La Deutsche Nationalbibliografie en données ouvertes liées [linked open data] : applications et perspectives

Jürgen Kett

Sarah Beyer

Mathias Manecke

Yvonne Jahns and

Lars G. Svensson
Deutsche Nationalbibliothek
Frankfurt am Main, Allemagne

Traduit de l'anglais par *Sylvie Sollier*
Bibliothèque nationale de France

Session: **215 — *What is a national bibliography today and what are its potential uses?* — Bibliography**

Session : 215 – Qu'est-ce qu'une bibliographie nationale aujourd'hui et quels en sont les usages potentiels ? – Bibliographie

Résumé :

Cet article traite des caractéristiques attendues d'une bibliographie nationale au 21e siècle. Aux critères traditionnels de complétude et de fiabilité des données, de fraîcheur, de référencement et de persistance des données, les auteurs ajoutent le fait qu'une bibliographie nationale doit s'intégrer dans le World Wide Web, puisque c'est là qu'ont lieu aujourd'hui les échanges d'information. On utilisera pour cela les technologies de données liées et les données devront être publiées sous licence ouverte. Le travail effectué à la Deutsche Nationalbibliothek, où 70% environ de la base de données sont déjà publiés dans le web de données, est présenté comme étude de cas.

Introduction

On peut définir une bibliographie nationale comme la liste complète des publications provenant d'une aire géographique délimitée[1]. Toutefois, à l'époque d'Internet, il n'est pas

[1] Cf Anderson (1974) p. 12

sans intérêt de pousser plus loin cette définition, puisque Internet modifie à la fois la signification de « publication » et la façon de publier.

Alors qu'un livre imprimé traditionnel a un contenu indépendant et donc statique, il existe sur Internet des publications sans limites bien définies, qui peuvent être dynamiques et interactives, et au contenu en perpétuelle évolution. De plus, les moyens techniques d'enregistrement des textes produits ont changé : la possibilité de rechercher en plein texte a marginalisé le recours aux catalogues de métadonnées, thésaurus et systèmes de classification. Pas nécessairement parce que la recherche en plein texte est supérieure à la recherche à partir des métadonnées, mais parce qu'elle est plus facile à automatiser : les ressources en plein texte sont facilement disponibles sur le Web, alors que les métadonnées de haute qualité doivent être produites. La création de métadonnées, ou catalogage, nécessite soit des processus automatisés complexes, soit une intervention humaine.

Mais maintenant, une grande partie de ce processus nécessitant une intervention humaine s'effectue en dehors des bibliothèques : les éditeurs utilisent leurs propres métadonnées pour augmenter leur visibilité et des plateformes comme Wikipédia ou l'OpenLibrary proposent au grand public de publier des articles et des publications avec des métadonnées descriptives, créant ainsi un complément idéal à la recherche en plein texte. L'augmentation récente de la numérisation de masse et de l'OCR a encore plus modifié la situation en rendant possible la recherche plein texte dans des documents (auparavant) non numériques. Pour un moteur de recherche comme Google cela semblera bien sûr plus efficace que les méthodes traditionnelles de catalogage – dans leur moteur de recherche, la recherche par métadonnées est considérée comme un simple complément[2].

Dans ce contexte, les bibliothèques doivent se demander quelle valeur ajoutée une bibliographie nationale basée sur le catalogage traditionnel peut apporter, maintenant et dans le futur.

Conditions requises pour une bibliographie nationale

Traditionnellement, la bibliographie nationale visait trois groupes différents : les éditeurs et le commerce du livre, les bibliothèques et les utilisateurs finaux (en particuliers les chercheurs et les intellectuels). Ces trois groupes ont comme point commun de tirer parti des quatre propriétés élémentaires d'une bibliographie nationale, qui sont les suivantes :

1) Complétude et fiabilité des données

Pour les libraires, les éditeurs et les chercheurs, il était et il est encore vital que la bibliographie nationale décrive la production (professionnelle) complète, sans aucun parti pris politique ou relatif au contenu. En outre, le respect des règles de catalogage a joué un rôle important, en particulier lors de l'établissement des cumulatifs.

