Exigences des métadonnées

Qu'une donnée puisse en représenter un jeu entier, qu'à son tour ce dernier puisse constituer un corpus cohérent de connaissances, ordonné et exploitable – cela identifie assez, depuis Alexandrie, la nécessité classificatoire des objets de pensée, leur invention[1] et leur mémoire.

Les métadonnées, puisqu'il s'agit bien de ces données *indexantes* du discours, ont connu une fortune invariable de renouvellement, que leur pouvoir d'orientation puis d'interprétation n'a cessé d'affirmer. Elles imposent encore aujourd'hui un maillage des contenus de la science – dans la mesure extensive de leur exposition numérique.
C'est justement sous le couvert empirique d'un univers insensé – le web de données – que la production des métadonnées trouve sa justification tant heuristique qu'économique.

Métadonnées de propriété

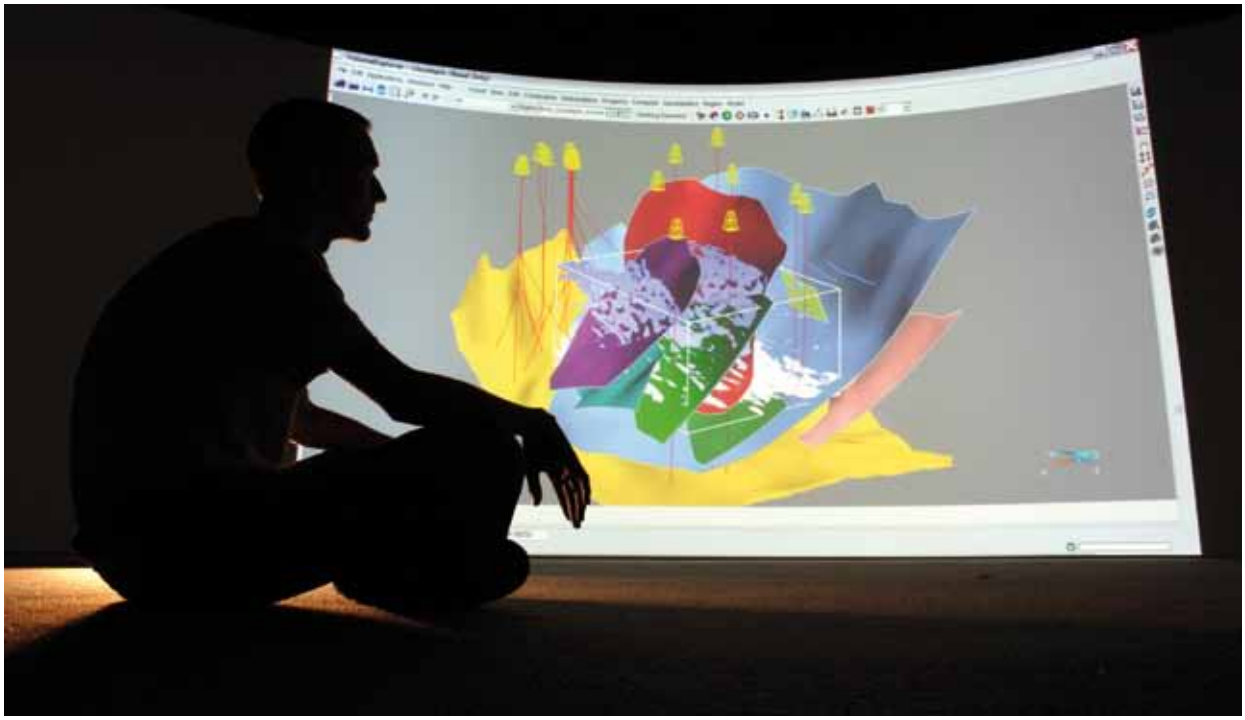
Tout objet de discours (de science), toute collection de données, en définitive toute unité de sens peut recevoir une annotation minimale, une cote, un identifiant qui renvoie à un système partagé – un référentiel qui a valeur de norme et contre lequel ils sont rapportés, puis reconnus. C'est à ce premier niveau de propriété qu'une métadonnée distribue les objets comme autant d'objets uniques.
Historiquement, de tels identifiants normés ont été attribués et maintenus par des communautés d'intérêts (intellectuelles ou commerciales) exerçant un droit de propriété sur les contenus visés. Une autorité auteur par exemple, au sens bibliothéconomique du terme, peut étendre sa représentation dans une définition stricte, portée par un identifiant numérique unique. L'initiative *Orcid*[2] constitue à ce titre l'effort sans doute le plus significatif à l'établissement d'un registre mondial de contributeurs scientifiques, y compris d'œuvres orphelines ou posthumes. L'enjeu que représente l'établissement d'une telle métadonnée intéresse directement la propriété intellectuelle et les ayants droit d'une œuvre de l'esprit – tout en permettant de lier les objets qu'ils ordonnent à l'échelle d'une e-science dont les contributions connaissent un développement exponentiel.
Autant l'ISSN constitue une métadonnée de propriété captive – et de même le DOI qui demeure une attribution éditoriale –, autant la nécessité de métadonnées de propriété d'un autre ordre, visant tous les objets de la recherche scientifique, apparaît comme un enjeu aux nombreuses ramifications. Une extension aux données brutes de la science de ces identifiants spécifiques est notamment portée par le consortium *DataCite*[3], dès lors accessibles à la citation comme aux liaisons sémantiques.

