arbido

2017/3 Métadonnées – Données de qualité

Morisod Pascal, historien et informaticien, spécialiste I+D

Données et métadonnées, aux sources de l'information numérique

Dans le sillage de la révolution numérique, les données et les métadonnées ont envahi nos paysages. Au prix d'une polyphonie technoïde et parfois névrotique, l'écosystème de la *cyberinformation* nous promet mille progrès, gérés par autant de métadonnées.

L'évolution des TIC et des métiers I+D sont technoïdes, car marqués par une tendance générale à l'accélération et à la multiplication, des beats comme des bits; ils sont névrotiques car ils se doivent d'absorber le choc des générations, la galaxie des nouveaux métiers, la multiplication des partenaires; ils confrontent le monolithique, du patrimoine séculaire à l'infobésité des services administratifs, au mouvement quasi perpétuel. Last but not least, *it's a game without frontiers* dans un monde globalisé, où les anglicismes font rage.

Un monde polymorphe

Nonobstant, les sciences de l'information sont un domaine fécond, dynamique et productif. De la gestion des documents patrimoniaux ou à valeurs probatoires, à ceux administratifs tout au long de leur cycle de vie, les enjeux sont considérables pour les métiers I+D. À condition d'éviter certains pièges et de franchir autant d'écueils avec succès, ce nouvel écosystème est nécessairement porteur de progrès. Stratégies de numérisation à des fins d'optimisation, traitement et gestion des informations en masse, diffusion et médiation numériques pour un public renouvelé ou nouvellement formé: c'est dans cette polymorphie ambiante que les (méta)données évoluent.

Clin d'oeil à Magritte. Dessin assisté par ordinateur, Pascal Morisod.

Une superposition d'états contradictoires

Un peu de philosophie : le préfixe «méta-» renvoie à un niveau d'abstraction supérieur ou à la réflexion et l'autoréférence, ce qui appelle certaines réticences lorsqu'il s'agit de la capacité d'un sujet à se référer à lui-même. La métaphysique serait soit la *«partie fondamentale de la réflexion philosophique qui porte sur la recherche des causes, des premiers principes»* (sens mélioratif, tiré du CNRTL), soit un *«abus de considérations abstraites, qui, au lieu d'éclairer la pensée, ne font que l'obscurcir»* (sens péjoratif, tiré du CNRTL). Par extension, les interrogations jaillissent : quelle est la valeur d'une métadonnée, est-elle nécessaire à la donnée, qu'apporte-elle de plus? ou est-elle redondante et peu rentable? à quels types de données s'applique-t-elle prioritairement et comment est-elle générée? Évidemment, il est impossible d'aborder les métadonnées sans définir les données, comme il est impossible de discourir sur la nature des choses sans en connaître ses lois.

L'une d'entre elles vient de la métaphysique déduite de la physique quantique, qui nous installe face à une superposition d'états contradictoires[1]. Cette approche, en tant que grille de compréhension de l'évolution des métiers I+D, paraît pertinente à la vue des nombreux paradoxes que nous pouvons y déceler, dans un univers des possibles qui semble sans limites et qui trouve son revers dans l'augmentation de l'incertitude et le technostress. Néanmoins, à l'ère du *benchmarking*, nombreux sont les garde-fous, puisque tout est comptable et mesurable à souhait. *Que nul n'entre ici s'il n'est géomètre*: la formule (soi-disant) gravée à l'entrée de l'Académie de Platon est toujours d'actualité!

Définitions

Il est a priori simple de donner des exemples de données à propos de données : les coordonnées topographiques d'une photo, la durée d'une séquence audiovisuelle, le titre ou l'auteur d'une œuvre ou d'un document d'archives. Néanmoins, quelques exemples ne suffisent pas à circonscrire l'entier du concept. Les métadonnées sont présentes dans les systèmes informatiques et le Web sémantique pour l'interconnectivité d'un côté, dans les métiers I+D de l'autre. Elles permettent de faciliter l'accès aux ressources et d'opérer dessus. Au niveau I+D, elles servent plus spécifiquement à décrire et contextualiser des documents ou des données afin de pouvoir en disposer de manière efficiente. Entre les deux domaines, il y a des convergences, des superpositions et parfois des frottements, dans une sorte de tectonique des plaques.