2) Fraîcheur des données

Notamment pour le commerce du livre et les bibliothèques la fraîcheur des données de la bibliographie nationale était extrêmement importante. Dans la seconde moitié du 20e siècle un grand effort a été fait sur le traitement des données pour assurer des cycles de publication courts malgré un nombre croissant de publications.

[2] Cf. http://books.google.com/intl/fr/googlebooks/about.html

3) Référencement

Étant donné la complétude, la fiabilité et la fraîcheur de ses données, une bibliographie nationale pouvait servir de référence à des travaux scientifiques : si un livre était répertorié dans la bibliographie nationale, son existence était attestée, et un livre non répertorié était considéré comme n'ayant jamais été publié.

4) Persistance

Pour servir de point de référence aux documents cités, il ne suffisait pas que les entrées présentent les critères de qualité indiqués ci-dessus, il fallait qu'elles aient aussi une certaine persistance. Jusqu'au début du 21e siècle ce n'était pas un problème, puisque la bibliographie était en soi une publication imprimée. On ne pouvait corriger une entrée bibliographique incorrecte que dans un cumulatif, mais à part cela, il n'y avait aucun moyen de faire disparaître la preuve qu'un livre avait été publié.

Les conditions requises pour une future bibliographie nationale doivent aller plus loin et se mesurer à l'utilisation et réutilisation des données dans le World Wide Web.

De plus en plus, le WWW évolue vers un espace ouvert aux échanges de données : ce qu'on appelle le nuage de données liées[3]. Depuis 2008, cette accumulation d'ensembles de données interconnectées a énormément cru en volume, mais on sait peu de choses sur la qualité et la persistance des données. Quoi qu'il en soit, le fonctionnement d'un web sémantique nécessite un certain seuil de fiabilité, concernant non seulement la qualité de l'information mais aussi et particulièrement la persistance de l'information. Ce réseau ne pourra se développer durablement que si nous pouvons être certains que l'information que nous annotons aujourd'hui sera disponible la semaine prochaine : on doit pouvoir la citer[4]. Ce problème concerne particulièrement les publications en ligne et leurs métadonnées. Pour assurer une fiabilité de longue durée, il est nécessaire de stabiliser cette partie de l'Internet. Cela pourrait être la tâche future des bibliothèques et autres organisations de préservation du patrimoine culturel.

Nous pouvons en conclure que :

1) Complétude et fiabilité des données

La hausse du nombre des publications en ligne rend virtuellement impossible de collecter toute la production originale relevant d'une bibliographie nationale. Il est très probable qu'une agence bibliographique nationale ne parviendra même pas à évaluer de façon fiable le degré d'incomplétude de la liste des publications qu'elle a répertoriées. Il lui faudra plutôt définir pour quels types de publications elle veut atteindre la complétude (par exemple les ressources imprimées, les blogs de haute qualité et certaines catégories de publications savantes indépendantes) et quand peut-elle se contenter d'instantanés périodiques d'un ensemble de sites web.

Certaines métadonnées, créées par traitement automatique, seront loin d'être parfaites. Pour que ces métadonnées soient utilisables, les traitements appliqués doivent être bien documentés, le fait que l'information résulte d'un traitement automatique devra être transparent pour l'utilisateur des données et il faudra déterminer dans quels cas on peut utiliser ce type de données.

[3] Pour une introduction aux données liées [Linked Data] et au nuage de données liées [Linked Data Cloud] cf. Heath and Bizer (2011)

[4] Cf. Schuster and Rappold (2006)

Pour appliquer des jeux de règles spécifiques, on se focalisera moins sur des règles de catalogage comme AACR2[5] ou RAK[6] que sur l'interopérabilité technique et sémantique des données pour intégrer des services externes.

