ÉTAT DES LIEUX SUR LES MÉTADONNÉES RELATIVES AUX CONTENUS CULTURELS

GLOSSAIRE

OBSERVATOIRE
DE LA CULTURE ET DES
COMMUNICATIONS
DU QUÉBEC

Québec

Pour tout renseignement concernant l'ISQ
et les données statistiques qui y sont disponibles,
s'adresser à :

Institut de la statistique du Québec
200, chemin Sainte-Foy
Québec (Québec)
G1R 5T4
Téléphone : 418 691-2401

ou

Téléphone : 1 800 463-4090
(sans frais d'appel au Canada et aux États-Unis)

Site Web : www.stat.gouv.qc.ca

Dépôt légal
Bibliothèque et Archives nationales du Québec
1er trimestre 2018
ISBN : 978-2-550-80447-5 (en ligne)

Février 2018

GLOSSAIRE DES TERMES

Apprentissage profond *(deep learning)*	Ensemble de méthodes d'apprentissage automatique permettant de modéliser avec un haut niveau d'abstraction des données à partir d'architectures impliquant l'articulation de différentes transformations non linéaires. Ces techniques ont permis des progrès importants et rapides dans les domaines de l'analyse du signal sonore ou visuel et notamment de la reconnaissance faciale, de la reconnaissance vocale, de la vision par ordinateur, du traitement automatisé du langage (Source : Wikipédia).
Chaîne de blocs *(blockchain)*	Base de données sécurisée, qui stocke l'historique de tous les échanges effectués entre ses utilisateurs depuis sa création, et partagée entre ces derniers sans aucun intermédiaire qui centralise ces données. Dit autrement, la *blockchain* est une technologie de stockage et de transmission d'informations, transparente, et fonctionnant sans organe central de contrôle. Les *blockchains* peuvent être publiques ou privées. Pour plus d'information, voir : trends.cmf-fmc.ca/fr/blog/blockchain-la-prochaine-perturbation. Cette technologie est encore embryonnaire, mais elle a fait beaucoup parler d'elle depuis 2016. La première *blockchain* est apparue avec la monnaie numérique (*bitcoin*). Son caractère décentralisé, sa sécurité et sa transparence promettent des applications plus larges que le domaine monétaire (ex. transfert d'actifs, registre, contrats « intelligents »). Les champs d'exploitation sont immenses : banques, assurances, immobilier, santé, énergie, transports, vote en ligne, etc. De façon générale, des *blockchains* pourraient remplacer la plupart des « tiers de confiance » centralisés (banques, notaires, cadastre, etc.) par des systèmes informatiques distribués. Pour plus de détails, voir : blockchainfrance.net/decouvrir-la-blockchain/c-est-quoi-la-blockchain/. De belles avenues avec la *blockchain* sont envisagées pour l'industrie des contenus comme la musique ou les médias, notamment en faisant disparaitre les intermédiaires entre le créateur/producteur et l'usager. Voir à ce sujet le texte suivant : trends.cmf-fmc.ca/fr/blog/blockchain-la-prochaine-perturbation.
Coconstruction (coproduction, cocréation)	Terme utilisé lorsqu'il y a implication d'une pluralité d'intervenants dans l'élaboration et la mise en œuvre d'un projet ou d'une action. Dans sa stratégie sur les métadonnées, le ministère de la Culture et de la Communication de France définit les processus de coproduction et de coconception (*codesign*) comme un nouveau paradigme des institutions publiques dans le développement et le partage de connaissances.
Compatibilité	La compatibilité se distingue de l'interopérabilité. La compatibilité fait référence à l'idée qu'un matériel ou un logiciel est conforme aux règles d'interface d'un système informatique donné, ce qui fait que l'intégration de ce logiciel ou matériel n'altère pas les conditions de fonctionnement du système en question. (Dans l'interopérabilité, au contraire, le système s'adapte pour collaborer avec un autre).
Découvrabilité	Capacité intrinsèque d'un contenu, d'un produit ou d'un service disponible sur le Web à être découvert facilement par l'internaute, ou à ressortir spontanément du lot sans que l'internaute ait recherché ce contenu en particulier. Il s'agit du potentiel à capter l'attention de l'internaute, à se positionner, à l'aide de différentes techniques et outils, de manière à être facilement repérable et découvrable.

Dictionnaire de données

Collection de métadonnées ou de données de référence nécessaires à la conception d'une base de données relationnelle. Dans le cas d'une entreprise, par exemple, le dictionnaire décrit des données, comme les clients, les nomenclatures de produits et de services. Il est souvent représenté par un tableau à quatre colonnes contenant le nom, le code et le type de donnée ainsi que des commentaires. Un dictionnaire de données doit respecter les contraintes suivantes.

- Tous les noms doivent être monovalués et non décomposables.
- Il ne doit pas y avoir d'homonymes ni de synonymes.
- Les données y sont regroupées par entité.
- Les identifiants sont complètement précisés.
- Les commentaires doivent être pertinents.

(Source : Wikipédia)

Données ouvertes
(open data)

Ensembles de données sur support numérique, d'origine publique ou privée, produites par une collectivité, une agence publique, un gouvernement ou une entreprise. Les données ouvertes sont diffusées de manière structurée selon une méthode et une licence ouverte garantissant leur libre accès et leur réutilisation par tous, sans restriction technique, juridique ou financière (donc gratuitement). Elles sont « lisibles par machine », c'est-à-dire que leur format de présentation les rend lisibles et interprétables par un ordinateur.

Voici les trois principes fondamentaux en matière de données ouvertes :

- les données sont facilement disponibles ;
- les données sont librement accessibles (publiées sous licence libre) ;
- les données peuvent être réutilisées.