Données

Au niveau informatique, les données sont *«l'ensemble des indications enregistrées en machine pour permettre l'analyse et/ou la recherche automatique des informations»*. En ce sens, tout document doit être numérique ou numérisé, ou encore bénéficier d'une extraction de contenu pour devenir «donnée», i.e. manipulable par un ordinateur. Par extension, il en va de même pour les métadonnées, qui peuvent porter sur des documents ou des objets assimilés à des données à partir du moment où ils acquièrent une existence numérique, en procédant par exemple à une cotation.

Métadonnées

Le terme «métadonnée» est un anglicisme qui s'est largement répandu à partir des années 1990 avec l'apparition du Web sémantique. Il s'est également diffusé au sein des systèmes informatiques, en convergeant avec les métadonnées archivistiques ou bibliothéconomiques, deux domaines où, depuis des lustres, l'information était structurée sous diverses formes analogues aux métadonnées.

En tant qu'informations (semi-)structurées, les métadonnées doivent répondre à un ensemble de critères qualitatifs comme l'authenticité, la fiabilité, l'exploitabilité, l'intégrité, etc. Elles sont en principe regroupées dans des ensembles classés par typologies. Les anglo-saxons en distinguent trois types: descriptives, structurelles, administratives[2]. Des normes, ISO ou pas, permettent de structurer l'information suivant un domaine donné, comme l'information et la documentation traité par le groupe de travail ISO TC 46. Afin de ne pas s'égarer dans la jungle des normes, des sources secondaires sont souvent indispensables[3].

Métadonnées archivistiques

ISAD(G) fournit des lignes directrices pour la description archivistique, indépendamment de la forme ou du support des documents d'archives. Le modèle définit 26 éléments de descriptions répartis dans 7 zones d'information comme l'identification, le contexte, le contenu, ou encore les conditions d'accès. La norme inclut un modèle hiérarchique de description à plusieurs niveaux, avec un ensemble de règles à appliquer comme la description du général au particulier et la non-redondance. Ce modèle, intégré à un système informatique, peut faire l'objet d'une interface en forme d'arborescence en accordéon, comme dans la plupart des inventaires en ligne des institutions cantonales d'archives (inventaires DAVEL). Plus généralement, les concepts de non-redondance et de récupération de l'information ou des processus préenregistrés sont primordiaux en informatique. Citons les notions d'instanciation d'objets et d'héritage en programmation orientée objet, l'utilisation de bibliothèques préprogrammées pour les interfaçages Web, les bases de données relationnelles (SGBDR).

Métadonnées et records management

La norme propriétaire ISO 23081[4] est consacrée aux métadonnées qui accompagnent les processus de gestion du *records management* (ISO 15489). Elle s'applique aux enregistrements, aux processus qui les affectent, au système dans lequel ils se trouvent, aux organismes responsables de leur gestion. Les métadonnées, en tant que *«données qui décrivent le contexte, le contenu et la structure des documents d'activité, ainsi que leur maintenance dans le temps»*, permettent la création, l'enregistrement, la classification, l'accès, la conservation ainsi que le choix du sort final des documents dans le temps. Les métadonnées assurent l'authenticité, l'exploitabilité et l'intégrité des objets d'informations, qu'ils soient analogiques ou numériques, et en permettent la gestion et la compréhension. En terme de management des métadonnées, des rôles et des responsabilités sont définis; au niveau des systèmes, des prescriptions sur leur structure et leur rôle sont données. Enfin, la norme propose une typologie des métadonnées ainsi qu'un ensemble de règles et de procédures concernant aussi bien les documents (processus de gestion) que les acteurs (processus métier).

Normes et réalité...