2) Fraîcheur des données

De nombreux utilisateurs s'attendent à ce que les publications en ligne soient répertoriées dans la bibliographie nationale au moment où elles sont publiées. Ceci va à l'encontre des conditions de référencement et de persistance : si la bibliographie nationale s'appuie sur des métadonnées décrivant les premières versions d'une publication, le risque est relativement élevé qu'il y ait des changements dans la description bibliographique. Il faudra donc que les systèmes d'information des bibliothèques puissent archiver différentes versions des métadonnées pour en conserver différents états.

3) Référencement

La bibliographie nationale conservera à l'avenir son rôle capital de référencement des publications. Pour les publications électroniques en particulier, on peut s'attendre à ce que ce rôle soit de plus en plus important. Pour proposer ce service, il est nécessaire de fournir des identifiants uniques aux données bibliographiques fonctionnant dans un environnement électronique.

4) Persistance

Assurer la disponibilité des descriptions bibliographiques sur le long terme et garantir leur intégrité sont les pré-requis d'un référencement de longue durée. Il n'est pas seulement question d'assurer l'identification pérenne des données bibliographiques et d'autorité, mais aussi de contrôler une version donnée, de gérer les corrections et les destructions, et d'identifier des états particuliers, par exemple au moyen d'une marque de datation.

La topologie de la Bibliographie nationale allemande à l'ère numérique

Étant donné ces caractéristiques nécessaires, une bibliographie nationale ne doit pas, à l'évidence, se tenir à l'écart du World Wide Web. En Allemagne, la bibliographie nationale est passée d'un ensemble de fiches catalographiques à une base de données (électronique) qui utilise des structures de données de plus en plus complexes, où les descriptions bibliographiques sont liées aux données d'autorité et souvent à des ressources extérieures au système. La prochaine étape est de fondre la bibliographie nationale dans le WWW, de façon à faire intégralement partie du web à tous points de vue : topologique, fonctionnel, technique et organisationnel.

Au vu de ses caractéristiques, on pourrait se représenter notre bibliographie nationale sous la forme d'un graphe dans le World Wide Web. Ce graphe est relié à lui-même mais aussi à d'autres parties du WWW. Représenter les données des bibliothèques comme un graphe dans

[5] Anglo American Cataloguing Rules, cf. http://www.aacr2.org/

[6] Regeln für die alphabetische Katalogisierung, code de catalogage utilisé en Allemagne et en Autriche, cf. http://www.dnb.de/EN/Standardisierung/Regelwerke/regelwerke_node.html#doc3132bodyText1

le WWW ne s'applique pas seulement à la Bibliographie nationale allemande : il faut que d'autres répertoires d'objets appartenant au patrimoine culturel, des fichiers d'autorité, des thésaurus et des systèmes de classification fassent également partie du Web.

