A propos de la standardisation des données

Laurence Dierickx - ohmybox.info

26 décembre 2018

La standardisation fait partie des bonnes pratiques (Boydens 20014) permettant d'assurer une interopérabilité optimale en vue de la réutilisation des données (Batini 2009, Janssens 2012, Boydens). Elle participe au cycle d'amélioration des données. Elle peut être interprétée en trois niveaux (Boydens, Hulstaert et Van Dromme 2011) : "*(1) comme la résolution des problèmes de qualité des données dus au manque de standardisation (recensés à l'aide d'un Data Profiling ou non) ; (2) comme l'ensemble des actions (manipulations des données) exécutées en préparation à un Data Matching (corrections automatiques, en interne, dans la mémoire...) ; (3) comme un projet de nettoyage en lots destiné à améliorer la qualité des données*".

Dans ses recommandations portant sur la publication de données publiques ouvertes, le W3C préconise un encodage des données en utilisant des normes ouvertes. Si les problèmes d'encodage peuvent aussi relever d'erreurs humaines, cette formalisation permet une homogénéisation des pratiques et donc des données. Standardiser, c'est appliquer la "commune mesure" (Boydens 2012),; c'est disposer de conventions univoques pour une représentation correcte des données. La standardisation porte non seulement sur les formats (ouverts et donc accessibles de manière indépendante d'un logiciel) mais aussi sur les valeurs des données. Boydens (2014) rappelle que les données doivent être documentées pour être compréhensibles, et que les standards doivent aussi s'appliquer aux formats des méta-informations correspondantes. La standardisation, poursuit-elle, passe par la mise en œuvre du web sémantique.

Outre les recommandations relatives aux formats, les standards consistent dans des règles pour un étiquetage uniforme des données : un ensemble de données disposant de la même étiquette aura le même contenu sémantique et syntaxique (Sadiq 2013). Sadiq met toutefois en garde : même lorsque les standards sont observés, les risques de "collision" dans l'étiquetage des données ne sont pas nuls. Pour Madnik et Zhu (2006), la standardisation peut s'avérer infructueuse pour différentes raisons : une hétérogénéité représentationnelle (à propos des attributs), une hétérogénéité ontologique (à propos des entités), et une sémantique susceptible d'évoluer avec le temps.

Normes ISO L'Organisation Internationale de Normalisation (ISO) a pour principale mission de rationnaliser et de promouvoir l'échange de biens et de services par le biais de normes internationales à adhésion volontaire (Grolleau et Mzoughi 2005). Ces normes sont prévues pour un usage général dans les divers domaines scientifiques et techniques. Les différentes normes détaillées ci-dessous sont identifiées comme susceptibles de concerner l'encodage de données

publiques ouvertes [1]. A noter qu'une norme ISO est mise à jour tous les cinq ans.

— **Références bibliographiques (ISO 690 :2010) :** a pour objet de donner des principes directeurs pour la rédaction des références bibliographiques, en organisant un ordre dans les mentions. La norme ne donne aucune consigne de mise en forme des références ou des citations. Elle ne dit rien non plus à propos de la ponctuation, qui doit donc faire l'objet d'une convention interne.

— **Représentation des pays (ISO ISO 3166-3 :2013) :** énonce les principes et procédés de mise à jour d'un code pour la représentation des noms de pays, par exemple FR pour la France. La norme prévoit les subdivisions territoriales, dont la nomenclature comprend le code du pays suivi par une lettre, suivant l'ordre alphabétique : FR-A pour l'Alsace, FR-B pour l'Aquitaine, FR-E pour la Bretagne ou FR-S pour la Picardie. Les subdivisions départementales ont une nomenclature qui débute par le code du pays, suivi par le numéro attribué au département : FR-22 pour Côtes d'Armor, FR-35 pour Ille et Vilaine. Une nomenclature est également prévue pour les DOM-TOM : FR-GP pour la Guadeloupe, FR-MF pour Saint-Martin.

