Documentation et bibliothèques

Étapes pour le développement d'un projet de données ouvertes et liées en bibliothèque

The Stages in the Development of an Open Data Project Linked to the Library

Marielle St-Germain

Volume 63, numéro 4, octobre–décembre 2017

Les données et les sciences de l'information

URI : https://id.erudit.org/iderudit/1042309ar
DOI : https://doi.org/10.7202/1042309ar

Aller au sommaire du numéro

Éditeur(s)

Association pour l'avancement des sciences et des techniques de la documentation (ASTED)

ISSN

0315-2340 (imprimé)
2291-8949 (numérique)

Découvrir la revue

Citer cet article

St-Germain, M. (2017). Étapes pour le développement d'un projet de données ouvertes et liées en bibliothèque. *Documentation et bibliothèques*, *63*(4), 35–45. https://doi.org/10.7202/1042309ar

Résumé de l'article

On constate depuis quelques années un intérêt de plus en plus marqué pour le Web sémantique et les avantages qu'il présente pour les institutions documentaires. Cependant, un flou persiste quant au potentiel, au fonctionnement et aux étapes de développement liées à des projets qui mettent en oeuvre les différents standards et composantes qui le caractérisent. L'objectif du présent article est de proposer une liste d'étapes pour la mise en oeuvre d'un projet de données ouvertes et liées (Web de données) en bibliothèque afin de faciliter l'appropriation de ces technologies par les professionnels de l'information.

ÉTAPES POUR LE DÉVELOPPEMENT D'UN PROJET DE DONNÉES OUVERTES ET LIÉES EN BIBLIOTHÈQUE

Marielle St-Germain
Étudiante au doctorat
École de bibliothéconomie et des sciences de l'information (EBSI)
Université de Montréal
marielle.st-germain@umontreal.ca

RÉSUMÉ | ABSTRACT

On constate depuis quelques années un intérêt de plus en plus marqué pour le Web sémantique et les avantages qu'il présente pour les institutions documentaires. Cependant, un flou persiste quant au potentiel, au fonctionnement et aux étapes de développement liées à des projets qui mettent en œuvre les différents standards et composantes qui le caractérisent. L'objectif du présent article est de proposer une liste d'étapes pour la mise en œuvre d'un projet de données ouvertes et liées (Web de données) en bibliothèque afin de faciliter l'appropriation de ces technologies par les professionnels de l'information.

The Stages in the Development of an Open Data Project Linked to the Library

A marked interest in the Semantic Web and its advantages for libraries has been observed in recent years. However, the potential, the operations and the stages in the development of such projects using different standards and components remains nebulous. This article suggests a list of stages in the implementation of open and linked data (Web data) in the library in order to facilitate the appropriation of these technologies by information professionals.

S'adapter aux évolutions technologiques : une réalité bien connue des professionnels de l'information

L'arrivée du Web, et du numérique par le fait même, a non seulement modifié les comportements informationnels, mais aussi notre façon de consommer et de traiter les connaissances. De plus, la multiplication des formats de documents a eu pour conséquence directe la prolifération des données les décrivant, nécessitant de nouvelles compétences et connaissances de la part des professionnels de l'information (Stuart 2011). On a aussi vu la perception du rôle de bibliothécaire changer : les tâches ont été redéfinies au fil des ans pour laisser plus de place à l'aide à la recherche ainsi qu'à une collaboration étroite entre professionnels de l'information et informaticiens. Les qualités reconnues aux professionnels de l'information telles que le fait d'accorder une importance particulière aux besoins des usagers et au partage de connaissances, la capacité d'identifier et d'organiser des ressources physiques et numériques ainsi que la connaissance des concepts relatifs à la gestion leur ont permis de saisir rapidement les avantages découlant de l'arrivée du Web traditionnel (Rao & Babu 2001 ; Stuart 2011). Malgré les différentes évolutions technologiques auxquelles elles ont dû faire face, la raison d'être et les missions des bibliothèques restent les mêmes : offrir l'accès à une collection de documents pour leur communauté, acquérir des ressources ainsi que produire des services et, finalement, agir comme des intermédiaires entre l'usager et les ressources (Leroux *et al.* 2009). On comprend ainsi aujourd'hui l'importance que peuvent prendre les données relatives aux ressources et l'on saisit le potentiel d'utilisation qu'elles présentent et qui permettrait de répondre à ces missions plus efficacement et en innovant. Le Web sémantique et ses applications viennent répondre à cette nécessité et proposent une nouvelle méthode visant à traiter de façon automatique l'information et les données qui la constituent sur le Web.

Quelques définitions

Pour toute définition complémentaire, nous invitons le lecteur à se référer au glossaire.

Web sémantique

L'expression « Web sémantique » (*Semantic Web*) est attribuée à Berners-Lee, Hendler & Lassila (2001) qui le décrivent non pas comme un Web distinct de celui que l'on connaît, mais plutôt comme une extension de celui-ci. Il s'agit d'un ensemble de technologies dont l'objectif est d'exploiter plus efficacement la quantité importante d'informations que l'on retrouve sur le Web, c'est-à-dire de trouver, partager, réutiliser ou modifier de l'information de manière plus rapide et efficace. Dans cet espace virtuel, l'information est disponible dans une grande variété de langues naturelles et dont le public est humain tandis que, dans le Web sémantique, l'information est exprimée dans un langage destiné à être compris et interprété par des machines (DeWeese & Segal 2015). De grandes quantités de données structurées sont présentement conservées dans des bases de données isolées du Web, que l'on pourrait comparer à des silos d'informations, dont font partie les métadonnées des catalogues de bibliothèques. L'interprétation par les machines permet de mieux gérer l'information de manière automatique, de permettre la création de nouveaux services et applications ainsi que d'assurer l'accessibilité aux connaissances à la fois par l'humain et par la machine. D'ailleurs, dès les balbutiements du Web, on évaluait déjà la possibilité d'organiser l'information afin que celle-ci soit « comprise » par les machines (Berners-Lee, Cailliau, Luotonen, Nielsen & Secret 1994 ; Shadbolt, Berners-Lee & Hall 2006).