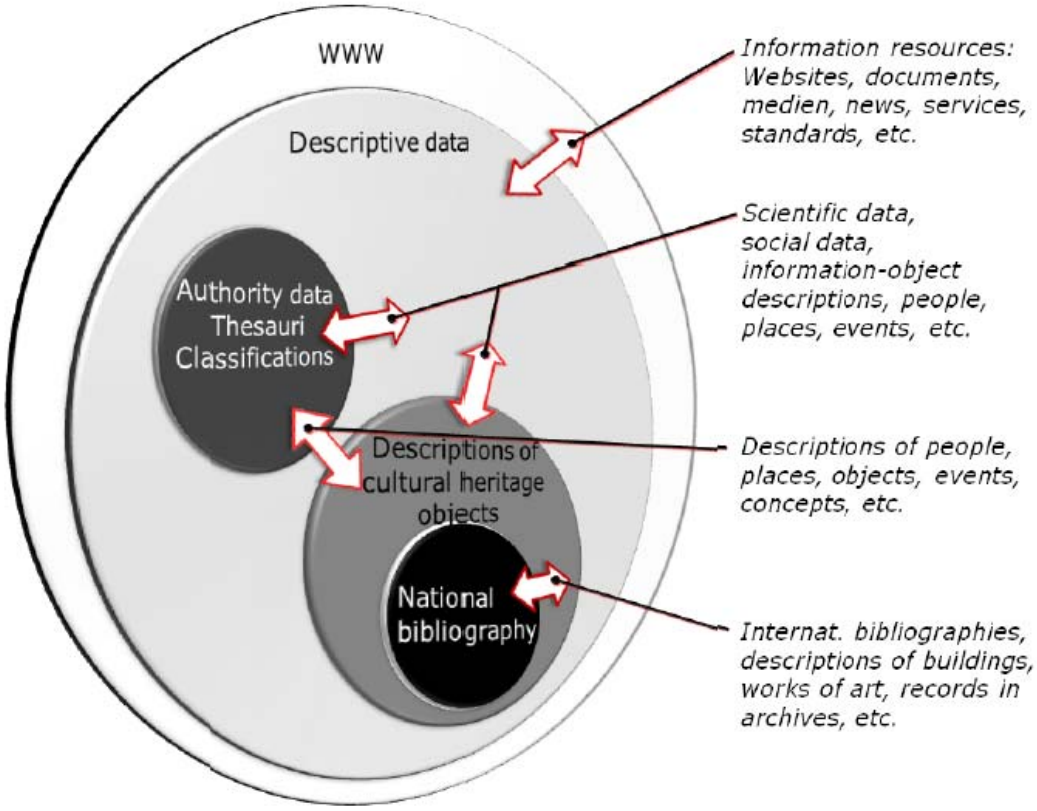

Figure 1 : la bibliographie nationale du futur peut être modélisée sous forme d'un graphe dans le WWW. *Les ellipses représentent les sous-graphes. Les flèches entre les sous-graphes indiquent les interconnexions possibles entre les nœuds dans les sous-graphes.*

Ainsi, la bibliographie nationale du futur ne peut pas être un système fermé et ne peut pas sans dommage être isolée du Web. Une version imprimée de la bibliographie ou même une sauvegarde complète des données de la bibliothèque nationale ne serait qu'un produit dérivé lacunaire, puisque la complétude de la bibliographie nationale dépend de données extérieures aux bases de données de la bibliothèque.

Dans ce graphe les nœuds représentent des artefacts issus d'un travail littéraire ou artistique. Ces artefacts – nous les appellerons des « biens culturels » - sont *grosso modo* les mêmes que les entités du groupe 1 des FRBR[7]. Chacun de ces nœuds constitue un paquet des propriétés significatives de l'objet qu'il représente et utilise ces propriétés pour décrire l'objet. Ces paquets de propriétés ont beaucoup de points communs avec les descriptions qui figurent sur les fiches catalographiques parce que ce sont des données (ou des faits) saisis selon un ensemble de règles précises. Le titre principal d'une publication est toujours la même chaîne

[7] Pour une description des entités des FRBR cf. IFLA (2009)

de caractères et son ISBN la même suite de chiffres. Cependant, la plupart des autres propriétés sont des références à d'autres entités, c'est-à-dire des arêtes orientées pour connecter un nœud à d'autres nœuds dans le WWW. Ces autres nœuds peuvent être eux aussi des nœuds de métadonnées : descriptions de biens culturels, de personnes, de collectivités, d'événements, de lieux, de concepts, etc.

Quelques-uns de ces nœuds sont édités par les bibliothèques (par exemple GND[8], RAMEAU[9] et LCSH[10]). Il est aussi possible de faire des renvois de métadonnées de bibliothèques à d'autres nœuds, par exemple des liens à l'objet lui-même (pour une ressource du Web) ou des renvois à des nœuds dans des ensembles de données gérées par des organisation extérieures à la communauté des bibliothèques, tels que des liens entre l'autorité d'un nom et l'entrée correspondante dans Wikipédia.