Métadonnées descriptives

C'est une société d'informatique dans les domaines de la santé et de la médecine qui dépose la marque « Metadata » en 1986. La NASA l'utilise pour la première fois en 1988 dans une édition du *NASA's Directory Interchange Format Manual*[4]. Dans les années 1990, le vocable entre peu à peu dans le domaine public et s'impose comme définissant la description des données. En 1995, après qu'une première spécification ait été réalisée pour les données géographiques, le terme apparaît dans le monde bibliothéconomique avec la création du *Dublin Core Metadata Set Element*. Les professionnels de l'information prennent alors conscience qu'ils sont créateurs, depuis des décennies, de nombreuses métadonnées et que celles-ci peuvent tout autant décrire des documents papier que des ressources électroniques.
C'est ainsi que les données catalographiques, depuis leur socle descriptif MARC (*Machine Readable Cataloguing*), n'ont cessé de décliner de nouveaux modèles de représentation. Aujourd'hui, le modèle conceptuel FRBR (*Functional Requirements for Bibliographic Record*) entend formaliser les relations sémantiques entre une œuvre et ses différents types d'expressions, en s'affranchissant de la notion de « document » comme première unité signifiante. Les enjeux de ce modèle verront se démultiplier de nouveaux points d'accès à l'information et permettront de nouvelles liaisons sémantiques entre des sources de données hétérogènes.
Ces nouveaux principes, organisés autour d'entités et de relations, offriront un cadre de données structurées, appuyé sur de nombreux schémas de métadonnées XML – formats créés en fonction de l'usage des ressources[5]. Nous voici à l'aube de la construction du web de données, où il ne s'agit plus seulement d'assurer des liens entre des objets ou des documents, mais d'offrir un accès logique et pérenne entre les données elles-mêmes. L'accès à la connaissance suppose d'ores et déjà l'appropriation de ces nouvelles cartes du savoir, affranchi de ses limites propres.

Métadonnées d'enrichissement

Pour porter le paradigme documentaire au sens, le seul signalement des ressources, fût-il normé et extensif, ne saurait servir qu'à un seul effet d'accumulation de données. C'est la structuration des métadonnées dans de véritables systèmes d'information, puis leurs liaisons avec d'autres univers de données, leur interopérabilité, qui permettront leur ouverture tant aux moteurs de recherche qu'aux routines des bases de connaissances. Toutes opérations qui requièrent un apport d'intelligence : le coût du facteur humain dans l'analyse et l'enrichissement des données oblige à des contournements ou à la mise en œuvre de relais sémantiques qui aient quelque valeur probante dans les différents champs de l'économie numérique.



© Inria/Photo C. Lebedinsky

Une première approche tient dans l'imposition de règles de production des médias d'information eux-mêmes, à des fins d'extraction automatique et contrôlée de métadonnées utiles : une telle disposition éditoriale augmente la performance des systèmes de fouille de données et des logiciels d'appariement lexicaux associés. Courante dans la chaîne éditoriale du document, de l'image ou du son (œuvre ou fragment), elle reste à généraliser aux objets complexes de la culture et de la connaissance (objets de la recherche scientifique, flux de données primaires, bruits des différents fonds de données archivées). Des ontologies y répondent, mais la méthode semble introuvable[6], tant l'expression du sens d'un objet (textuel, multimédia, sériel…) semble hors de la portée des outils qui le manipulent (quand ceux-ci seraient fondés sur des recommandations du web sémantique). En clair, un jeu de métadonnées ontologiques, qui expose un contenu à l'interprétation d'agents logiciels, ne saurait épuiser le sens qu'il porte, sauf à le rapporter à un segment de science (une discipline captée par son lexique) ou de culture (une œuvre représentée par ses codes) – c'est-à-dire à des référentiels fermés, sans prétention herméneutique (interprétative)[7].