Web de données

Le Web de données (*Linked Data*), aussi appelé « données liées », pour sa part, est l'une des composantes du Web sémantique et probablement l'application la plus connue. Afin d'éviter toute confusion entre Web sémantique et Web de données, il est possible de dire que le Web de données est l'une des applications qui entrent dans la grande famille du Web sémantique. Par exemple, les données structurées internes, qui permettent d'ajouter du contenu sémantique à même les pages HTML, sont aussi considérées comme l'une des applications du Web sémantique. Le Web de données doit donc être considéré comme l'ensemble des pratiques et des standards permettant de publier des données structurées sur le Web afin que celles-ci puissent être liées entre elles et interrogées. Il s'agit aussi de promouvoir l'idée selon laquelle les données sont comparables aux documents et qu'il est possible de les relier entre elles de la même façon que les documents sont liés entre eux grâce aux liens hypertextes sur le Web traditionnel (Bermès, Isaac & Poupeau 2013 ; Bizer & Heath 2011). Berners-Lee (2006) a défini quatre principes de base du Web de données sur lesquels la majorité des acteurs du domaine s'appuiera par la suite pour déployer les technologies qui les caractérisent :

1. Nommer les ressources avec des *Uniform Resource Identifier* (URI) ;
2. Utiliser des URI déréférençables (protocole HTTP afin qu'il soit possible d'accéder à des informations sur les ressources ;
3. S'assurer que les URI déréférençables fournissent des informations pertinentes à l'aide des standards tels que *Resource Description Framework* (RDF) et *SPARQL Protocol and RDF Query Language* (SPARQL) ;
4. Créer un réseau de liens avec d'autres URI provenant d'autres bases de données.

Dans le milieu des bibliothèques, on constate la possibilité de jumeler les principes du Web de données à ceux des données ouvertes et, ainsi, non seulement rendre accessibles les données, mais aussi les lier entre elles. C'est ce qu'on appelle les données ouvertes et liées (DOL) (plutôt qu'uniquement « Web de données » ou encore « données liées »). Dans sa note sur les données liées (2006), Berners-Lee a, en 2010, ajouté une échelle de qualité basée sur cinq étoiles qui permet d'évaluer jusqu'à quel point l'information est facilement réutilisable. Les cinq étapes (ou étoiles) sont les suivantes et la figure 1 permet de les illustrer :

1. Rendre vos données disponibles sur le Web (quel que soit leur format) en utilisant une licence ouverte (*Open Licence*) ;
2. Rendre vos données disponibles sous forme de données structurées (par exemple, en format Excel [.xlsx] plutôt que sous forme d'image numérisée d'un tableau) (*Reusable*) ;
3. Utiliser des formats non-propriétaires (par exemple utiliser le comma-separated values [.csv] plutôt que le format Excel [.xlsx]) (Open Format) ;
4. Utiliser des URI pour identifier vos données afin que les autres utilisateurs puissent pointer vers elles (*URI*) ;
5. Relier vos données à d'autres données pour fournir du contexte (*Linked Data*).

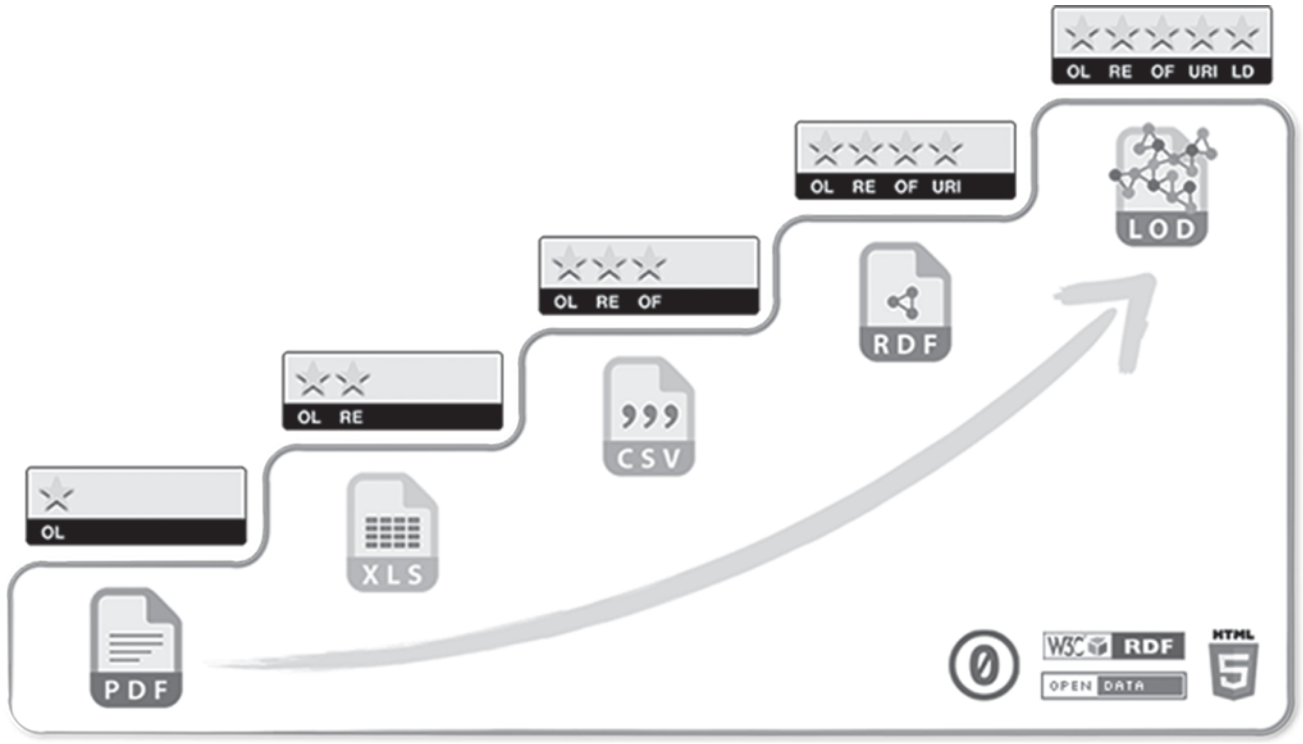
Il s'agit d'un ensemble de technologies dont l'objectif est d'exploiter plus efficacement la quantité importante d'informations que l'on retrouve sur le Web, c'est-à-dire de trouver, partager, réutiliser ou modifier de l'information de manière plus rapide et efficace.

Potentiel des données ouvertes et liées en bibliothèque et impacts

Les professionnels de l'information sont entre autres formés de façon à être aptes à analyser l'information structurée, à effectuer la description de contenus en suivants des standards et des normes et à participer à l'évolution et à la

FIGURE 1

Échelle de qualité des données ouvertes propose par Berners-Lee (2010)[1]



définition de formats informatiques qui permettront de conserver, manipuler et échanger les données (Bermès, Isaac & Poupeau 2013). Il est important de miser sur cette habileté déjà acquise quant à l'encodage des données structurées afin de permettre une évolution des catalogues de bibliothèques. Il est en effet maintenant nécessaire de mettre de plus en plus de l'avant l'aspect d'interopérabilité, non seulement dans le contexte bibliothéconomique, mais aussi dans le contexte infiniment plus large qu'est le Web. Les données ouvertes et liées se présentent comme l'une des solutions envisagées pour répondre aux nouveaux défis associés aux évolutions technologiques.