De ce point de vue, la Bibliographie nationale allemande est déjà un ensemble de données fortement interconnectées dans un graphe appelé le réservoir de données de la Deutsche Nationalbibliografie. Par « réservoir de données » nous entendons le graphe complet disponible aujourd'hui et à l'avenir. C'est-à-dire que le réservoir de données comprend non seulement les données dont la Deutsche Nationalbibliothek est responsable d'un point de vue éditorial, mais aussi tous les ensembles de données en relation avec les données de la Deutsche Nationalbibliografie. Comme la Deutsche Nationalbibliothek expose ses ensembles de données sur le Web (par exemple en proposant des services de données ouvertes liées [linked open data] dans le Web de données), notre réservoir de données contiendra inévitablement une quantité croissante de données hors bibliothèques. Les relations dans le réservoir de données proviennent de plusieurs mécanismes : l'utilisation de numéros normalisés internationaux tels que l'ISBN ou l'ISMN crée des relations implicites avec les exemplaires d'autres bibliothèques portant le même numéro normalisé ; des relations sont créées en mettant en correspondance des vocabulaires contrôlés ou des fichiers d'autorité, soit intellectuellement (par exemple GND, LCSH et RAMEAU) ou par des algorithmes informatiques.(par exemple VIAF[11]) ; ou des relations provenant de coopérations avec des organisations tiers (par exemple Wikipédia).

Frontières et limites

Il est difficile de définir où commence et où finit exactement une bibliographie nationale, mais on peut très facilement trouver de grands ensembles de données contenant toute l'information pertinente pour sa création. De façon générale, une bibliographie nationale est un sous-ensemble spécifique de toutes les données descriptives du WWW. En outre, c'est un sous-ensemble de toutes les données descriptives du domaine culturel. Ces données

[8] Le Gemeinsame Normdatei (GND – Fichier d'autorité intégré) est un fichier d'autorité allemand qui couvre tous les types d'entités et qui sert de système d'autorité de référence commun aux données bibliographiques des bibliothèques et aux données catalographiques d'autres utilisateurs de fichiers d'autorité tels que les archives, les musées, les projets, les institutions scientifiques et culturelles.
Voir http://www.dnb.de/EN/Home/home_node.html

[9] RAMEAU (Répertoire d'autorité-matière encyclopédique et alphabétique unifié) est un langage documentaire utilisé par la Bibliothèque nationale de France, les bibliothèques universitaires françaises et plusieurs bibliothèques publiques en France. Voir http://rameau.bnf.fr/

[10] Les Library of Congress Subject Headings (LCSH) sont utilisés pour indexer les documents à la Bibliothèque du Congrès et dans le monde entier. Voir http://id.loc.gov/authorities/subjects

[11] VIAF – le Virtual International Authority File (http://viaf.org/) "est un service international destiné à fournir un accès pratique aux principaux fichier d'autorité de noms » Cf. http://www.oclc.org/viaf/

comprennent par exemple des descriptions de personnes, d'objets, de concepts et de lieux pertinents dans le contexte du patrimoine culturel. S'y ajoutent des contraintes telles que la limitation à une région géographique particulière (la nation) et la restriction à un genre spécifique d'objets du patrimoine culturel (par exemple des livres, des cartes, des cartes postales, des enregistrements sonores etc.)[12]

Un autre aspect des limites d'une bibliographie nationale est la responsabilité administrative. Pour satisfaire les exigences de fiabilité à long terme et de grande qualité des données, l'organisme responsable doit avoir un appui gouvernemental suffisant et disposer de l'équipement adéquat pour gérer et traiter les données dans le futur proche. Pour une bibliothèque nationale ou une agence bibliographique nationale, c'est un élément fondamental de leur raison d'être et cela devrait faire partie de leurs missions, souvent – mais pas toujours – en s'appuyant sur une législation de dépôt légal[13]. Il ne doit cependant pas en découler que toutes les données gérées par une agence bibliographique nationale font automatiquement partie de la bibliographie nationale : la Deutsche Nationalbibliothek gère plusieurs ensembles de données, y compris des collections spécialisées, qui ne sont pas pertinentes pour la Bibliographie nationale allemande.