Les données ouvertes sont rendues accessibles au public par l'entremise de portails et d'outils de recherche, de manière à ce qu'elles puissent être réutilisées par les gouvernements, les citoyens, les organismes sans but lucratif et le secteur privé, qui peuvent les exploiter de manière inédite ou originale (Source : Wikipédia).

Données ouvertes liées
(linked open data)

Il faut distinguer les « données ouvertes » et les « données ouvertes liées » (DOL). Une donnée ouverte est une donnée librement utilisée, réutilisée et redistribuée par quiconque. Une donnée ouverte liée est unie à au moins une autre donnée par une combinaison « donnée-lien-donnée ». Une donnée peut être liée sans être ouverte, et vice versa. Une DOL associe donc deux modes différents de gestion de la donnée : ouverte d'une part, liée d'autre part.

Le concept de données ouvertes liées est étroitement associé à l'essor du Web sémantique. Les données liées du Web sémantique utilisent des URI (*Uniform Resource Identifier*) et le langage RDF (*Resource Description Framework*). Pour plus d'information, voir : linkeddata.org/.

Voir aussi les termes « Resource Description Framework (RDF) », « URI » et « Web sémantique ».

Économie numérique

Le gouvernement du Québec, dans son *Plan d'action en économie numérique*, définit l'économie numérique par ses grandes tendances mondiales.

- L'émergence de grands moteurs de changements, par exemple le téléphone mobile, l'infonuagique, les mégadonnées et les technologies financières.
- L'ouverture des données d'origine publique ou privée.
- La promotion par les gouvernements des données gouvernementales ouvertes.
- La convergence des secteurs des services et de la fabrication de TIC.
- La naissance et la croissance d'entreprises numériques faisant apparaître de nouveaux modèles d'affaires, de nouveaux produits et de nouveaux services, et remettant en question les cadres réglementaires.
- L'émergence d'une économie collaborative qui contribue à la création de nouveaux modèles de consommation, privilégiant les échanges et le partage de biens et de services.
- L'amélioration de la performance des réseaux de télécommunication grâce au déploiement de la fibre optique et de la quatrième génération (4G).
- Une augmentation du trafic Internet mondial et de l'utilisation des téléphones intelligents ou des tablettes numériques.
- La hausse du commerce électronique chez les personnes et les entreprises, facilitant l'achat et l'offre de produits et de services sur le marché mondial et représentant des défis logistiques importants.
- Des questionnements liés à la gouvernance (exigences relatives à l'acheminement, à la coproduction ou à la coconception).

Économie numérique *(suite)*

L'économie numérique se caractérise aussi par la place qu'occupe le secteur des TIC. Celle-ci est de plus en plus étendue, de sorte que ce secteur n'est plus associé à un ensemble défini d'entreprises et d'organisations, s'étant introduit dans tous les autres secteurs industriels.

Finalement, certains auteurs mentionnent que l'économie numérique se structure autour de l'échange de données, notamment de mégadonnées.

Source : www.economie.gouv.qc.ca/fileadmin/contenu/documents_soutien/strategies/economie_numerique/paen.pdf

Flux de métadonnées

Un flux de métadonnées est un flux d'information, soit un « transfert d'information entre deux acteurs du système d'information. Souvent représenté par des flèches, un flux part d'un acteur source pour aboutir à un acteur de finalité » (Source : www.linternaute.com/dictionnaire/fr/definition/flux-d-information/).

La circulation de l'information entre les acteurs du système est souvent représentée par un diagramme, appelé flux de métadonnées. Ces acteurs peuvent être internes ou externes. Les notions d'interne et d'externe peuvent varier selon le système à l'étude : une organisation, un domaine culturel, une ville, etc. L'acteur interne fait partie du système. L'acteur externe ne fait pas partie du système, mais a des échanges avec les acteurs internes dans le cadre de l'activité étudiée.

Le rapport *Étude de faisabilité relative à la mise en place de registres ouverts de métadonnées*, de la firme française BearingPoint (2014), identifie différentes étapes dans le flux de métadonnées.

- La création : le fait de créer la métadonnée (ex. : création du code ISWC par la SACEM).
- L'agrégation : l'agglomération de métadonnées (il peut s'agir d'une agrégation horizontale de métadonnées relatives à un même contenu, ou d'une agrégation verticale de métadonnées homogènes décrivant des contenus différents). Lors de cette phase, certains acteurs valident et modifient les métadonnées pour garantir leur exactitude.
- L'utilisation : les métadonnées sont utilisées à plusieurs fins
  - Pour donner plus d'information aux utilisateurs, plus de valeur et certifier les informations transmises
  - Pour faciliter la recherche des contenus culturels
  - Pour gérer la répartition des droits
  - Pour donner de la visibilité aux contenus les moins exposés (théorie de la longue traîne).

Folksonomie

« Folksonomie » est une adaptation française du terme anglais *folksonomy*, qui combine les mots *folk* (le peuple, les gens) et *taxonomy* (taxonomie). Ainsi, une « folksonomie », ou indexation personnelle, est un système de classification collaborative décentralisé et spontané, basé sur une indexation effectuée par des non-spécialistes. À l'inverse des systèmes hiérarchiques de classification, les contributeurs d'une « folksonomie » ne sont pas contraints à une terminologie prédéfinie, mais peuvent adopter les termes qu'ils souhaitent pour classer leurs ressources. Ces termes sont souvent appelés « mots-clés » ou « *tags* » (Source : Wikipédia).