Si l’ensemble des prescriptions peut paraître abscons et leur application représenter des difficultés qu’une collaboration entre professionnels I+D et ingénieurs-informaticiens se doit d’atténuer, un autre obstacle résulte dans l’identification des besoins réels d’une institution, sachant qu’une norme peut vite être un océan sur lequel on navigue à vue, embarqué dans une galère. Avant tout, les professionnels de l’information agissent avec ces compendiums comme les pilotes d’avion avec leurs check-lists: dans un système complexe, ils les aident à ne rien oublier. Dans les deux cas, un des enjeux majeurs est l’adéquation entre la théorie et la réalité et comme souvent dans le monde numérique fortement globalisé, seule une partie des ressources théoriques à disposition sera véritablement utile au sein d’une institution.

Document

Par rapport aux données et aux métadonnées, qu'en est-il du document? En archivistique, il existe plusieurs normes et lois qui le définissent, comme ISAD(G) et sa variante suisse, ou encore la loi suisse sur l’archivage (article 3) et la loi jurassienne, qui définit également le document actif, semi-actif et inactif (article 4). S’il apparaît important, à la lumière des définitions, que le document soit irréductible à son support, la loi fédérale précise encore que les (bases de) données en font partie, tout comme les métadonnées, comprises dans les données complémentaires. Ces définitions visent à tenir en lisière de l’archivistique les questions liées au monde numérique, en maintenant une forme de souveraineté corporative. Ainsi, le concept de document, pour un registre de notaire, englobe aussi bien le registre en tant qu’objet matériel, sa version électronique, et l’ensemble des métadonnées qui servent à le définir et le contextualiser. Le sens étymologique d’acte écrit ou déposé sur un support est conservé, les photos étant des écritures de lumière et les records (le vocable provient du vieux français!) des témoignages écrits.

Deux projets des archives jurassiennes

En mettant en ligne les enregistrements audionumériques des séances de l’Assemblée constituante (1976-1978) et des trois premières législatures du Parlement du canton du Jura (1979-1990) (projet AGORA), les Archives cantonales jurassiennes, en partenariat avec Memoriav et Wikimedia CH, ont dû réfléchir aux métadonnées nécessaires à la recherche d'extraits sonores et à l'interopérabilité des systèmes (inventaires des archives avec scopeQuery et espace de stockage des fichiers sur Wikimedia Commons).

Un autre projet des Archives cantonales jurassiennes est intéressant concernant l'utilisation des métadonnées. Une collection de plus de 3000 photographies du facteur-photographe jurassien Eugène Cattin a été numérisée et téléversée dans Wikimedia Commons sous la forme de fichiers TIF, accompagnés de métadonnées archivistiques. Grâce à l'accès libre (*open data*) dans Commons, l'ajout de métadonnées par les internautes est à présent continuel, tout Wikipédien ayant la possibilité d'indexer les photos en y ajoutant des informations en tous genres (noms des personnes figurant sur les photos, objets spécifiques, lieu précis de la prise de vue).

Remarquons le côté novateur de cette gigantesque base de données multimédia qu'est Wikimedia Commons: l'indexation se fait autant par des personnes que par des «bots», des robots programmés par des Wikipédiens avertis, qui scannent le contenu du site afin de reconnaître des objets spécifiques représentés sur les photos, ce qui permet un catalogage automatique (lire aussi l'article sur structurer Wikimedia Commons avec Wikidata). En tout, la production de métadonnées va de l'archiviste à l'internaute féru d'histoire local, de photographies ou de n'importe quel sujet traité, en passant par la production de métadonnées automatisées grâce à de puissants algorithmes.

Périodiques en ligne

La bibliothèque numérique e-periodica héberge et présente l'ensemble des Actes de la société jurassienne d'émulation, de 1879 à 2015. Le projet a été mené par la Bibliothèque cantonale jurassienne. Les actes ont été numérisés et océrisés à l'aide du procédé de reconnaissance optique de caractères (ROC ou OCR en anglais). Cela signifie que les recherches plein-texte sont possibles et qu'en conséquence, chaque terme, tronqué ou non, peut être repéré en une seule requête au sein de toute la collection. La bibliothèque numérique présente une interface utilisateur où deux ensembles de métadonnées ont été extraits afin d'encadrer la navigation au sein d'une somme considérable d'informations: la table des matières et les titres des articles. Les requêtes peuvent être élargies à toutes les revues de la bibliothèque, ce qui offre de nouvelles perspectives d'analyses, particulièrement du côté du *big data*.