D'abord, les données ouvertes et liées impliquent la réutilisation de données par les moteurs de recherche, donc une meilleure visibilité. En effet, les ressources des catalogues de bibliothèques seraient ainsi disponibles pour les usagers du Web, ce qui ferait en sorte que l'expertise des bibliothécaires et professionnels de l'information pourrait être mise de l'avant grâce à la qualité et la fiabilité de ces ressources et de leurs descriptions (Coyle 2010 ; Gonzales 2014). Étant donné que l'accès aux ressources dans les bibliothèques numériques dépend des métadonnées qui les décrivent, le Web de données permettrait de naviguer d'une ressource à l'autre, même si ces ressources sont externes au catalogue de bibliothèque. Contrairement aux catalogues en ligne qui sont basés sur le fait d'effectuer une recherche textuelle pour obtenir des résultats (par exemple, avec le nom d'un auteur ou le titre d'une ressource), le Web de données ouvre la porte à des recherches plus larges et qui peuvent être basées sur des concepts plutôt que sur des informations précises. En effet, l'usager pourrait faire des découvertes inattendues et naviguer d'une bibliothèque numérique à une autre ou encore d'une bibliothèque numérique à des sources externes telles que Wikipédia, l'exposition virtuelle d'un musée, une base de données thématique, etc. On constate ainsi du même coup la possibilité d'effectuer des recherches fédérées ainsi que des recherches interdisciplinaires plus performantes, qui permettraient d'accéder à l'information provenant d'un plus grand nombre de bases de données ouvertes.

Ensuite, contrairement à la pratique établie où les données sont partagées sous la forme de notices, il s'agit donc ici de la possibilité de créer des graphes sur des ressources données dont les informations proviennent de sources différentes. Par le fait même, on constate la possibilité de diminuer la redondance en ce qui a trait aux descriptions bibliographiques et aux métadonnées. Il s'agit ainsi d'une occasion d'optimiser les coûts liés aux activités de description des fonds. Conséquemment, les bibliothèques peuvent partager entre elles les données qui sont déjà accessibles

1. Source : <5stardata.info/en>.

par l'intermédiaire de sources fiables et ainsi éviter la duplication du travail (Alemu, Stevens, Ross & Chandler 2012 ; Tillett 2013).

Le fait d'avoir un format universel pour le partage de données présente des avantages, car celui-ci permettrait une meilleure interopérabilité d'une institution à une autre, mais aussi d'une institution à d'autres instances publiant leurs données, et ce, peu importe le système utilisé. Les technologies développées pour la gestion des données en bibliothèques sont présentement produites par les fabricants de systèmes intégrés de gestion de bibliothèque (SIGB) et les normes bibliographiques sont conçues uniquement pour la communauté des professionnels de l'information, ce qui limite les possibilités de partages et contribue largement à l'isolement de leurs données.

Dans le monde, plusieurs bibliothèques nationales ont déjà incorporé les technologies relatives aux données liées à leurs catalogues […].

Finalement, Baker *et al*. (2011) soulignent aussi qu'étant donné que les données liées sont caractérisées par le fait qu'elles décrivent la sémantique, et ce, indépendamment des formats et de la syntaxe, même si un nouveau format voit le jour, les données garderont leur signification et sont par conséquent plus pérennes.

Dans le monde, plusieurs bibliothèques nationales ont déjà incorporé les technologies relatives aux données liées à leurs catalogues, soit la Bibliothèque nationale de France (BnF) avec son projet dédié[2], la Bibliothèque nationale d'Allemagne[3], la Library of Congress[4], la Bibliothèque nationale d'Espagne[5] et la Bibliothèque nationale du Royaume-Uni[6]. Au Québec et au Canada, notons entre autres l'existence de l'Initiative canadienne sur les données liées[7] (ICDL), mise sur pied en 2015 à la suite de la constatation que certaines institutions américaines et européennes ont déjà une longueur d'avance dans le développement de projets de données liées en bibliothèques. L'ICDL est née d'une collaboration entre l'Université de Toronto, l'Université McGill, l'Université de Montréal, l'Université de l'Alberta et l'Université de la Colombie-Britannique, et est menée en partenariat avec Bibliothèque et Archives Canada (BAC) et Bibliothèque et Archives nationales du Québec (BAnQ). En octobre 2016, le Sommet canadien sur les données liées[8] a eu lieu à Montréal. Celui-ci visait à sensibiliser différents acteurs du domaine à l'importance des données liées en misant sur le partage d'expériences et à permettre aux membres de se rencontrer. L'ICDL est composée de différents groupes de travail, dont le Groupe de travail francophone qui a pour mission d'étudier les défis que pose l'adoption des technologies du Web sémantique en milieu francophone.

De plus, quelques projets ont été développés ou sont en cours de développement dans le but de démontrer que les données ouvertes liées représentent une bonne option pour le partage de données et de ressources d'une institution à l'autre. Prenons par exemple le projet Au-delà des tranchées[9] mené par le Réseau pancanadien du patrimoine documentaire (RPCPD), le dépôt des notices de BAnQ dans le fichier d'autorité international virtuel (VIAF), 150 années d'art canadien[10] mené par Artefacts Canada et quelques projets toujours en cours de développement liés au Plan culturel numérique du Québec[11]. Bien que l'on constate un intérêt de plus en plus marqué au sein de différentes institutions et sur le plan de la recherche, il reste du travail à effectuer et il est nécessaire d'encourager la diffusion des connaissances, le partage de compétences et le retour sur les expériences vécues.

Il est aussi nécessaire de mentionner le modèle de données pour la description bibliographique *Bibliographic Framework*[12] (BIBFRAME) initié par la Library of Congress, développé dans l'objectif clair de remplacer les formats MARC et de fournir un format qui serait compatible avec le Web de données. En constante évolution, il est basé sur RDF, sur les principes des données liées et sur trois niveaux d'abstraction : l'œuvre qui reflète l'essence conceptuelle de la ressource à cataloguer, l'instance qui est la réalisation matérielle d'une œuvre et l'item qui est la copie physique ou électronique de l'instance. Comme on peut le constater, le modèle n'est pas sans rappeler un autre modèle, le *Functional Requirements for Bibliographic Records* (FRBR). Il pourrait possiblement se présenter comme une solution à l'implémentation de projets de Web de données dans les institutions documentaires par le fait qu'il a été réfléchi expressément pour répondre aux besoins des bibliothèques.