FRBR
(*Functional Requirements for Bibliographic Records* ou spécifications fonctionnelles des notices bibliographiques)

Les FRBR sont une modélisation conceptuelle des renseignements contenus dans les notices bibliographiques des bibliothèques. Elles ne sont pas une norme de notice bibliographique, mais décrivent les renseignements d'une notice bibliographique d'un point de vue logique en utilisant un modèle « entité-association ».

Les FRBR organisent les différentes composantes de la description bibliographique en trois groupes d'entités liées entre elles.

- Entités du groupe 1 : Représentent les différents aspects de ce qu'un utilisateur peut trouver dans les produits d'une activité intellectuelle ou artistique, c'est-à-dire les documents et leurs différentes versions. Ces entités sont : œuvre, expression, manifestation, document.
- Entités du groupe 2 : Correspondent à la modélisation des personnes physiques ou morales qui ont une responsabilité dans le contenu intellectuel ou artistique, la production matérielle et la distribution, ou la gestion juridique des entités du premier groupe. Elles sont de deux ordres, soit les personnes et les collectivités.
- Entités du groupe 3 : Entités qui sont le sujet des œuvres (concept, objet, événement, lieu). La relation de sujet peut aussi fonctionner avec le groupe 2 et le groupe 1.

Pour plus d'information, voir : fr.wikipedia.org/wiki/Sp%C3%A9cifications_fonctionnelles_des_notices_bibliographiques.

Voir aussi « WEMI ».

Format d'enregistrement	Il y a deux grandes catégories de format d'enregistrement d'un document numérique. 1. Le format « natif » ou « propriétaire » de chaque application ou logiciel. Par défaut, une application encode les documents qu'elle fabrique sous une forme qui lui est propre et que les autres applications ne comprennent pas : c'est ce qu'on appelle le format natif ou propriétaire du logiciel. Un document enregistré dans le format natif d'une application n'est pas lisible par une autre application. 2. Le format standard à l'échelle internationale (aussi appelé « format d'échange »). Il permet d'enregistrer des documents qui pourront être ouverts avec d'autres applications que celle qui a servi à les créer. On utilise des formats d'échange différents pour les documents textuels, les images, les documents audio, etc. Exemples de format d'échange : rtf, gif, jpeg, tiff, eps, mp3.
Gouvernance	Notion à définitions multiples, le terme « gouvernance » recouvre l'idée de « bien gouverner » ou « bien gérer » dans un contexte de décentralisation des décisions et de multiplication des instances de décision, et ce, afin d'atteindre un but ou de réaliser un projet. Il renvoie aussi à l'idée de mise en place de modes de pilotage ou de régulation qui sont souples et éthiques, fondés sur un partenariat ouvert et éclairé entre différentes parties prenantes (Source : Wikipédia).
Granularité	La notion de granularité fait référence à la taille du plus petit élément d'un système, à sa finesse. Quand on atteint le niveau maximal de granularité d'un système, on ne peut plus découper l'information. À titre d'exemple, dans la description d'un document de bibliothèque, le niveau de granularité indique à quel point la description est détaillée dans la hiérarchie des composantes bibliographiques : on peut avoir simplement le titre du périodique, avoir aussi l'information quant au numéro de ce périodique, avoir aussi l'information quant à l'article de ce numéro (Source : Wikipédia).
Identifiant (identifiant unique)	Chaîne de caractères alphanumériques qui a pour fonction d'identifier de manière stable un document, une ressource ou une entité, quelle que soit sa nature. En principe, un identifiant devrait être unique pour chaque ressource. Voici, à titre d'exemples, quelques identifiants utilisés dans le contexte de la culture numérique : EIDR, IPI, ISAN, ISLI, ISNI, ISBN, ISRC, ISTC, ISWC, UN/LOCODE.
Intelligence collective	Désigne les capacités cognitives d'une communauté résultant des interactions multiples entre ses membres. Suppose quelques conditions, dont : 1. une communauté d'intérêts avec une libre appartenance, une structure horizontale, une gestion collective ; 2. un espace collaboratif avec des outils de coopération, un système d'information, un processus d'apprentissage (Source : Wikipédia).
Internet des objets	L'Internet des objets (ou IdO ; en anglais, *Internet of Things* (ou IoT)) représente l'extension du réseau à des choses et à des lieux du monde physique. Alors qu'Internet ne se prolonge habituellement pas au-delà du monde électronique, l'Internet des objets connectés représente les échanges d'information et de données provenant de dispositifs présents dans le monde réel vers le réseau Internet. Il revêt un caractère universel pour désigner des objets connectés aux usages variés, dans le domaine de la santé, de la domotique ou du *quantified self*. L'Internet des objets est en partie responsable d'un accroissement exponentiel du volume de données généré sur le réseau, à l'origine du *big data* (ou mégadonnées en français) (Source : Wikipédia).
Interopérabilité	Capacité d'un système à s'adapter afin de fonctionner ou collaborer avec d'autres systèmes indépendants, avec des plateformes matérielles et logicielles différentes, et ce, sans restriction d'accès ou de mise en œuvre, en vue de créer un réseau et de faciliter le transfert des données avec un minimum de pertes de contenu et de fonctionnalité. L'interopérabilité nécessite que les communications obéissent à des normes clairement établies et univoques (ex. : format de données échangées, tensions et courants à utiliser, types de câbles à utiliser). Chaque système les intègre dans son propre fonctionnement. Ces normes jouent un double rôle : d'abord, elles indiquent la façon dont le dialogue entre les différents éléments va s'opérer et, ensuite, elles permettent une passerelle de communication qui sera en mesure de s'adapter aux besoins changeants des éléments.

Interopérabilité *(suite)*

Dans le cadre du schéma de métadonnées Dublin Core, par exemple, le paysage des métadonnées est caractérisé par les quatre niveaux d'interopérabilité suivants.

