches plein texte (skos:hiddenLabel), ce qui permet de retrouver le concept, même après saisie d'une faute d'orthographe. **La recherche plein texte porte également sur les notes d'application** (skos:scopeNote). Elle ne tient pas compte de la casse mais tient compte des accents. La page de résultat contient une navigation dans les pages de résultats, la présentation des résultats dans un tableau et une ligne par élément de réponse avec un lien vers la page d'affichage du concept. Pour chaque résultat sont affichés : le référentiel dont le concept est issu, l'identifiant du référentiel, le terme

[7] La plate-forme de consultation des thésaurus du MCC est accessible à : <http://data.culture.fr/thesaurus/>

représentant le concept et le terme correspondant à la recherche (on ne précise pas le type de terme qui peut être un descripteur, un non descripteur ou une forme lexicale « cachée »). Les lignes du tableau sont triées par ordre alphabétique du référentiel source, puis par ordre alphabétique du descripteur représentant le concept.

3. Perspectives et améliorations possibles

Les informations respectives du format SKOS et de la nouvelle norme ISO 25964 sur les thésaurus[8] sont proches mais pas tout à fait identiques. C'est ainsi que la norme ISO-25964 ne définit pas la forme que prendront les informations, mais uniquement la manière de les structurer, tandis que SKOS, construit sur la base du RDF, est défini par une ontologie OWL de représentation des connaissances. Par ailleurs, la norme ISO 25964 introduit des précisions qu'il n'est pas possible de représenter nativement avec SKOS, comme par exemple la notion de relais virtuels, permettant de regrouper les concepts indépendamment de leur organisation hiérarchique, ou encore la gestion des différentes versions d'un thésaurus.

Du fait des limites du langage SKOS et des outils de production et de gestion des vocabulaires dont dispose actuellement le MCC, l'affichage du concept est donc imparfait puisque n'apparaissent ni le statut des concepts (candidats, valides, obsolètes), ni les thèmes ou facettes auxquels ils peuvent se rattacher indépendamment de l'organisation hiérarchique du thésaurus. Concernant le statut des concepts, une solution provisoire a été adoptée, consistant à indiquer, dans des notes de mise à jour (skos:changeNote), que certains concepts avaient été rendus obsolètes par rapport à la dernière version du *Thésaurus pour la description et l'indexation des archives locales*[9]. Toutefois, cette solution n'est pas pleinement satisfaisante dans la mesure où ces informations restent seulement lisibles par un humain et ne sont pas exploitables par la machine.

Plus fondamentalement, le *Thésaurus pour la description et l'indexation des archives locales* pourrait être rendu plus conforme par rapport à la norme ISO 25964-1 afin de faciliter son rapprochement avec d'autres vocabulaires. Outre la distinction entre le concept et les termes qui les représentent, ce qui permet une manipulation plus aisée des concepts, et la prise en compte du multilinguisme, auparavant traité dans une norme distincte de celle des thésaurus monolingues, la norme offre en effet la possibilité de caractériser beaucoup plus finement les relations sémantiques ou lexicales. C'est ainsi qu'elle définit trois types de relations hiérarchiques : génériques (genre à espèce ; par exemple : salade/laitue, oiseau/moineau, etc.), partitive (tout à partie ; par exemple : armée/armée de terre, rosaire/grain de rosaire, etc.) et d'instance (entre une catégorie générale de choses ou d'événements, exprimée par un nom commun, et un spécimen individuel de cette catégorie ; par exemple : région montagneuse/Alpes, etc.). Or, dans le *Thésaurus pour la description et l'indexation des archives locales*, les relations hiérarchiques peuvent être de nature différente sous un même terme générique, voire ne rentrer dans aucune des trois catégories définies par la norme, ce qui est le cas par exemple, de la relation entre les concepts « environnement » et « équipement », peu explicite pour des non utilisateurs de ce référentiel. Le *Thésaurus pour la description et l'indexation des archives locales* relève plutôt d'une logique classificatoire et le positionnement des concepts dans la hiérarchie ne permet pas

---

[8] *ISO 25964-1:2011 -- Information et documentation -- Thésaurus et interopérabilité avec d'autres vocabulaires -- Partie 1: Thésaurus pour la recherche documentaire.* La première partie de cette norme a été publiée en juillet 2011, la deuxième sur l'interopérabilité avec d'autres vocabulaires est attendue avant la fin de l'année 2012.

[9] C'est le cas par exemple des concepts : « aide sociale facultative », fusionné avec « aide sociale légale » pour donner « prestation d'aide sociale légale » ; « discipline scientifique », remplacé par « sciences de la vie et de la terre » et « sciences humaines » ; « monument historique » et « édifice classé », devenus non descripteurs de « patrimoine architectural » ; « objet d'art », devenu non descripteur de « œuvre d'art » ; « hygiène sociale » devenu non descripteur de « action sanitaire » ; « alcoolique », remplacé par « alcoolisme » ; « marché d'intérêt national », devenu non descripteur de « marché en gros » ; « gitan », devenu non descripteur de « nomade ».

toujours de désambiguïser les synonymes, ce qui explique en partie que l'alignement des concepts du thésaurus-matières avec les concepts représentés dans RAMEAU et dbpedia soit incomplet.

Le directeur, chargé des Archives de France

Hervé LEMOINE