2. <data.bnf.fr/>.
3. <www.dnb.de/EN/Service/DigitaleDienste/LinkedData/linkeddata_node.html>.
4. <id.loc.gov/>.
5. <datahub.io/dataset/datos-bne-es>.
6. <datahub.io/dataset/bluk-bnb>.
7. <connect.library.utoronto.ca/display/U5LD/Canadian+Linked+Data+Initiative+Home>.
8. <www.mcgill.ca/clds/fr/propos>.
9. <www.canadiana.ca/rpcpd-dol>.
10. <chin-rcip.canadiana.ca/aclod/?lang=fr>.
11. <culturenumerique.mcc.gouv.qc.ca/>.
12. <www.loc.gov/bibframe/docs/bibframe2-model.html>.

Étapes pour la mise sur pied d'un projet de données ouvertes et liées

Parmi la littérature disponible sur le Web de données et le Web sémantique, on retrouve quelques recommandations et méthodologies sur lesquelles se baser afin de publier des jeux de données en respectant les normes et standards (notamment Bizer & Heath 2011 ; Van Hooland & Verborgh 2014 ; W3C 2014 ; Zengenene, Casarosa & Meghini 2014). Cependant, il est difficile d'obtenir une version claire et condensée d'une marche à suivre afin d'appliquer ces technologies aux données bibliographiques. Nous proposons ici les étapes nécessaires à la mise en place d'un projet de Web de données. La formulation de ces étapes est le résultat d'une synthèse effectuée à la suite d'une revue de la littérature approfondie. Les informations étant disséminées à travers différents textes, un travail d'analyse, de regroupement, de classement et de fédération de celles-ci a été effectué. Le but de l'exercice est de proposer un processus clair de publication dans le Web de données en se basant sur les standards préconisés par le World Wide Web Consortium (W3C) afin de faciliter l'appropriation des technologies du Web sémantique par les professionnels de l'information. Les deux premières étapes ne sont pas spécifiques à un projet de Web de données et constituent plutôt des exigences liées au facteur humain. En effet, elles pourraient s'appliquer à n'importe quelle gestion de projet d'implémentation nouvelle. Cependant, elles sont tout de même primordiales à son fonctionnement et à son acceptation au sein de l'institution.

Comprendre la motivation et obtenir l'autorisation des parties prenantes

Avant de commencer la mise en place d'un tel projet, il est nécessaire que tous les acteurs impliqués puissent avoir une vision globale de ce que signifient les données ouvertes et liées et les technologies qui s'y rattachent, mais aussi les avantages et défis qui en découlent. Ce type d'initiative permet souvent de répondre à un besoin spécifique en lien avec un jeu de données pertinent. L'importance réside dans la compréhension du rôle des sciences de l'information dans cette nouvelle extension du Web, dans l'amélioration de l'expérience de l'usager, mais aussi dans le perfectionnement du processus de traitement de l'information en général (Zengenene, Casarosa & Meghini 2014). La vulgarisation ainsi que la capacité à communiquer les bénéfices pour les institutions sont primordiales afin de permettre aux instances de s'approprier les technologies du Web de données et de constater les bénéfices à courts, moyens et longs termes. L'un des défis les plus importants est le fait qu'il est, pour le moment, difficile, voire impossible, d'évaluer les retombées économiques d'un tel projet et même d'en mesurer l'efficacité de manière concrète. Ainsi, il est nécessaire de se concentrer sur les autres avantages tels que l'augmentation du nombre de visiteurs sur les plateformes Web de l'institution, la présence et la visibilité des données dans les résultats présentés par les moteurs de recherche, l'amélioration des services aux usages, les possibilités de réutilisation des données ouvertes permettant la création de nouvelles applications et la participation à un mouvement collaboratif d'envergure internationale visant, entre autres, l'évolution du catalogue de bibliothèque.

Une fois les différentes instances préparées, le projet présenté de façon claire et simple et le document rédigé, il devient plus facile de communiquer les objectifs et d'évaluer les ressources nécessaires. Pour obtenir l'autorisation des parties prenantes, il est aussi possible de présenter des exemples de projets de publication de jeux de données dans le Web de données qui permettront de constater concrètement ce que ces applications peuvent apporter comme nouveaux services et nouvelles applications. Afin de rendre possible une association aux approches traditionnelles de gestion de l'information, il est recommandé de présenter des modèles identifiants le cycle de vie des données ouvertes et liées.

Établir une licence d'utilisation

Il est primordial de statuer sur les droits d'utilisation des données publiées en indiquant le propriétaire des données et dans quelle mesure il est possible de les réutiliser, surtout lorsqu'il s'agit de données provenant d'institutions culturelles publiques (Bermès 2013 ; Bizer & Heath 2011 ; Hyvönen 2012 ; Villazón-Terrazas, Vilches-Blázquez, Corcho & Gómez-Pérez 2011 ; W3C 2014 ; Zengenene, Casarosa & Meghini 2014). Les usagers et les développeurs auront d'ailleurs plus tendance à s'approprier les données si une licence présente les spécificités légales de manière claire et précise.

De plus en plus, on constate l'utilisation du modèle de licences *Creative Commons*[13]. Par exemple, Données Québec[14], la Bibliothèque nationale d'Espagne, la Bibliothèque nationale d'Allemagne et plusieurs autres utilisent les licences *Creative Commons* pour encadrer légalement la réutilisation de leurs données. Il existe plusieurs autres types de licences disponibles. On soulignera au passage le projet RightsStatements.org[15], une initiative de la Digital Public Library of America[16] (DPLA)et de Europeana[17], qui propose un ensemble d'énoncés de droits normalisés réfléchis spécifiquement pour les institutions du patrimoine culturel. Ceux-ci ont été conçus en gardant en tête à la fois

13. <creativecommons.org>.

14. <www.donneesquebec.ca/fr/licence>.

15. <rightsstatements.org/en/>.

16. <dp.la/>.

17. <www.europeana.eu/portal/en>.

les usagers et les machines et déploient les technologies du Web sémantique.