1. Partage des définitions de termes (*Shared term definitions*). À ce niveau, l'interopérabilité suppose une utilisation de définitions des métadonnées partagées. Ainsi, les participants conviennent des termes à employer dans leurs métadonnées ainsi que de leur définition. Ce premier niveau fait référence principalement à des environnements d'application, comme un intranet, un système de bibliothèque. L'interopérabilité avec « le reste du monde » n'est généralement pas une priorité. Plusieurs applications de métadonnées existantes fonctionnent pour le moment à ce niveau d'opérabilité.
2. Interopérabilité sémantique formelle (*Formal Semantic Interoperability*). L'interopérabilité entre les applications d'utilisation de métadonnées est basée sur le modèle formel de partage fourni par RDF, qui est utilisé pour soutenir le Web des données liées (*linked data*). La pratique du *linked data* consiste à exposer, partager et connecter des données d'information.
3. *Description Set Syntactic Interoperability*. Les applications sont compatibles avec le modèle de données liées et, en plus, elles partagent une syntaxe abstraite pour l'enregistrement de métadonnées pouvant être validées, le *description set*.
4. *Description Set Profile Interoperability*. Les dossiers/enregistrements échangés entre les applications de métadonnées sont soumis à un ensemble commun de contraintes, utilisent les mêmes vocabulaires et reflètent un modèle partagé du monde.

Les niveaux 1 et 2 sont plus fréquents dans la pratique que les niveaux 3 et 4.

L'interopérabilité concerne trois aspects de l'échange d'informations :

- Interopérabilité technique ou informatique : « Pouvoir communiquer ». Elle concerne les problèmes techniques de liaison entre systèmes, la définition des interfaces, le format des données et les protocoles, y compris les télécommunications. Elle décrit la capacité pour des technologies différentes à communiquer et à échanger des données basées sur des normes d'interface bien définies et largement adoptées. Elle est considérée comme importante, voire déterminante. Elle déterminerait l'interopérabilité globale. Fait référence notamment à des protocoles, comme HTTP, OAI-PMH, etc.
- Interopérabilité sémantique : « Savoir se comprendre ». Elle assure que la signification exacte des renseignements échangés soit comprise par n'importe quelle autre application, même si celle-ci n'a pas été conçue initialement dans ce but précis. En effet, des conflits sémantiques surviennent lorsque les systèmes n'utilisent pas la même interprétation de l'information, définie différemment d'une organisation à l'autre. Pour réaliser l'interopérabilité sémantique, les deux parties doivent se référer à un modèle commun d'échange d'informations, tels que les jeux de métadonnées Dublin Core, MAR-SML, MODS, etc.
- Interopérabilité syntaxique : « Savoir communiquer ». Elle concerne la façon dont sont codées et formatées les données, en définissant notamment la nature, le type et le format des messages échangés. Des langages structurés, tels que XML ou RDF, permettent l'interopérabilité syntaxique.

Langage informatique

Langage formel utilisé lors de la conception, la mise en œuvre ou l'exploitation d'un système d'information. Il s'agit d'un ensemble organisé de symboles, de mots-clés, de caractères et de règles (instructions et syntaxe) utilisé pour adresser des commandes à l'ordinateur et assurer la communication avec la machine.

Selon le type d'information à communiquer à la machine, on distingue le langage de commande, le langage de programmation, le langage d'interrogation, etc. Par exemple, un langage de requête est un langage informatique utilisé pour accéder aux données d'une base de données ou à d'autres systèmes d'information (Source : Wikipédia).

Mégadonnée
(big data)

Terme utilisé pour décrire des lots extrêmement volumineux de données numériques structurées, mi-structurées et non structurées, ayant le potentiel de contenir une mine de renseignements, mais si volumineux qu'ils deviennent difficiles à stocker, à gérer, à archiver et à analyser avec des outils classiques de gestion et d'analyse de bases de données. Le concept de « mégadonnées » fait également référence à la capacité de mettre en relation plusieurs bases de données et, grâce à l'utilisation d'analyses avancées, d'identifier des *patterns* d'information qui demeureraient autrement invisibles.

Les mégadonnées présentent trois caractéristiques clés, communément appelées 3Vs.

- Volume : Les lots de mégadonnées sont gigantesques, d'où le préfixe « méga ». Les termes souvent utilisés en référence au volume des mégadonnées sont « pétabytes » (ou pétaoctet) et « exabytes » (ou exaoctet) de données.
- Vélocité : Les mégadonnées sont caractérisées par la grande vitesse à laquelle elles se constituent (souvent de manière automatique, au fil des « clics » ou des transactions Web) et par le fait qu'elles sont en constant changement, reflétant une situation en temps réel.
- Variété : Il s'agit d'ensembles composés de données structurées (généralement du texte organisé dans des bases de données relationnelles traditionnelles) et de données non structurées (photos, vidéos, données textuelles).

L'étude *Profil du Big Data au Québec*, publiée par Montréal international en 2016, y associe une quatrième caractéristique :

- Véracité : Précision sur les sources, la qualité et la validité des données.

R. Concessao, quant à lui, dans son livre « *What is Big Data really ?* », reconnaît deux autres dimensions au Big Data[1].

- Variabilité : Les flux de données varient grandement selon les périodes de pointe (journalière, saisonnière, ponctuelle due à un événement précis), ce qui représente un défi pour les gestionnaires de ces données.
- Complexité : Puisqu'elles proviennent de multiples sources, les données peuvent être difficiles à lier, harmoniser, transformer.

Pour plus d'information, voir : trends.cmf-fmc.ca/media/uploads/reports/37-rapport-tendances-accessibilite-accrue.pdf.