Une terre commune et un new deal

Les métadonnées sont omniprésentes au sein des nouvelles pratiques informationnelles, lesquelles induisent de nombreux changements de paradigmes. Au final, elles servent à cultiver une terre commune à l'aide de puissantes machines, générant un contenu facilement accessible au plus grand nombre. Au sein d'un écosystème I+D qui se doit d'être aussi stable que novateur, cette évolution, pour reprendre la conclusion du récent ouvrage *Archives en Suisse*, appelle à formuler un *new deal*[5]. Avec la mondialisation, la multiplication et la complexification des canaux d'informations, le potentiel prospectif des archives ou leurs usages originaux - en art, par exemple -, les mutations sont plurielles. Il en va de même pour les humains, où collaborations et complémentarités des savoirs deviennent incontournables, dans un monde où toute *«phase de transformation technologique possède ses pionniers, ses 'bousculés', ses transfuges, mais aussi, malheureusement, ses laissés-pour-compte»* [6]. Dans ce milieu se croisent les classiques, les mutants et les novateurs [7]. Ne rien perdre du savoir des anciens, tout en adoptant les technologies du troisième millénaire, non sans tempérance, afin que les publics et les professionnels I+D en retirent les bénéfices escomptés: tel me paraît être un bon début.

1 Voir à ce propos «L'histoire en chute libre», Alexandre Lacroix, Philosophie magazine 108, avril 2017, pp. 54-58, ainsi que le dossier complet consacré à l'incertitude, pp. 54-71.

2 Voir Wikipédia, article «metadata».

3 Soulignons à ce propos que le groupe de travail "normes et standards" de l'AAS fournit un catalogue raisonné des normes archivistiques, tandis que la société docuteam publie un grand nombre de ressources, notamment dans le cadre d'un Système ouvert d'archivage de l'information (SOAI/OAIS).

4 Voir présentation: partie 1; partie 2; la norme est payante, et donc pas accessible librement en ligne.

5 Gilbert Coutaz, *Archives en Suisse, conserver la mémoire à l'ère numérique*, Lausanne: PPUR, 2017, pp. 88-112.

6 Tiré de D. Cotte,«Les métiers, entre traditions et modernité», in *Documentaliste, sciences de l'information - Les métiers de l'information et la donnée*, numéro 3, Septembre 2013, pp. 42-46.

7 *Ibidem.*

Pascal Morisod

Après une formation en Histoire et Informatique et méthodes mathématiques à l'Université de Lausanne, Pascal Morisod rejoint la jurassique et jurassienne Paléontologie A16 et l'équipe du Système d'information documentaire, en charge de la gestion du patrimoine informationnel et physique issu des creusements de la Transjurane (2009-2015). Il occupe ensuite le poste d'Archiviste cantonal adjoint du canton du Jura, en charge des projets «nouvelles technologies» (2016-2017). Depuis 2018, il est archiviste et référent TIC aux Archives cantonales vaudoises.

Résumé

Français

Dans le sillage de la révolution numérique, les données et les métadonnées ont envahi nos paysages. Au prix d'une polyphonie technoïde et parfois névrotique, l'écosystème de la *cyberinformation* nous promet mille progrès gérés par autant de métadonnées. Mais que sont-elles? Est-ce que la métadonnée est nécessaire à la donnée et que lui apporte-t-elle de plus? Ou est-elle redondante et peu rentable? Enfin à quels types de données s'applique-t-elle prioritairement et comment est-elle générée? Cet article essaie d'aborder ces aspects.

Deutsch

Als Folge der digitalen Revolution haben Daten und Metadaten unsere Cyberlandschaft geradezu überflutet. Das Ökosystem der Cyberinformation verspricht uns grosse Fortschritte, generiert durch eben so viele Metadaten - auf Kosten einer technoiden, manchmal gar neurotischen Polyphonie. Aber sind Metadaten eigentlich wichtig für Daten? Und was bewirken sie? Oder sind sie redundant und wenig rentabel? Und bei welchen Datentypen werden sie vorzugsweise angewendet? Und wie werden sie überhaupt generiert? Auf diese Aspekte versucht dieser Artikel einzugehen.