Lorsque vient le temps de choisir une licence, il faut tenir compte de trois concepts différents, soit l'attribution, la restriction de l'usage commercial et la redistribution à l'identique (Bermès, Isaac & Poupeau 2013). Les clauses d'une licence liées à l'attribution (« *by* ») sont celles qui précisent comment doit être reconnu le détenteur de la propriété intellectuelle de la ressource. Ainsi, dans certains cas, l'usager peut être dans l'obligation de citer la source des données dans le contexte d'une réutilisation. La restriction de l'usage commercial (« *nc* ») implique que l'utilisateur ne peut réutiliser des données à des fins commerciales. Une licence peut aussi préciser les différentes possibilités pour l'usage commercial. Finalement, la redistribution à l'identique (« *share-alike* ») précise si l'usager peut ou non réutiliser et transmettre les données en utilisant une licence semblable à celle utilisée par l'institution.

Évaluer les compétences

À cette étape, il est nécessaire de commencer à évaluer quel processus de conversion sera le plus pertinent pour les jeux de données à publier. Pour ce faire, une analyse de la situation qui présente les compétences des différents acteurs impliqués dans le processus doit être effectuée. Selon Zengenene, Casarosa & Meghini (2014), les compétences nécessaires pour les bibliothécaires et autres professionnels de l'information sont triples :

- Systèmes d'information :
  - Avoir des connaissances en ce qui a trait au téléchargement, à l'installation et à la configuration de systèmes d'information, de bases de données et de serveurs ;
  - Être aptes à écrire et lire les formats XML et RDF.
- Métadonnées :
  - Connaître le processus de catalogage, l'importance et la signification des métadonnées.
- Modélisation :
  - Comprendre la structure des données et être aptes à évaluer la meilleure façon de convertir les données d'une structure donnée vers RDF.

[…] la place des professionnels de l'information au sein de l'équipe de développement est absolument nécessaire. Les compétences doivent donc être hétérogènes, diversifiées et bien développées.

Ensuite, dans la mesure où l'institution se voit dans l'obligation d'engager des techniciens à l'externe (développeurs de logiciels, par exemple), les bibliothécaires et professionnels de l'information doivent être aptes à communiquer leurs besoins et à décrire les ontologies et vocabulaires qu'ils souhaitent utiliser pour décrire leurs données. Mettre sur pied une équipe dont l'objectif sera de mettre en place un projet de Web de données peut s'avérer particulièrement ardu. En effet, on constate d'un côté le besoin de développeurs de logiciels et de programmeurs qui seront aptes à construire l'architecture, mais aussi la nécessité d'avoir au sein de l'équipe des individus qui maîtrisent parfaitement bien le contenu qui doit être manipulé. Ainsi, la place des professionnels de l'information au sein de l'équipe de développement est absolument nécessaire. Les compétences doivent donc être hétérogènes, diversifiées et bien développées.

Évaluer les jeux de données

En ce qui a trait à l'identification des jeux de données à publier, on doit analyser lesquels on souhaite rendre disponibles, les formats dans lesquels ils sont encodés ainsi que les technologies nécessaires à chaque situation. On recommande en général d'accorder la priorité à la publication de données qui sont uniques à l'institution et qui présentent un intérêt de réutilisation par des usagers. Trois types de données sont à considérer (Bermès, Isaac & Poupeau 2013) :

- Notices se trouvant dans le catalogue ou la base de données ;
- Référentiels ou notices d'autorité ;
- Éléments de métadonnées ou ontologies développées.

Les référentiels ou les données d'autorité présentent pour leur part un avantage intéressant, dans la mesure où ils sont souvent le résultat d'efforts importants de normalisation et offrent une grande possibilité de réutilisation par d'autres institutions. Par exemple, comme mentionné précédemment, les thésaurus et listes d'autorité RAMEAU (BnF) et Library of Congress Subject Headings (LCSH) ont été rapidement publiés sur le Web de données. Pour ce qui est de la troisième option, soit les ontologies ou les éléments de métadonnées développés au sein de l'institution, il est plutôt recommandé d'utiliser des ontologies existantes plutôt que de procéder à la création de celles-ci afin d'assurer une meilleure interopérabilité. Il peut cependant arriver que certaines institutions ressentent le besoin de créer leur ontologie lorsque leurs données présentent des spécificités particulières. Finalement, il est intéressant d'envisager la publication de données administratives et techniques afin de s'inscrire dans le mouvement des données ouvertes, telles que des données de nature statistique et transactionnelle. Dans tous les cas, il est important de statuer quant au choix des jeux de données à publier, car cela facilitera les tâches suivantes et l'analyse du travail qui sera à effectuer.

Choisir le modèle de publication et évaluer les outils nécessaires

Cette étape est importante dans la mesure où elle permet d'avoir une idée d'ensemble du processus de publication des données. D'abord, si les documents sont en format texte en langue naturelle (par exemple, des rapports ou des articles), il est possible d'utiliser un extracteur d'entités nommées et d'annotations Web sémantique, qui permettra la reconnaissance et l'extraction d'informations dans un corpus donné ainsi que l'annotation en RDF (Bizer & Heath 2011). Le fonctionnement de ce type d'extracteur est le suivant : les documents sont annotés par l'extracteur à l'aide des URI des entités nommées dans le document. La publication de ces documents accompagnés de ces annotations augmente les chances de découverte et permet aux applications du Web sémantique d'effectuer des liens vers ceux-ci. Notons cependant que les extracteurs d'entités nommées ne permettront pas d'extraire les relations entre les entités. Ainsi, on aura la possibilité de définir les URI, mais non les triplets RDF. Un travail supplémentaire est donc nécessaire afin d'obtenir des fichiers RDF qui pourront être versés sur un serveur Web. Cette façon de faire est la plus simple et est principalement utilisée lorsque la quantité de fichiers est limitée, car le travail se fait principalement de façon manuelle.

Ensuite, dans la mesure où les données sont déjà structurées, ce qui est le cas dans la plupart des situations liées au domaine des sciences de l'information, tout dépend alors de la façon dont les données sont stockées. Ainsi, si les données sont stockées dans une base de données relationnelle, deux technologies de conversion peuvent être utilisées : RDB-to-RDF[18] ou l'utilisation d'un système de gestion de contenu qui permet d'exprimer les données en RDFa. Un système de gestion de contenu permet la création et la gestion de contenu numérique via une interface conviviale. Ainsi, certains outils comme Drupal[19], lui-même un système de gestion de contenu, permettent donc de transformer les données structurées qui se trouvent dans sa base de données en RDFa en ajoutant des attributs RDF aux éléments HTML. Mentionnons aussi l'existence des langages de mise en correspondance de données qui permettent de transformer les données provenant de bases de données relationnelles en RDF[20]. Lorsque les données sont encodées dans des bases de données relationnelles, il est en général préférable de ne pas procéder à un transfert des données vers un *triplestore* afin de conserver l'infrastructure de gestion de l'information déjà mise en place (Bizer & Heath 2011).