Métamoteur
(metasearch engine)

Un métamoteur (ou méta-moteur) ou un méta-chercheur est un moteur de recherche qui puise ses renseignements dans plusieurs moteurs de recherche généralistes. De manière plus précise, le métamoteur envoie ses requêtes à plusieurs moteurs de recherche et retourne les résultats de chacun d'eux. Il permet aux utilisateurs de n'entrer le sujet de leur recherche qu'une seule fois tout en accédant aux réponses de plusieurs moteurs de recherche différents. Il élimine les résultats similaires et trie les résultats pour faire ressortir en premier les pages fournies par plusieurs moteurs. Certains métamoteurs indiquent de quels moteurs de recherche provient chaque résultat (Source : Wikipedia).

Voir aussi « Métarecherche ».

Métarecherche
(metasearching)

Une métarecherche est une recherche qui puise des renseignements dans plusieurs bases de données, sources, plateformes et protocoles à la fois. En informatique, elle est généralement réalisée par un métamoteur.

Voir aussi « Métamoteur ».

Microformat

Approche de formatage de données dans des pages Web, qui cherche à rationaliser et standardiser le contenu existant, comme les métadonnées, en utilisant des classes et attributs de balises HTML et XHTML. Cette approche est conçue pour permettre à l'information destinée aux utilisateurs finaux, telle que le carnet d'adresses, les coordonnées géographiques, les numéros de téléphone, les événements et autres données ayant une structure constante, d'être traitée automatiquement par les logiciels.

Même si le contenu des pages Web était déjà capable techniquement d'être traité automatiquement depuis la conception du Web, il existait certaines limites. Les balises traditionnelles de marquage étaient en effet utilisées pour afficher l'information sur le Web et non pas pour la décrire. Les microformats sont destinés à combler ce fossé en attachant de la sémantique par la standardisation de la codification des balises HTML et XHTML. Cela permet d'éviter d'autres méthodes plus compliquées de traitement automatisé, comme le traitement du langage naturel ou le *screen scraping*. L'utilisation, l'adoption et le traitement des microformats permettent aux éléments de données d'être indexés, cherchés, sauvegardés ou référencés de manière que l'information puisse être réutilisée ou combinée (Source : Wikipédia).

Modèle conceptuel des données

Le modèle conceptuel des données (MCD) a pour but d'écrire de façon formelle les données qui seront utilisées par un système d'information. Il s'agit donc d'une représentation des données, facilement compréhensible, permettant de décrire le système d'information à l'aide d'entités.

1. CONCESSAO, R. (2015). *What is Big Data really?* En ligne : www.amazon.com/What-Big-Data-R-Concessao/dp/153516185X.

Moissonnage *(Web scraping* ou *harvesting)*	Technique d'extraction du contenu de sites Web, par l'intermédiaire d'un script ou d'un programme, dans le but de le transformer pour le réutiliser dans un autre contexte, par exemple le référencement. Cette opération se pratique le plus souvent de façon automatique, ce qui permet de constituer des pages à bon compte. Le moissonnage du Web (*Web scraping* ou *harvesting*) peut être utilisé pour récupérer des métadonnées (Source : Wikipédia).
Moteur de recherche	Un moteur de recherche est une application Web permettant de trouver des ressources à partir d'une requête sous forme de mots. Les ressources peuvent être des pages Web, des articles de forums Usenet, des images, des vidéos, des fichiers, etc. Certains sites Web (Google, par exemple) offrent un moteur de recherche comme principale fonctionnalité ; on appelle alors *moteur de recherche* le site lui-même (Source : Wikipédia).
Norme (standard)	Les termes « norme » et « standard » sont considérés comme des synonymes pour les besoins du présent document. (Cette simplification repose notamment sur le fait que la documentation consultée est majoritairement de langue anglaise et qu'en anglais, « norme » se dit « standard ».) Une norme (ou un standard) est un ensemble de règles ou de procédures qui définit la façon d'effectuer une activité. Dans le contexte des métadonnées, les normes (ou standards) établissent un ensemble commun de termes et de définitions pour les concepts et composantes. La normalisation ou la standardisation est le fait d'établir des normes ou standards techniques, c'est-à-dire d'élaborer un référentiel commun et documenté destiné à harmoniser l'activité d'un secteur. Elle est réalisée par des organismes spécialisés, par exemple des organisations créées par les professionnels ou les entreprises d'un secteur d'activité donné. Voici quelques exemples de normes (ou standards). • MARC : norme utilisée dans les bibliothèques pour soutenir le catalogage lisible par machine. Elle détermine les champs de métadonnées et inclut les exigences de codage pour la création d'enregistrements individuels. • Dublin Core : schéma générique de métadonnées utilisé surtout en bibliothéconomie. Il comprend 15 champs d'information (titre, créateur, éditeur, langue, titulaire du copyright, etc.). Il s'applique à des contenus aussi bien physiques que numériques. • VRA 4.0 : norme pour la description des ressources visuelles, comme une image. • GRid (*Global Release Identifier*) : norme pour l'industrie de la musique. Le GRid est un système de codes uniques qu'un distributeur assigne à chaque publication qu'il souhaite distribuer sur le Web. Le GRid vise à standardiser la manière d'identifier les produits et à pouvoir suivre la trace de leur distribution. • RDA : Ressources : Description et accès. Norme utilisée par les bibliothèques pour décrire les documents de tous types.
Notice d'autorité	En science de l'information, une autorité (ou notice d'autorité, ou forme d'autorité) sert à identifier sans ambiguïté des personnes, des collectivités, des noms géographiques, des œuvres, des objets ou des concepts. La forme retenue fait autorité, d'où son nom (calqué de l'expression *authority control*). Les objectifs de l'autorité sont doubles : • faire en sorte qu'une même réalité soit toujours indiquée de la même manière ; • éviter qu'une même désignation ne s'applique à deux personnes ou à deux lieux. Pour ces raisons, une liste d'autorité est contrôlée. Selon le cas, ou bien il est possible de créer de nouvelles notices d'autorité, mais en respectant des normes qui indiquent la forme à suivre (par exemple nom, prénom (date de naissance-date de décès)), ou bien la liste est confiée à un organisme chargé de la suivre, et seules les personnes travaillant dans cet organisme peuvent la modifier, au contraire d'une liste d'index. Les notices d'autorité sont utilisées dans les catalogues de bibliothèques, les fichiers, les bases de données et les systèmes d'information. Initialement utilisées dans les bibliothèques, elles sont également employées dans la description des documents d'archives ou dans la gestion des droits numériques (Source : Wikipédia).