Contenu

Lors d'un atelier à la Deutsche Nationalbibliothek, les participants devaient définir ce qui à leur avis faisait partie de la Bibliographie nationale allemande. Les réponses allaient de « toutes les données gérées par la bibliothèque » à « uniquement les titres et en excluant les collections spécialisées ». Cependant, même le groupe défendant la solution minimale ne souhaitait pas l'abandon total des données d'autorité. Ce n'est pas surprenant puisque les données d'autorité sont un produit du processus de catalogage et contiennent des informations – par exemple le nom de l'auteur – nécessaires à un affichage correct en ISBD, sans cette information, la notice bibliographique n'est pas assez renseignée. Par contre, toutes les données d'une notice d'autorité ne sont pas nécessaires à l'affichage en ISBD d'une description bibliographique dans une bibliographie nationale imprimée.

Cela conduit à supposer qu'on ne peut pas résoudre la question des données nécessaires pour fabriquer une bibliographie nationale en se situant au niveau des entités, mais plutôt à celui des attributs individuels : les attributs nécessaires à la notice d'une entité du patrimoine culturel n'appartiennent pas tous à l'entité elle-même mais à d'autres entités qui lui sont attachées[14]. Cela s'applique aussi à l'environnement de recherche. Dans l'idéal, un catalogue en ligne devrait utiliser toutes les données disponibles pour optimiser le processus de recherche. Cela peut inclure des données extérieures au monde des bibliothèques, créées sans aucune référence aux entités bibliographiques. Par exemple l'utilisation des coordonnées géographiques provenant d'ensembles de données comme geonames.org, qui permettent de chercher par lieu ou par proximité géographique.

Afficher en ISBD et rechercher dans un catalogue en ligne ne sont que deux des nombreux services accompagnant la bibliographie nationale. Les données utilisées par ces deux services

[12] Pour de plus amples commentaires sur ce point, cf. IFLA (2009)
[13] Cf. Andersen (1974) p. 11
[14] Ceci est avant tout un prolongement de l'affirmation qu' « une description bibliographique est normalement basée sur l'item qui représente la manifestation et peut inclure des attributs qui caractérisent l'œuvre ou l'expression contenue » provenant de l'IFLA (2009) p. 4

ne sont pas toutes considérées comme faisant partie de la bibliographie, nous ne pouvons donc pas déduire ce qui fait partie d'une bibliographie nationale en observant les services (bibliographie imprimée, recherche en ligne, services bibliographiques nationaux) qui sont construits dessus.

En construisant des services bibliographiques (nationaux) sur des données qui ne sont pas exclusivement utilisées pour la bibliographie nationale, on devrait supprimer complètement la dépendance entre service et données. Une image claire et cohérente émerge si on se concentre sur ce qui est au cœur de la bibliographie nationale : les publications. De cette façon, nous parvenons à un ensemble minimal de données, composé d'éléments textuels et de renvois, qui pourrait ne pas être très utile en tant que tel. Si on cherche plus loin, du côté des FRBR[15] et de RDA, nous pouvons l'énoncer plus précisément : une bibliographie nationale enregistre des manifestations, c'est-à-dire « la matérialisation de l'expression d'une œuvre »[16]. Les données « œuvre », « expression » et « item » – tout comme les notices d'autorité – appartiennent à leur propre ensemble de données, et les données de la bibliographie nationale contiennent uniquement des renvois à ces autres entités.

Processus de création

Les bibliographies nationales devraient participer de plus en plus à un écosystème global d'utilisateurs et de producteurs de données. Plus forts sont les liens avec des données indépendantes du domaine, provenant d'une plus grande variété d'institutions, plus les applications tireront profit de sa base de connaissance. Les services fournis par les bibliothèques s'appuieront de plus en plus sur des données de sources externes. La disponibilité d'informations supplémentaires sur les entités relatives au patrimoine culturel – que se soient des biens culturels, des personnes ou des concepts – nous permet d'améliorer nos services bibliographiques. Il se pourrait même que nous ne gérions plus nous-mêmes une partie de ces données mais que nous réutilisions à la place l'information fournie par un tiers.