Il est aussi possible d'accéder aux données stockées via une interface de programmation applicative (API). Dans ce cas-ci, la situation est plus complexe, dans la mesure où l'institution devra développer un adaptateur personnalisé (*wrapper*). Un adaptateur permettra de régler certaines limites qui sont associées aux API, comme, par exemple, le fait que leur contenu ne peut être repéré par les moteurs de recherche. En général, les adaptateurs permettent d'assigner des URI aux ressources sur lesquelles l'API fournit des données, de reformuler une requête pour que celle-ci soit comprise par l'API et de transformer les résultats de cette requête en RDF (Bizer & Heath 2011). Le développement d'un tel outil implique des connaissances informatiques qui dépassent en général celles déjà acquises par les professionnels de l'information, donc l'appel à une ressource externe peut alors être nécessaire.

Finalement, les données structurées peuvent aussi être stockées dans des *triplestores*. La plupart des *triplestores* offrent une interface de données liées, ce qui peut faciliter grandement le travail de programmation. On peut donc accéder directement à un point d'accès SPARQL et il est possible d'effectuer la configuration et la gestion du contenu à même le *triplestore*.

Attribuer les URI

Il est nécessaire de porter une attention particulière à la création d'URI pour décrire les ressources et les relations qui les unissent. Cette étape permet aussi de constater la quantité d'entités avec lesquelles on devra travailler et qu'il sera nécessaire d'identifier. Les URI doivent être construits de façon simple, stable, pérenne et gérable (Berners-Lee 2008 ; Villazón-Terrazas, Vilches-Blázquez, Corcho & Gómez-Pérez 2011).

On devra d'abord choisir un nom de domaine qui ne sera pas susceptible de changer au fil du temps. Ensuite, pour ce qui est de la description des ressources, il est préférable d'utiliser des identifiants déjà existants, tels que les référentiels ou les listes d'autorité, qui ont démontré leur persistance dans le temps. Par exemple, un ISBN pour un livre, des codes de bibliothèque ou, comme c'est le cas pour la Bibliothèque nationale de France, le numéro de notice associée à la ressource (Bermès, Isaac & Poupeau 2013 ; Villazón-Terrazas, Vilches-Blázquez, Corcho & Gómez-Pérez 2011 ; Zengenene, Casarosa & Meghini 2014). Dans la mesure où aucun identifiant n'existe, l'institution devra en créer de nouveaux et deux options s'offrent alors à elle. L'utilisation d'identifiants opaques, qui n'ont aucune signification particulière, ou d'identifiants signifiants, qui présentent une information lisible et compréhensible par l'humain.

Il est important de s'assurer de respecter le concept de négociation de contenu et de procéder à la création d'au moins trois URI pour identifier la même ressource (ou ses différentes représentations). Ainsi, il est nécessaire d'accorder une attention particulière à l'origine de l'identifiant (référentiel ou liste d'autorité), à la forme de l'identifiant (opaque ou

18. *Relational database to RDF*.

19. <www.drupal.org>.

20. D2RQ ou R2RML, par exemple.

signifiant) et à la création d'au moins trois identifiants pour une même ressource (un pour le concept, un pour la représentation HTML et un pour la représentation RDF).

Choisir son modèle de données et effectuer la mise en correspondance et le nettoyage des données

Cette étape, qui consiste à développer la structure sémantique, peut présenter certaines difficultés, surtout dans la mesure où les métadonnées sont encodées de manières différentes, dans le cas où, par exemple, elles proviendraient de catalogues différents. Un modèle de données découle lui-même de la modélisation des données, c'est-à-dire le fait d'analyser et de concevoir la structure des données qui sont contenues dans un système, mais aussi leur sémantique. D'ailleurs, la modélisation des données dépendra du mode de publication choisi et de leur nature. Ainsi, cette étape implique les choix relatifs aux classes et aux propriétés qui seront sélectionnées pour décrire les entités et les relations qui les unissent. Pour ce faire, il est donc important de bien connaître le niveau de description et les éléments de métadonnées utilisés pour, par la suite, effectuer une mise en correspondance entre les jeux de données. La plupart du temps, l'institution devra utiliser plusieurs ontologies et vocabulaires pour représenter son information.

Le choix de ces jeux de données dépend de plusieurs facteurs, tels que la valeur ajoutée que ceux-ci peuvent apporter aux données de l'institution, mais aussi la visibilité que ces liens peuvent apporter.

La difficulté réside aussi dans le fait que le modèle de données doit permettre de représenter toutes les entités et leurs relations, tout en assurant une logique durant le processus. Le choix des ontologies et des vocabulaires est donc l'une des parties les plus importantes du processus de publication de données ouvertes et liées sur le Web et l'on encourage la réutilisation de standards. Notons que le choix dépend grandement des objectifs de l'institution ainsi que du type de données qu'elle souhaite publier. De plus, il est fort probable que la mise en correspondance des données ait comme conséquence une perte de granularité au niveau des descriptions. Cet aspect peut décevoir, mais il est à noter que les données qui seront publiées dans le Web de données n'ont pas comme objectif de remplacer les notices bibliographiques qui sont souvent plus précises, mais plutôt d'apporter une couche supplémentaire d'interopérabilité et une possibilité de réutilisation.

Ensuite, une fois le ou les vocabulaires et ontologies choisis, et ainsi, le modèle de données défini, il est nécessaire de procéder à la mise en correspondance des données. Il s'agit d'une sous-étape qui peut donner du fil à retordre dans la mesure où elle nécessite une bonne maîtrise des technologies qui seront utilisées, du vocabulaire ou de l'ontologie et des formats. En général, on commence par la création d'un tableau de conversion qui permettra d'indiquer la source, la cible, la règle de conversion et, si possible, un exemple. Il est possible de créer la conversion selon différents niveaux de précision et de détail, le choix revenant à l'institution d'évaluer les différentes possibilités. Puis, on doit par la suite transformer les données en fonction des règles de conversion. Pour ce faire, on utilise généralement un programme informatique et le choix de cet outil dépend du format source.

