Le multilinguisme dans les projets européens

Rendre le patrimoine culturel numérisé accessible à tous dans sa diversité implique de partager normes et bonnes pratiques pour la description des collections et des documents, et pour leur interrogation par les moteurs de recherche sémantique et multilingue.

MARIE-VÉRONIQUE LEROI
SG / SCPCI / DREST

Aujourd'hui, à travers toute l'Europe, des centaines d'institutions culturelles numérisent leurs collections pour les conserver et, surtout, pour permettre au grand public d'y accéder sur Internet. Face à la diversité et à la richesse de l'offre numérique issue de ces programmes de numérisation, des réseaux d'institutions et de ministères, tel le réseau Minerva[1], se sont constitués avec l'appui de la Commission européenne, afin de donner une meilleure visibilité à l'ensemble des initiatives.

Le projet européen MICHAEL[2] *(Multilingual Inventory of Cultural Heritage in Europe)*, lancé en 2004, a pris en compte la question des langues en élaborant un guide européen multilingue de collections du patrimoine culturel numérisé. Initié par la France, l'Italie et le Royaume-Uni, MICHAEL s'est progressivement étendu à de nombreux autres pays en Europe ; aujourd'hui, 21 pays mettent en commun leurs inventaires, qui sont intégrés dans un portail européen accessible dans 12 langues.

Dans chaque pays, les partenaires de MICHAEL développent des réseaux de partenariat et mettent en place une organisation spécifique afin d'inciter les institutions culturelles à contribuer au catalogue national. L'instance française de MICHAEL, *Patrimoine numérique*[3], présente ses contenus statiques et les titres de ses collections dans trois langues.

En 2007, une association internationale sans but lucratif de droit belge, l'association Michael Culture, a été créée afin de pérenniser le projet, de favoriser le multilinguisme et de fédérer les initiatives nationales.

La problématique du multilinguisme s'est posée de manière encore plus évidente lors du lancement de la Bibliothèque numérique européenne, *Europeana*. Service d'accès aux ressources numériques et numérisées des musées, des bibliothèques, des archives et des collections audiovisuelles européennes, *Europeana*[4] compte en novembre 2010 plus de 14 millions d'objets numériques et s'appuie sur les contributions de plus de 1 500 institutions de toute envergure dans toute l'Europe. Les contenus accessibles dans *Europeana* représentent ainsi une grande diversité linguistique.

De nombreux projets soutenus par la Commission européenne visent non seulement à enrichir *Europeana* mais aussi à définir des méthodologies et des outils normalisés pour appréhender au mieux les spécificités métier des institutions et le caractère plurilingue de ces contenus. Le projet Athena[5], notamment, tend à renforcer et encourager la participation des musées et d'autres institutions patrimoniales qui ne sont pas encore impliqués dans *Europeana*. Un lot de travail dirigé par l'association Michael Culture est consacré au multilinguisme et aux terminologies, en vue de transmettre des recommandations et de bonnes pratiques aux institutions muséales.

Parallèlement à ces projets orientés « contenus », le projet Europeana Connect[6] se concentre davantage sur les utilisateurs, en développant des fonctionnalités de navigation et de recherche multilingues.

Thésaurus et normes : état des lieux et réflexions

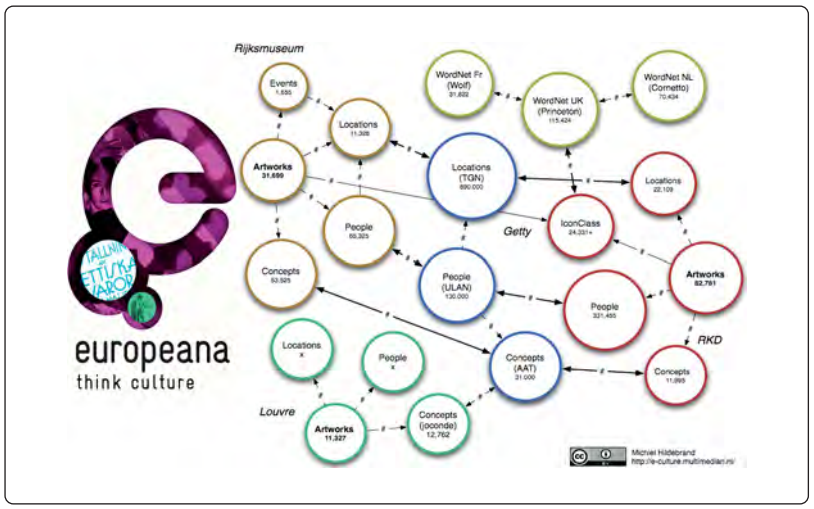
La terminologie recouvre un ensemble de pratiques et de ressources très variées. Dans le domaine d'activité des institutions culturelles, ce terme renvoie aux instruments de recherche et aux outils d'indexation utilisés pour le catalogage.

L'utilisation de vocabulaires contrôlés ou non contrôlés est inhérente à la gestion des objets ou des collections dans une institution culturelle. Une étude[7] menée dans le cadre du projet Athena a montré que ces vocabulaires étaient extrêmement variés tant du point de vue de leur forme que de leur contenu. Lexiques, classifications, thésaurus ou ontologies sont autant de types de ressources différentes utilisées par les musées européens. Cette étude a aussi montré que les institutions utilisent plus volontiers des ressources créées en interne et essentiellement des thésaurus monolingues.