Mais à défaut de sens, ces métadonnées référentielles peuvent opérer une aire de correspondances entre différents univers de données. Des objets liés à des référentiels hétérogènes peuvent entrer en résonance, sans corrompre leurs lexiques propres. Des initiatives comme SKOS cherchent les conditions d'interopérabilité de ces différents référentiels auxquelles de grands prescripteurs institutionnels apportent leur soutien[8].

Une exposition au plus grand nombre

Un des enjeux d'exploitation de la variété des gisements informationnels repose sur leur facilité d'exposition (*Open Data*[9]) et d'intégration dans des systèmes tiers, au-delà de leurs éventuelles incompatibilités natives pouvant résulter de leur structuration physique ou sémantique.

À titre d'exemple, la complexité de la description de l'objet muséal ou archéologique, dans toute la variété de ses formes, a conduit à la création de modèles, tel CRM-CIDOC[10] devenu norme ISO en 2006, et à plusieurs standards pour les décrire tels CDWA (*Categories for Description of Work of Art*)[11] et LIDO[12].

CRM-CIDOC, ontologie de domaine orientée objet, entend modéliser et formaliser la nature et les relations existant entre divers éléments descriptifs des objets de musée, qu'ils soient unitaires ou partie d'un tout, réels ou virtuels. Ce modèle s'accorde à donner le sens précis des métadonnées ainsi que des événements entrant en relation avec l'objet décrit. Ces descriptions fortement contextualisées permettent de préciser l'unicité de l'item (en restituant ses identifiants, ses marques, ses mesures…) tout en offrant des liens d'adjacence avec des ressources externes qui qualifient son environnement (géographique, historique, manifestations en rapport, intervenants…).

Appliqué à divers corpus documentaires, tels qu'images, textes, vidéos, CRM-CIDOC peut offrir de nouveaux espaces de médiation fédérant des ressources variées et favorisant ainsi la découverte, au-delà des frontières disciplinaires ou institutionnelles. L'initiative CLAROS (*CLassical Art Research Online Service*)[13], programme soutenu par l'UNESCO, réunit plus de 20 millions de notices d'objets de l'Antiquité gréco-romaine issus des musées du monde entier, grâce à l'utilisation du CRM-CIDOC et de RDF, le tout interrogeable *via* le langage de requête SPARQL.

La création de ces nouveaux points d'accès permet une plus vaste exposition des métadonnées, véritable enjeu de diversité culturelle et d'aide à la recherche scientifique, rendu possible grâce aux langages et vocabulaires sémantiques tels que RDF, SKOS…

Dans un contexte plus commercial, les métadonnées d'identification, telles ISAN (*International Standard Audiovisual Number*), ISRC (*International Standard Recording Code*) et ISMN (*International Standard Music Number*), prennent tout leur sens dans un contexte de reproductibilité des supports, de quantité de transactions et d'échanges dans le cadre mondial du e-commerce. En tant qu'identifiants uniques, ces formats répondent à une attente très forte de la part des ayants droit, des gestionnaires de col-

lections, des diffuseurs ou revendeurs de contenus. En général, ils sont attribués par une autorité fédérative gérant un répertoire central.

Et demain ?

Ces métadonnées, embarquées dans de nouveaux médias, utilisées *via* des dispositifs ubiquitaires (extensifs et mobiles), se proposeront demain spontanément aux utilisateurs que nous sommes, en fonction de nos usages passés – appréciations, prescriptions et consommations. Alors qu'à ce jour, l'usager numérique se devait d'être actif pour accéder à l'information, celle-ci pourrait être dotée d'une forme d'intelligence afin d'anticiper les besoins informationnels, préparant ainsi le web de demain : l'Internet des objets[14].