Ontologie	En informatique et en science de l'information, une ontologie est un ensemble structuré des termes et concepts, un réseau sémantique permettant de donner un sens aux informations relativement à un domaine de connaissances. Ces concepts sont liés les uns aux autres par des relations taxonomiques (hiérarchisation des concepts), d'une part, et sémantiques d'autre part. En quelque sorte, l'ontologie est aux données ce que la grammaire est au langage. En somme, une ontologie est un vocabulaire contrôlé qui est organisé de manière à représenter un domaine de connaissances dans son ensemble et à décrire les relations entre les termes. Il existe des ontologies qui peuvent servir de standards dans le Web sémantique, comme le FOAF (*Friend of a Friend*) pour la description de personnes[2], BIO (*Biography Ontology*) pour la description de biographies[3], TIME pour la description du temps. Par ailleurs, le « *Web Ontology Language* (OWL) » est un langage de représentation des connaissances construit sur le modèle de données de RDF, pour décrire des ontologies. Ce langage est en quelque sorte un standard informatique qui met en œuvre certaines logiques de description et permet à des outils, dont OWL, de travailler avec les données, de vérifier qu'elles sont cohérentes, de déduire des connaissances nouvelles, etc. Pour plus d'information, voir : liris.cnrs.fr/amille/enseignements/Master_PRO/TIA/RAPC/igc_rapc_Folder/fig41.gif.
Passerelle de métadonnées (*metadata crosswalk*)	Une passerelle (aussi appelée « mappage sémantique ») est un plan faisant le lien sémantique et syntaxique d'un standard de métadonnées à un autre. Elle permet donc que les métadonnées créées par un groupe soient utilisées par un autre qui emploie un standard différent. Le degré de succès d'une passerelle dépend, entre autres, de la similarité entre les deux standards, de la granularité des éléments comparés et de la compatibilité des règles de contenu utilisées dans chacun des standards. Pour plus d'information, voir : NISO Press (2004), *Understanding Metadata*, p. 12, www.niso.org/publications/press/UnderstandingMetadata.pdf.
Prescripteur de goût	Un « prescripteur de goût », ou « influenceur », est une personne ou un groupe de personnes ayant une influence sur le choix de produits, de services ou encore de contenus. À titre d'exemple, le prescripteur de goût peut exploiter un blog spécialisé, ou encore il peut s'agir d'un éminent critique de cinéma ou d'une figure publique, comme Oprah Winfrey. Le prescripteur fait des recommandations sur des contenus et fait autorité auprès de son public. Sur le Web, les prescripteurs ont un rôle significatif et ils sont nombreux, avec des expertises, des centres d'intérêt et des auditoires qui leur sont propres. Ils ont une grande influence sur la découvrabilité des contenus culturels. On peut voir les prescripteurs de goût comme constituant un réseau. Le premier en début de chaîne est le prescripteur zéro. Sa prescription à l'égard d'un produit culturel donné peut être déterminante dans le succès de ce produit auprès des consommateurs. Ses recommandations seront relayées par d'autres prescripteurs, et ainsi de suite.
Protocole	Dans l'univers de l'informatique et des télécommunications, un protocole est une méthode standard qui permet la communication entre des processus (s'exécutant, le cas échéant, sur différentes machines), c'est-à-dire un ensemble de règles et de procédures à respecter pour émettre et recevoir des données sur un réseau. Il en existe plusieurs, selon ce que l'on attend de la communication. Certains protocoles seront, par exemple, spécialisés dans l'échange de fichiers (comme le *File Transfert Protocol*, ou FTP) ; d'autres pourront servir à gérer simplement l'état de la transmission et les erreurs.

2. FOAF, xmlns.com/foaf/spec/
3. BIO, vocab.org/bio/

Registre (ou référentiel) de métadonnées

Système de gestion des métadonnées formel qui fournit l'information d'autorité sur la sémantique et la structure de chaque élément. Pour chaque élément, le registre donne la définition, les qualificatifs qui lui sont associés, ainsi que les *correspondances* avec des équivalents dans d'autres langues ou d'autres schémas.

C'est en quelque sorte la colonne vertébrale d'un système d'information. Il comprend la définition des données et les règles de gestion des données. Il comprend deux types de renseignements : 1) les données dont les applications de l'ensemble du système d'information ont besoin pour fonctionner ; soit les « données de référence » ; 2) les renseignements plus techniques qui seront utilisés pour faire évoluer une application en particulier.

Le registre de métadonnées est un élément essentiel pour garantir la protection du patrimoine informationnel, dans des contextes où l'on doit échanger à grande échelle des renseignements et faire circuler de l'information, avec de fortes contraintes d'interopérabilité. Dans un projet de gestion des métadonnées, la création d'un registre permet que les relations de confiance entre les parties prenantes reposent sur des définitions et des structures revues et approuvées par tous.