[15] Functional Requirements for Bibliographic Records. Cf. http://www.ifla.org/publications/functional-requirements-for-bibliographic-records
NdT : traduction française établie par la Bibliothèque nationale de France : *Spécifications fonctionnelles des notices bibliographiques* Cf. http://www.bnf.fr/documents/frbr_rapport_final.pdf
[16] IFLA (1998) p. 21

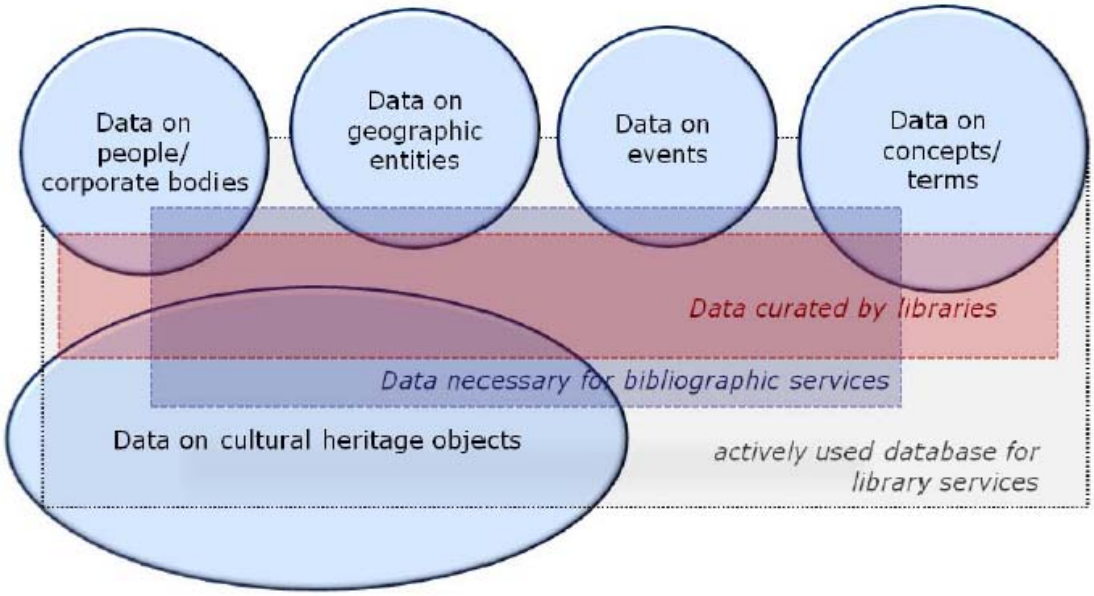

Figure 2 : future répartition possible de la création et de l'utilisation des données dans les bibliothèques.
1. Les éléments de données des catalogues de bibliothèques ne sont pas tous nécessaires aux produits bibliographiques. Ils peuvent néanmoins être indispensables, par exemple pour la communication ou l'archivage.
2. Les éléments de données nécessaires aux services bibliographiques ne sont par tous produits/catalogués par les bibliothèques.
3. La base des données utilisées par les bibliothèques est plus grande que ces deux ensembles de données réunis.

Cette approche n'est pas entièrement nouvelle. Même aujourd'hui, à la Deutsche Nationalbibliothek nous ne créons pas manuellement toutes les parties de la description bibliographique mais réutilisons de l'information produite par des tiers, telle que des métadonnées fournies par les éditeurs, ou de l'information créée par traitement semi-automatique comme les tables des matières ou des vedettes-matière générées automatiquement. Dans ce contexte, il semble logique que la prochaine étape soit l'extension de la réutilisation des données et l'ouverture contrôlée du catalogage à une sélection de communautés du Web.