Avant de convertir les données en RDF, il est nécessaire de s'assurer qu'on y retrouve le moins d'erreurs possible et qu'on en augmente la qualité. On pense à des erreurs typographiques, des valeurs manquantes, des doublons ou des contradictions (Rahm & Do 2000 ; Van Hooland & Verborgh 2014). Par exemple, il pourrait y avoir des informations manquantes en ce qui a trait à une notice bibliographique (date de publication, numéro de l'édition, etc.) ou la forme choisie pour le nom de l'auteur pourrait ne pas correspondre à la notice d'autorité. Il est évident que le résultat ne sera pas parfait, mais il est toujours préférable de faire quelques vérifications avant de procéder à la conversion. Il est donc nécessaire de s'assurer que les données sont dans un format structuré et l'on peut par la suite utiliser des outils tels que OpenRefine[21] ou DataWrangler[22] pour procéder au nettoyage.

Enrichir les données en créant des liens et convertir les données en RDF

Cette étape est primordiale et a comme objectifs de définir des triplets qui seront connectés à d'autres et de créer des triplets qui permettront de définir des relations, le tout à l'interne et à l'externe (Zengenene, Casarosa & Meghini 2014). Par exemple, il est possible de lier l'URI d'un auteur à d'autres notices d'autorité d'institutions reconnues comme la Library of Congress ou le Virtual International Authority File (VIAF). Lorsqu'on crée des liens au sein même du jeu de données, il est nécessaire de s'assurer que toutes les ressources seront connectées entre elles. Pour ce qui est des liens vers l'externe, il est préférable de se lier à des jeux de données qui sont fiables, grandement utilisés, stables et bien établis. C'est le cas de DBpedia[23], Geonames[24], VIAF et Library of Congress. Le choix de ces jeux de données dépend de plusieurs

21. <openrefine.org>.

22. <vis.stanford.edu/wrangler/>.

23. <wiki.dbpedia.org/>.

24. <www.geonames.org/ontology/documentation.html>.

TABLEAU 1

Points à vérifier avant de procéder à la publication des données sur le Web de données

Points à vérifier	☑
Est-ce que le jeu de données est lié à d'autres jeux de données ?	
Avez-vous rendu disponible la provenance des métadonnées ?	
Avez-vous rendu disponible une licence d'utilisation ?	
Utilisez-vous des termes provenant de vocabulaires ou d'ontologies bien établis ?	
Les URI sont-ils déréférençables ?	
Avez-vous effectué la mise en correspondance des données entre les termes de votre vocabulaire ou ontologie avec d'autres vocabulaires ou ontologies, si applicable ?	
Avez-vous rendu disponible une description des jeux de données ?	

facteurs, tels que la valeur ajoutée que ceux-ci peuvent apporter aux données de l'institution, mais aussi la visibilité que ces liens peuvent apporter. Notons la nécessité de prendre connaissance de la licence d'utilisation de chacune des entités externes vers lesquelles on souhaite effectuer des liens et, ainsi, réutiliser les données. Selon Bermès, Isaac & Poupeau (2013), la création de liens implique trois tâches liées, soit :

- identifier quels sont les points de contact entre ses propres jeux de données et ceux de l'externe ;
- identifier les liens qui unissent les entités de l'externe et celles de son jeu de données ;
- évaluer quelle est la meilleure méthode pour effectuer ces liens, soit de façon manuelle ou automatique.

Le stade de la conversion des données vers RDF ne pose en général pas trop de difficultés. Il est cependant nécessaire de choisir la ou les syntaxes de sérialisation dans lesquelles les données seront rendues disponibles pour les usagers qui voudraient peut-être les réutiliser, mais surtout pour les machines. Il est important de savoir que la procédure doit se faire de façon automatique ou semi-automatique étant donné que d'effectuer un tel travail manuellement serait beaucoup trop fastidieux. Il existe un bon nombre d'outils qui permettent la conversion.

Valider et publier les jeux de données

Avant de publier les jeux de données sur le Web, il est recommandé d'en vérifier la qualité. Bizer & Heath (2011) proposent une liste de points à vérifier avant de procéder à la publication. Le Tableau 1 propose une liste de vérification.

À cette étape, il est aussi pertinent de vérifier si les triplets RDF sont bien construits à l'aide d'outils de validation et de relire l'échelle de qualité basée sur les cinq étoiles présentée précédemment.

Une fois les données converties et évaluées, il est possible de les rendre accessibles sous forme de fichiers RDF, dans un *triplestore* présentant une interface d'interrogation SPARQL, via un point d'accès SPARQL, via un RESTful API (*Representational State Transfer*) ou encore directement dans les pages Web grâce aux données structurées internes.

Conclusion

L'objectif de cet article était de proposer une série d'étapes afin de faciliter l'implantation d'un projet de données ouvertes et liées en bibliothèque. On constate que le Web de données permet de jeter un regard nouveau sur les occasions d'innover qui se présentent aux professionnels de l'information. Les comportements et les besoins informationnels ayant grandement changé depuis l'arrivée du Web, les défis sont nombreux et les bibliothèques et institutions documentaires doivent s'adapter tout en travaillant au processus d'évolution du catalogue. Ce nouvel environnement potentiel présente de nombreux avantages. Cependant, le déploiement de ces technologies n'est pas sans défi. En effet, de tels changements impliquent de nombreuses modifications quant aux habitudes de travail et un temps non négligeable quant à la formation requise pour mettre en pratique ces connaissances. Ainsi, les ressources nécessaires sont importantes et une réflexion s'impose de la part des décideurs et des instances gouvernementales. On peut cependant croire que les projets de Web de données prendront de plus en plus d'ampleur lorsque les pratiques auront été établies de manière plus claire et qu'un plus grand nombre de projets concrets auront vu le jour.

Glossaire

***Application Programming Interface* (API) :** Interface de programmation qui permet d'accéder à l'ensemble de règles qui composent une application. Un développeur de logiciels peut ainsi y accéder et procéder à l'échange de données.

Données ouvertes : Données diffusées de manière structurée et en libre accès selon une licence d'utilisation permettant leur réutilisation par tous.

Données ouvertes et liées : Concept qui jumèle les principes du Web de données aux principes des données ouvertes.

Données structurées internes : Contenu sémantique ajouté à même les pages HTML afin d'en faciliter le repérage par les moteurs de recherche. Il s'agit de l'une des applications du Web sémantique.

Licence d'utilisation : Contrat dans lequel sont stipulées les conditions d'utilisation, de modification et de diffusion des jeux de données.

Métadonnée : Donnée normée présentant de l'information précise à propos d'une autre donnée.

Modèle de données : Modèle visant à définir la structure et la sémantique des données contenues dans un système.