— **Représentation des monnaies (ISO 4217 :2015) :** définit le code de trois lettres attribué aux devises utilisées das le monde : par exemple, EUR pour l'euro et USD pour le dollar.

— **Représentation normalisée de la localisation des points géographiques par coordonnées (ISO 6709 :2008) :** spécifie notamment la représentation des coordonnées, dont la latitude et la longitude, utilisées pour l'échange de données. Cette norme supporte la représentation de la localisation des points au moyen du format XML [2].

— L'ISO propose également la norme ISO 19157 :2013, relative à la qualité des données géographiques. L'organisme argumente l'opportunité de cette norme en raison de données géographiques "*de plus en plus partagées, échangées et utilisées à d'autres fins que celles initialement prévues par leur producteur*". Dans la notice présentant cette norme, l'organisme souligne que "*les informations se rapportant à la qualité des données géographiques disponibles sont essentielles au processus de sélection d'un jeu de données, la valeur des données étant directement liée à leur qualité. (...) La description de la qualité des données géographiques a pour objectif de faciliter la comparaison des jeux de données et la sélection de l'ensemble le mieux adapté aux besoins ou aux exigences de l'application. Une description exhaustive de leur qualité encouragera le partage, l'échange et l'utilisation des jeux de données appropriés*". Par ailleurs, la norme ISO 19101-1 :2014 définit le modèle de référence pour la normalisation dans le domaine de l'information géographique, lequel décrit la notion d'interopérabilité et établit les principes de base

1. Cette sous-section est basée sur la documentation fournie par l'ISO, dont les éléments ne relevant pas de la propriété intellectuelle sont publiés sur le site www.iso.org.

2. *Extensible Markup Language*, métalangage de balisage.

sur lesquels s'appuiera la normalisation. La norme ISO ISO 19115-1 :2014 porte quant à elle sur le schéma des métadonnées de l'information géographique.

— **Représentation de la date et de l'heure (ISO 8601 :2004) :** décrit un mode de représentation numérique de la date et de l'heure accepté à l'échelon international. Elle pour objet de lever l'ambiguïté d'interprétation lorsque les dates sont exprimées en chiffres.

— **Grandeurs et mesures (ISO 80000 1 :2009) :** donne des informations générales et des définitions à propos des grandeurs, des systèmes de grandeurs, des unités, des symboles de grandeurs et d'unités, et des systèmes cohérents d'unités. La norme s'appuie notamment sur le Système international de grandeurs (ISQ) et sur le Système international d'unités (SI). La norme précise également les conventions de rédaction/d'encodage pour les chiffres : les décimaux sont séparés des unités par une virgule pour répondre au conventions internationales. Un point peut être utilisé mais, dans ce cas, il convient d'observer toujours la même convention (point ou virgule). La langue française privilégiera toutefois l'usage de la virgule, celui du point étant davantage répandu dans l'univers anglo-saxon. Il est d'usage, dans la langue française, de séparer les milliers (trois chiffres) par un espace, de manière à permettre une meilleure lecture.

— Des normes relatives à la documentation des données peuvent encore être appliquées dans le cadre de la gestion et de la diffusion d'open data public. A l'instar de la norme ISO/CEI 11179 :2005, qui a pour objet la représentation des métadonnées pour leur organisation dans un registre. Le Dublin Core, qui consiste en un schéma de métadonnées génériques – comportant quinze descripteurs – pour la documentation de documents numériques, est l'objet de la norme ISO 15836 :2009 [3]. La norme ISO/IEC 19503 :2005 porte quand à elle sur l'interchangeabilité des métadonnées au format XML, lequel fait par ailleurs l'objet de recommandations spécifiques du W3C [4].

— La norme ISO/TS 14048 :2002 relative à l'analyse du cycle de vie et au format de documentation des données "*fournit des exigences et une structure pour un format de documentation de données à utiliser en vue d'une documentation transparente et claire, afin de faciliter l'échange de données d'analyse de cycle de vie et d'analyse d'inventaire de cycle de vie. (…) Le format de documentation des données précise une liste d'exigences sur la façon dont la documentation des données doit être répartie entre les champs de données, chacun de ces derniers comportant une description explicative, (…) ensuite précisée par la structure du format de documentation des données*". Cette norme s'applique spécifiquement à la gestion de données environnementales.