Les thésaurus ont la particularité de distinguer les termes selon deux catégories : les descripteurs et les non-descripteurs. Les descripteurs sont les termes principaux et les non-descripteurs sont les termes liés aux descripteurs par des relations sémantiques, hiérarchiques ou associatives. La structuration des termes constitue la caractéristique la plus importante et aussi la plus appréciée des thésaurus.

Plusieurs normes définies par l'ISO (Organisation internationale de standardisation) régissent l'élabora-

1. Ministerial Network for Valorising Digitisation Activities : **www.minervaeurope.org**
2. Michael : **www.michael-culture.org**
3. Patrimoine numérique : **www.numerique.culture.fr**
4. Europeana : **www.europeana.eu**
5. Athena : **www.athenaeurope.org**
6. Europeana Connect : **www.europeanaconnect.eu**
7. Athena Livrable D4.1 *Identification of existing terminology resources in museums* (Identification des ressources terminologiques existantes dans les musées)



Nuage de ressources *(datacloud)* d'*Europeana*, par Michiel Hildebrand. **http://e-culture.multimedian.nl**

tion des thésaurus. La norme ISO 2788:1986, *Guidelines for the establishment and development of monolingual thesauri* fournit des recommandations pour le développement de bonnes pratiques en matière d'indexation pour l'élaboration d'un thésaurus monolingue. La norme ISO 5964:1985, *Guidelines for the establishment and development of multilingual thesauri* étend le spectre de la norme ISO 2788 en se concentrant sur les thésaurus multilingues.

Parallèlement à ces normes, des standards ont été développés et proposés par le W3C (Word Wide Web Consortium), ces dernières années, afin de prendre en compte la gestion informatisée des terminologies et l'émergence des technologies du Web sémantique. Cette notion introduite par le créateur du Web, Tim Berners-Lee, représente une évolution du Web actuel, Web de documents, vers un Web de données où chaque document serait formalisé de sorte que les technologies du Web sémantique pourraient interpréter les liens entre ces données et donc entre ces documents.

Le Linked Data, littéralement « données liées » s'apparente à une mise en œuvre du Web sémantique par le biais de bonnes pratiques, de méthodes et d'outils normalisés qui ont vocation à lier les données du Web. Pour faire partie du diagramme de ressources généralement utilisé pour représenter le Linked Data, une ressource se doit d'être formalisée dans le format RDF *(Resource Description Format*[8]*)* et chacun des éléments qui composent cette ressource (concepts, personnes, objets, etc.) doit être identifié de façon unique par une URI *(Uniform Ressource Identifier)*. Cette URI permet ensuite de connecter ces éléments les uns aux autres pour constituer le Linked Data et adopter une approche plus sémantique que linéaire et donc plus « données » que « document ».

En aout 2009, SKOS *(Simplified Knowledge Organisation System)*, un format de modélisation des terminologies, a notamment été reconnu comme standard par le W3C. Ce standard qui s'appuie sur les caractéristiques principales des thésaurus constitue pour une institution le format de transition idéal depuis le thésaurus « classique » vers une ressource normalisée et structurée partie intégrante du Web sémantique. La transformation de leur terminologie en SKOS représente un effort et un investissement importants pour les institutions culturelles qui peuvent ainsi adopter une démarche normalisée pour une mise à disposition de leurs contenus en dehors de leurs murs, et donc favoriser la compréhension de leurs contenus dans un environnement multilingue tel qu'*Europeana*.

Ce format SKOS, outre la structuration sémantique des concepts et des termes, permet de modéliser assez simplement et pourtant assez précisément les équivalences entre termes au sein d'une même terminologie (équivalences sémantiques) mais aussi les équivalences entre concepts de terminologies différentes (équivalences linguistiques). Cette gestion ontologique des thésaurus permet de dépasser l'usage premier du thesaurus, qui se réduit à la seule indexation, pour converger vers des usages plus orientés « Web sémantique » tels que la recherche multilingue sémantique. Ce type de recherche permettrait de structurer les résultats d'une recherche mais aussi de l'élargir ou de la préciser : selon les souhaits de l'utilisateur, une requête « Mona Lisa » permettrait d'aboutir aux éléments ayant pour indexation « la Joconde » et ayant comme auteur « Léonard de Vinci » ou encore « Leonardo Da Vinci ».

Des ressources majeures telles que Rameau[9] (Répertoire d'autorité-matière encyclopédique et alphabétique unifié) géré par la BNF, ou le thésaurus Eurovoc[10], terminologie de référence de l'Union européenne, ont franchi cette étape de « SKOSification », c'est-à-dire de conversion en SKOS de leurs ressources. Un second rapport[11] comportant des lignes directrices pour la SKOSification des terminologies a été produit dans le cadre du projet Athena, afin de guider les institutions dans ce processus.

Ces normes et standards visent à harmoniser les ressources linguistiques et terminologiques et ne constituent qu'un premier pas vers le Linked Data[12] qui permettra de relier l'ensemble des ressources disponibles sur le Web les unes aux autres. ■

8. RDF : **www.w3.org/TR/rdf-primer**

9. Langage d'indexation matière RAMEAU : **http://rameau.bnf.fr**

10. EUROVOC : **http://eurovoc.europa.eu**

11. Athena Livrable D4.2 : *Guidelines for mapping into SKOS, dealing with translations*

12. Linked Open Data : **http://linkeddata.org**