Négociation de contenu : Mécanisme qui permet de proposer pour une même ressource différentes visualisations en fonction de la provenance de la requête.

Ontologie : Ensemble structuré visant à définir les concepts et les relations afin de classifier l'utilisation de certains termes dans une application donnée, caractériser les relations entre les ressources et clarifier les contraintes quant à l'utilisation d'un terme.

Point d'accès SPARQL : Service qui respecte le protocole SPARQL et qui présente la possibilité d'effectuer des requêtes dans un *triplestore*.

RDFa : Recommandation du W3C définissant une méthode permettant l'ajout de données structurées dans une page HTML ou un document XML.

***Resource Description Framework* (RDF) :** Modèle de graphe ayant comme objectif de décrire les ressources et leurs métadonnées. Un document RDF est basé sur la formation de triplets.

***SPARQL Protocol and RDF Query Langage* (SPARQL) :** Langage de requête permettant d'effectuer des recherches au sein de données encodées en RDF, de les récupérer et de les manipuler.

Syntaxe de sérialisation : Syntaxe permettant l'encodage de données et le partage de celles-ci entre les systèmes.

Triplestore : Base de données conçue pour stocker et récupérer des données en RDF. On retrouve donc dans un *triplestore* uniquement des triplets.

Triplet RDF : Énoncé créé à partir d'un sujet (ressource décrite), d'une prédicat (propriété de la ressource ou relation) et d'un objet (donnée ou autre ressource).

***Uniform Resource Identifier* (URI) :** Suite alphanumérique qui identifie de manière univoque une ressource physique ou abstraite.

URI déréférençable : URI qui se définit principalement par le fait qu'il met en pratique le mécanisme de négociation de contenu.

Vocabulaire : Schéma permettant l'organisation des connaissances de manière standardisée afin de faciliter la recherche d'information.

Web 2.0 : Mouvement découlant du World Wide Web misant sur la participation des usagers, la création de contenu et l'interopérabilité.

Web de données : Application du Web sémantique visant la publication de données en format RDF afin de permettre la création de liens entre celles-ci et leur compréhension par la machine. On parle aussi de « données liées ».

Web sémantique : Extension du Web tel qu'on le connaît basée sur des standards définis par le W3C qui vise la publication de données dans des formats standardisés afin de les exploiter de manière plus efficace.

World Wide Web Consortium (W3C) : Organisme de standardisation pour le Web.

SOURCES CONSULTÉES

Alemu, Getaneh, Brett Stevens, Penny Ross & Jane Chandler. 2012. Linked data for libraries : Benefits of a conceptual shift from library-specific record structures to RDF-based data models. *New Library World* 113 (11/12) : 549-570.

Baker, Thomas *et al*. 2011. Library linked data incubator group final report. Consulté le 13 juillet 2017. <www.w3.org/2005/Incubator/lld/XGR-lld-20111025/>.

Bermès, Emmanuelle. 2013. Enabling your catalogue for the semantic web. In *The future of the library catalogue*, sous la direction de Sally Chambers. Chicago : Neal-Schuman, an imprint of the American Library Association, 117-142.

Bermès, Emmanuelle, Antoine Isaac & Gautier Poupeau. 2013. *Le web sémantique en bibliothèque*. Paris : Électre/Éditions du Cercle de la Librairie.

Berners-Lee, Tim. 2006. Linked data. Consulté le 13 juillet 2017. <www.w3.org/DesignIssues/LinkedData.html>.

Berners-Lee, Tim. 2010. Is your linked open data 5 star ? Consulté le 13 juillet 2017. <www.w3.org/DesignIssues/LinkedData.html>.

Berners-Lee, Tim, Robert Cailliau, Ari Luotonen, Henri Frystyjk Nielsen & Arthur Secret. 1994. The world-wide web. *Communications of the ACM* 37 (7) : 76-82.

Berners-Lee, Tim, Jim Hendler & Ora Lassila. 2001. *The semantic web*. Scientific American 284 (5) : 28-37.

Bizer, Christian & Tom Heath. 2011. *Linked data : Evolving the web into a global data space*. San Rafael, Californie : Morgan & Claypool.

Coyle, Karen. 2010. Library data in a modern context. *Library Technology Reports* 46 (1) : 5-13.

DeWeese, Keith P. & Dan Segal. 2015. *Libraries and the semantic web*. San Rafael, Californie : Morgan & Claypool.

Gonzales, Brighid M. 2014. Linking libraries to the web : Linked data and the future of the bibliographic record. *Information Technology and Libraries* 33 (4) : 10-22.

Hyvönen, Eero. 2012. *Publishing and using cultural heritage linked data on the semantic web*. San Rafael, Californie : Morgan & Claypool.

Leroux, Éric *et al.* 2009. Les professions et les institutions. In *Introduction aux sciences de l'information*, sous la direction de Jean-Michel Salaün et Clément Arsenault. Montréal : Presses de l'Université de Montréal, 16-52.

Rahm, Erhard & Hong Hai Do. 2000. Data cleaning : Problems and current approaches. *IEEE Data Engineering Bulletin* 23 (4) : 3-13.

Rao, K. Nageswara & K. H. Babu. 2001. Role of librarian in internet and world wide web environment. *Information science* 4 (1) : 25-34.

Shadbolt, Nigel, Tim Berners-Lee & Wendy Hall. 2006. The semantic web revisited. *Intelligent Systems, IEEE* 21 (3) : 96-101.

Stuart, David. 2011. *Facilitating access to the web of data : A guide for librarians*. Londres : Facet.

Tillett, Barbara. 2013. RDA and the semantic web, linked data environment. *JLIS.it* 4 (1) : 139-145.

Van Hooland, Seth & Ruben Verborgh. 2014. *Linked data for libraries, archives and museums : How to clean, link and publish your metadata*. Chicago : Neal-Schuman, an imprint of the American Library Association.

Villazón-Terrazas, Boris, Luis Vilches-Blázquez, Oscar Corcho & Asunción Gómez-Pérez. 2011. Methodological guidelines for publishing government linked data. In *Linking government data*, sous la direction de David Wood. New York : Springer, 27-49.

W3C. 2014. Best practices for publishing linked data. Consulté le 13 juillet 2017. <www.w3.org/TR/ld-bp/>.

Zengenene, Dydimus, Vittore Casarosa & Carlo Meghini. 2014. Towards a methodology for publishing library linked data. In *Bridging between cultural heritage institutions*, sous la direction de Tiziana Catarci, Nicola Ferro & Antonella Poggi. Berlin : Springer, 81-92.