Jacques Millet
Gaëlle Rivérieux

Inria, établissement public de recherche dédié aux sciences du numérique http://www.inria.fr
Jacques Millet est délégué à l'information scientifique de l'Inria et membre du Comité scientifique de l'ABES
jacques.millet@inria.fr
Gaëlle Rivérieux est responsable IST du Centre Inria Grenoble Rhône-Alpes gaelle.riverieux@inria.fr

[1] *Inventio* – au sens latin de découverte.
[2] *Open Researcher and Contributor ID* – proposé comme identifiant unique et ouvert pour les contributeurs d'œuvres de l'esprit, similaire à ce que pourrait être un DOI [*Digital Object Identifier*] d'auteur (cf. http://www.orcid.org). Voir également, sur un périmètre élargi des œuvres de création, l'ISNI [*International Standard Name Identifier*] qui ambitionne l'interopérabilité de systèmes d'identification propriétaires.
[3] Datacite.org – et son *Metadata Store* : https://mds.datacite.org/
[4] Caplan (P.), *Metadata for all librarians*, Chicago, American Library Association, 2003, p. 1.
[5] Dublin Core, EAD, MODS, METS...
[6] C'est cependant l'horizon que se donne RDF [*Resource Description Framework*], avec l'établissement de représentations sémantiques des métadonnées descriptives des objets de recherche et de culture – sans le filtre d'une lecture humaine.
[7] Cf. Zacklad (M.), « Évaluation des systèmes d'organisation des connaissances », *Les Cahiers du numérique*, 2010, vol. 6, n° 3, p. 133-166.
[8] Bibliothèques nationales, ministères de la Culture, registres de publications officielles. Voir http://www.w3.org/2001/sw/wiki/SKOS/Datasets
[9] http://www.data.gouv.fr
[10] http://www.cidoc-crm.org
[11] Développé au début des années 1990 : http://www.getty.edu/research/publications/electronic_publications/cdwa/index.html
[12] http://www.lido-schema.org/documents/LIDO-Handout.pdf (2010).
[13] http://www.clarosnet.org
[14] Mitton (N.), Simplot-Ryl (D.), *From the Internet of things to the Internet of physical world*, http://hal.archives-ouvertes.fr/hal-00598395

Les métadonnées, ces grandes indiscrètes

Chacun a encore en mémoire l'affiche de campagne de Nicolas Sarkozy pour les dernières élections présidentielles : le portrait du candidat avec le slogan « Une France forte » au premier plan, une mer calme et un ciel lumineux, au second.
Patrick Peccatte, chercheur associé au Laboratoire d'histoire visuelle contemporaine (Lhivic/EHESS), dans un billet publié le 17 février 2012 sur Culture visuelle, explique comment un internaute a réussi à établir que la photo du paysage marin de l'affiche n'était autre qu'un cliché de la mer Égée en Grèce.
C'est la diffusion de l'affiche au format numérique qui a permis de remonter ainsi jusqu'à la prise de vue initiale.
L'auteur du billet rappelle les 3 types de métadonnées stockées dans un fichier image : **les informations IPTC/IMM** (*Information Interchange Model*), standard adopté en 1994 par Adobe pour définir les informations associées à une image et dont les champs sont renseignés manuellement lors de son indexation ; **les données EXIF** (*EXchangeable Image File*), informations d'ordre technique correspondant à un standard supporté pas tous les fabricants d'appareils numériques, renseignées automatiquement lors de la prise de vue ; **les données XMP** (*eXtensible Metadata Platform*), créées par Adobe en 2001, version simplifiée de RDF (*Resource Description Framework*), modélisation de base du web sémantique.
Le croisement de ces trois types de métadonnées a permis de retrouver l'agence qui a commercialisé la photo (Tetra Images) et d'établir que cette dernière a été prise en Grèce.
Des investigations un peu plus approfondies fournissent la date de prise de vue du cliché initial (28 mai 2011) et l'examen précis des données XMP de l'affiche numérique dévoile que la photo a subi 38 modifications du 12 au 14 février 2012.
La révélation de ces investigations sur les métadonnées de l'affiche a bien évidemment donné lieu à un *buzz* sur Internet pointant la maladresse du slogan « une France forte » associé à une photo de la Grèce, pays dont on connaît la situation économique actuelle. Un *buzz* dont les concepteurs de l'affiche se seraient sans doute fort bien passé mais qui prouve que désormais les communicants devront apprendre à se méfier des indiscrétions des métadonnées.

Béatrice Pedot

Pour en savoir plus : http://culturevisuelle.org/dejavu/1118