Un registre de métadonnées a typiquement les caractéristiques suivantes.

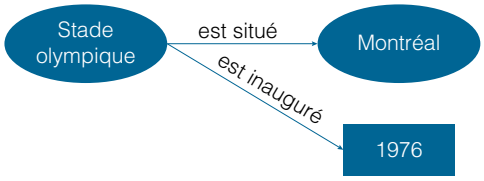
1. C'est une zone protégée où seules des personnes autorisées peuvent faire des modifications.
2. Il enregistre des éléments qui incluent à la fois la sémantique et les classes de représentation.
3. Les zones sémantiques d'un registre de métadonnées contiennent la définition précise d'un élément.
4. Les zones de représentation d'un registre de métadonnées définissent comment la donnée est représentée dans un format déterminé, comme une base de données ou une structure de format de fichier de type XML.

Un registre ou référentiel clair, logique et précis est un des gages de bonne interopérabilité d'un système d'information.

***Resource Description Framework* (RDF)**
(canevas de description de ressources)

Modèle de graphe destiné à décrire de façon formelle les ressources Web et leurs métadonnées, de façon à permettre le traitement automatique de telles descriptions. Élaboré par le W3C, le RDF est le langage de base pour représenter l'information sur des ressources dans le Web sémantique (Source : Wikipédia).

Le RDF est un modèle où les données sont structurées par des triplets (sujet, prédicat, objet). Le sujet est une ressource, le prédicat est une propriété et l'objet est un élément littéral ou une autre ressource. Le graphique suivant exprime clairement l'idée de triplet.



Ressource culturelle numérique

Fixation numérique de l'expression d'une œuvre abstraite, comme un enregistrement sonore, un enregistrement audiovisuel, une photographie, un logiciel, une image graphique ou un passage de texte.

Schéma de métadonnées

Modèle composé d'un ensemble déterminé de champs de métadonnées, conçu dans un but spécifique. Un schéma de métadonnées spécifie généralement les noms des champs et la sémantique des renseignements contenus dans ces champs. Il peut également contenir des règles de syntaxe, c'est-à-dire des règles indiquant comment les champs et leurs contenus doivent être encodés.

Plusieurs schémas de métadonnées ont été conçus pour une variété de disciplines et d'utilisateurs. Un des plus connus est le Dublin Core.

Science des données
(data science)

La science des données est une nouvelle discipline qui s'appuie sur des connaissances en mathématiques, en statistiques, en informatique et en visualisation des données. Elle est principalement une « science des données numériques ». Elle est en plein développement dans le monde universitaire et trouve des applications dans les secteurs privé et public.

Le *data scientist* produit des méthodes (automatisées, autant que possible) de tri et d'analyse de mégadonnées et de sources plus ou moins complexes ou disjointes de données, afin d'en extraire des renseignements utiles.

Pour plus d'information, voir : fr.wikipedia.org/wiki/Science_des_donn%C3%A9es.

Tag (mot-clé, étiquette)	Les *tags* sont des mots-clés insérés, par exemple, dans des pages html. Ils sont couramment utilisés dans les blogues. Les gestionnaires de site ou auteurs de pages html choisissent de façon personnelle des mots-clés et les attachent à des éléments textuels ou à des images faisant partie de leur site, afin de les marquer comme appartenant à une catégorie ou à un sujet donné. Les sites Web et les blogues avec des *tags* identiques peuvent alors être liés ensemble, ce qui permet aux internautes de trouver des contenus similaires ou liés. Les *tags* ne sont pas directement liés au Web sémantique. Bien que l'utilisation de *tags* comme système de classement ait l'avantage d'être souple et facile, sa principale faiblesse est qu'un même *tag* peut avoir différents sens.
Tatouage numérique (*digital watermark* ou, en France, *watermarking*)	Technique permettant d'ajouter des informations de copyright ou d'autres types à un fichier ou signal audio, vidéo, une image ou un autre document numérique. Le message inclus dans le signal hôte, généralement appelé « marque », peut être le nom du créateur ou un identifiant du créateur, du propriétaire, de l'acheteur ou encore une forme de signature décrivant le signal hôte. Le nom anglais de cette technique provient du marquage des documents papier et des billets (Source : Wikipédia).
Taxonomie	Sous-ensemble de vocabulaire contrôlé organisé selon une structure hiérarchique. En tant que vocabulaire contrôlé structuré, un thésaurus est un exemple classique de taxonomie.
Thésaurus	Vocabulaire contrôlé et dynamique de termes ayant entre eux des relations sémantiques et génériques, et qui s'applique à un domaine particulier de la connaissance. Si, du point de vue de sa structure, un thésaurus est un langage documentaire, du point de vue de sa fonction, c'est un instrument de contrôle destiné à éliminer les ambiguïtés du langage naturel, et qui exploite une liste exclusive de termes à utiliser obligatoirement pour la caractérisation du contenu des documents à analyser, à enregistrer, à indexer et à classifier (Source : *Grand dictionnaire terminologique* de l'Office québécois de la langue française). Terme voisin de « ontologie ». Pour plus d'information, voir : www.lattice.cnrs.fr/sites/itellier/poly_info_ling/thesaurus.jpg. Exemple de thésaurus : le thésaurus de l'activité gouvernementale du gouvernement du Québec : www.thesaurus.gouv.qc.ca/tag/accueil.do ;jsessionid=03EDF48BC336DCAFAEFA059C96F27EA0.
Traçabilité (des données)	La traçabilité renvoie au fait de disposer de l'information nécessaire et suffisante pour connaître (parfois de façon rétrospective) la composition d'un matériau ou d'un produit tout au long de sa chaîne de production et de distribution. Lorsqu'on parle de traçabilité des données, on fait référence à la capacité de valider la qualité des informations, notamment en répondant aux questions suivantes : Qui a produit les données ? Dans quel contexte ? À quel moment ? Quel est le degré d'expertise de cette source ? Quel est l'historique des interventions opérées sur les données (date de dernière mise à jour, par exemple) ? (Source : Wikipédia).
***Uniform Resource Identifier* (URI)** (identifiant uniforme de ressource)	Un URI étend le principe d'URL. L'URL représente un document Web, tandis que l'URI représente un objet unique (une personne, un lieu, un livre, etc.). Ainsi, l'URI est une courte chaîne de caractères qui identifie une ressource physique ou abstraite sur un réseau, et dont la syntaxe respecte une norme d'Internet mise en place par le *World Wide Web*. Par exemple, l'URI de la tour Eiffel est : http://dbpedia.org/resource/Eiffel_Tower Ainsi, le triplet « Tour Eiffel – est situé – Paris » se lirait comme suit : *(http://dbpedia.org/resource/Eiffel_Tower, http://dbpedia.org/ontology/location, http://dbpedia.org/resource/Paris)* Un URI doit permettre d'identifier une ressource de manière permanente, même si elle est déplacée ou supprimée.
***Uniform Resource Locator* (URL)** (localisateur uniforme de ressource)	Une URL est un format de nommage universel pour désigner une ressource sur Internet. Il s'agit d'une chaîne de caractères qui se décompose en cinq parties. • Le nom du protocole, c'est-à-dire le langage utilisé pour communiquer sur le réseau. Le protocole le plus largement utilisé est HTTP (*HyperText Transfer Protocol*), celui-ci permettant d'échanger des pages Web en format HTML. • L'identifiant et le mot de passe : permettent de spécifier les paramètres d'accès à un serveur sécurisé. • Le nom du serveur : il s'agit d'un nom de domaine de l'ordinateur hébergeant la ressource demandée. • Le numéro de port : il s'agit d'un numéro associé à un service permettant au serveur de savoir quel type de ressource est demandé. • Le chemin d'accès à la ressource : cette dernière partie permet au serveur de connaître l'emplacement auquel la ressource est située, soit le répertoire et le nom du fichier demandé.