— On citera enfin la norme ISO/TS 8000-150 :2011 qui spécifie les principes fondamentaux du management de la qualité des données (*master data quality management*). Il

3. A noter que le Dublin Core émane de la Dublin Core Metadata Initiative, voir www.dublincore.org.
4. Lire infra.

propose notamment un "Master Data Vocabulary", en vue d'améliorer la qualité des données en proposant des listes de vocabulaire standardisé, dont l'objectif est une modélisation uniforme de l'information (Boydens 2012). Boydens souligne toutefois que le recours à un tel lexique est délicat " *en raison du caractère très volatile tant du langage naturel que des processus et des réalités représentés, cela dans tous les domaines empiriques qui sont par essence sujets à interprétations dans le temps et l'espace*".

Standards d'adressage En matière d'adresse postale, chaque pays édicte ses règles. L'UPU (Universal Postal Union) [5] ressemble les informations relatives à ses membres. Elle propose la norme S42, destinée à faciliter les échanges en uniformisant les pratiques d'encodage. La nome S42 permet ainsi de transformer des éléments d'adresse en une adresse formatée en fonction du pays, à partir de la nomenclature prénom – NOM – numéro de rue – type de rue - numéro –ville – région - code postal – pays. Pour la France, le format sera appliqué comme suit [6] : prénom – nom – numéro de rue – rue – code postal – ville.

Nomenclature pour la numérotation téléphonique L'Union Internationale des Télécommunications (UIT/ITU) propose la recommandation E123 an vue d'harmoniser la nomenclature des numérotations téléphoniques [7]. Elle prévoit deux modes de numérotations : national (dix chiffres séparés par des espaces, dont les deux premiers numéros correspondent au code territorial tel que défini par l'Autorité de régulation des communications électronique et des postes), international (la numérotation débute par deux chiffres, correspondant à l'indicatif du pays, précédés du sigle "+". Le troisième chiffre est celui du code territorial).

Recommandations du W3C Le W3C propose une série de recommandations relatives aux formats des données numériques, ainsi qu'une recommandation spécifique à la publication en ligne d'open data public. Les recommandations du W3C portent sur l'utilisation de données tabulaires (format CSV), sur l'organisation de données dans le métalangage XML [8], sur l'usage de métadonnées et de données liées [9] (formats RDF et OWL), de vocabulaire contrôlé SKOS, portant sur l'organisation des thésaurus), sur le langage de requête SPARQL (qui interagit avec le format RDF) [10].

5. L'organisation, fondée en 1874, compte 192 pays membres. Elle a son siège à Berne. Source : www.upu.int .
6. Source : www.upu.int/fileadmin/documentsFiles/activities/addressingUnit/fra.pdf
7. Source : www.itu.int
8. Extensible HyperText Markup Language
9. En anglais, les linked data, caractérisées par le triplet sujet-prédicat-objet. La promesse des données liées est de créer des liens qui ont du sens mais leur implémentation actuelle, souligne van Hooland et Verborgh, tend à être complexe.
10. Le W3C édicte également des recommandations relatives à la mise en forme des pages web : les formats HTML (*HyperText Markup Language*), pour la structure, et CSS (*Cascading Style Sheet*), pour la mise en forme). Le W3C propose également des recommandations sur les formats d'images SVG (*Scalable Vector Graphics*) et PNG (*Portable Network Graphics*)

Références

BATINI, C., CAPPIELLO, C., FRANCALANCI, C. et MAURINO, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR)*, 41(3):16.

MADNICK, S. et ZHU, H. (2006). Improving data quality through effective use of data semantics. *Data & Knowledge Engineering*, 59(2):460–475.

SETH, v. H. et RUBEN, V. (2014). *Linked Data for Libraries, Archives and Museums : How to Clean, Link and Publish Your Metadata*. Amer Library Assn.