Vocabulaire contrôlé	Liste de termes, titres ou codes prescrits, chacun représentant un concept à désignation unique. Exemples : listes de codes ou listes d'autorités de noms (noms de pays, par ex.), systèmes de classement, systèmes de vedettes-matière, thésaurus, ontologies.
Web des données *(linked data)*	Le Web des données est une des applications du Web sémantique, caractérisée par une série de principes relatifs à la publication de données et à la création de liens entre celles-ci en utilisant, entre autres, le modèle de graphe *Resource Description Framework* (RDF), afin que les machines puissent les interpréter. Voir aussi « Interopérabilité » et « Web sémantique ».
Web sémantique	Le Web sémantique est une extension du Web traditionnel qui implique l'application de standards définis par le *World Wide Web Consortium* (W3C) et qui visent à permettre aux machines de comprendre l'information sur le Web et à faciliter l'échange, la modélisation, l'encodage et l'interrogation des données au sein des applications, des organisations et des communautés. Voir aussi « Données ouvertes liées ».
WEMI *(Works, Expressions, Manifestations, Items)*	Issues du modèle FRBR, les WEMI sont des entités qui ont des attributs de même que des relations avec d'autres entités, y compris d'autres œuvres, expressions et manifestations. Les termes « Works » (« œuvres », en français) et « Expressions » renvoient à l'activité intellectuelle ou artistique et au contenu, tandis que « Manifestations » et « Items » font référence aux caractéristiques physiques. Une « œuvre » est réalisée par une « expression », qui est incarnée dans une « manifestation », qui est exemplifiée par un « item ». De façon plus détaillée : • L'œuvre (*work*) est une idée abstraite ou une création intellectuelle distincte. Les attributs d'une œuvre sont : le titre, la date, l'identifiant, l'auditoire visé, la forme de travail, le moyen de représentation, la désignation numérique, etc. • L'expression est l'accomplissement de cette idée à travers les mots, le son, l'image, etc. L'expression est une entité intellectuellement / artistiquement concrète. Elle est l'expression d'une œuvre sous forme de notation alphanumérique, musicale, chorégraphique, cartographique, etc. Une expression n'a pas de caractéristiques physiques. Les attributs d'une expression sont : le titre, la forme, la date, la langue, le mode de représentation, l'identifiant, etc. • La manifestation est le mode de réalisation physique (c'est-à-dire la publication ou la fixation) d'une expression d'une œuvre. Les attributs d'une manifestation sont : le titre, l'énoncé de responsabilité, l'édition, l'empreinte (lieu, éditeur, date), la forme / l'étendue et les dimensions, les conditions de disponibilité, le mode d'accès, l'identifiant, etc. • L'item est la copie réelle de la manifestation que prend l'expression qui appartient à une personne ou à une personne morale. C'est la seule entité absolument concrète. Les attributs d'un item sont la provenance, le propriétaire, l'emplacement, la condition, les restrictions d'accès, l'identifiant, etc. Voici un exemple de WEMI : • Œuvre : *Germinal*, d'Émile Zola. • Expression : la traduction anglaise de *Germinal*, par Roger Pearson. • Manifestation : *Germinal*, de Zola, traduit par Roger Pearson et publié chez Penguin Books en 2004. • Item : l'exemplaire de *Germinal*, de Zola, traduit par Roger Pearson et publié chez Penguin Books en 2004, qui se trouve à la bibliothèque municipale de Perpignan.

Institut
de la statistique
Québec