Pour les bibliothèques, il s'agit moins d'abandonner des tâches particulières que de se recentrer sur nos points forts, tels que la gestion des données de haute qualité, l'organisation du savoir et le contrôle d'autorité. Il en résultera une division du travail plus efficace où chaque fournisseur de données pourra se concentrer sur ses compétences fondamentales tout en partageant les données comme un bien commun. Le centre national bibliographique a la responsabilité des objets du patrimoine culturel publiés sous forme textuelle, de tous les renvois de ces objets vers d'autres entités, et de leurs propriétés quand celles-ci sont nécessaires aux services bibliographiques et si aucune autre institution ne peut les fournir au niveau de qualité adéquat. D'autres fournisseurs de données sont responsables d'autres propriétés.

Du fait du nombre croissant des publications, les bibliothèques devront restreindre la création manuelle, la sauvegarde et la vérification des métadonnées aux domaines où ces procédures ajoutent de la valeur par rapport à la création automatisée de données. La définition de ce qui constitue la valeur ajoutée ne dépendra pas, dans ce cas de figure, de la conformité à des

règles de catalogage particulières mais se concentrera plutôt sur l'utilité accrue de la sélection et du filtrage des données, d'une part, et des modèles économiques des usagers de la bibliothèque, d'autre part.

Caractères qualitatifs

Pour être sûres que la qualité des services de la bibliothèque ne souffre pas de l'intégration de données fournies par un tiers, mais qu'elle peut, au contraire, leur être profitable, les bibliothèques doivent étudier soigneusement quelles sources de données extérieures elles peuvent réutiliser. La fiabilité (référencement, persistance, utilisation de normes et garantie de qualité) des données est cruciale. L'origine des données et – s'il est disponible – le processus à l'origine de leur création doivent être transparents et des processus de contrôle de la qualité doivent être mis en place. En particulier, chaque ensemble de données externe doit être évalué par rapport à ses perspectives à long terme.

La réutilisation d'ensembles de données fournis par un tiers dans une bibliographie nationale peut provoquer des divergences avec les règles de catalogage existantes. Beaucoup – ou même la plupart – des fournisseurs de données extérieurs à la communauté des bibliothèques ne connaissent pas les AACR2, RAK ou RDA et il ne faut pas espérer de leur part un investissement dans ce domaine. Il nous faut plutôt épurer les règles de catalogage existantes et les évaluer pour savoir si la règle permet à la bibliographie nationale d'atteindre son objectif : rendre compte d'une façon exhaustive du paysage éditorial d'un pays. C'est seulement si la règle permet cela qu'elle sera utilisée comme référence pour évaluer la qualité des ensembles de données fournis par un tiers.

Si une institution utilise des métadonnées provenant d'un tiers, elle devra définir les caractéristiques d'un fournisseur de données digne de confiance. Pour y parvenir, une liste de critères est nécessaire, semblables à ceux qui existent par exemple pour les archives numériques digne de confiance. Le critère principal, pour établir cette liste, sera de savoir si les données fournies remplissent les conditions nécessaires à l'interopérabilité. Les données ne provenant pas de fournisseurs dignes de confiance devront être marquées en conséquence.

Les mêmes critères s'appliquent aux données générées par traitement automatique. Là, il est important que les processus soient bien documentés et les résultats reproductibles. Si l'on peut parvenir à produire des données homogènes pour des sous-ensembles de la bibliographie nationale, l'acceptation du produit par le consommateur sera bien plus grande. La rédaction en commun d'une liste des critères pour des processus automatiques dignes de confiance peut être un moyen d'y parvenir. Il va sans dire qu'il faudra contrôler régulièrement ces processus.

Réalisation technique

Nous avons examiné ci-dessus comment une bibliographie nationale peut être produite à partir d'un ensemble minimal de métadonnées de manifestations, combiné à de l'information provenant de données d'autorité et d'ensembles de données reliés. Ceci, allié à la proposition de considérer la bibliographie comme un graphe dans le World Wide Web, mène tout naturellement à utiliser les techniques de données liées – essentiellement un réseau http pour que les relations entre entités soient lisibles par des machines – pour réaliser techniquement la bibliographie.

Si on considère les caractéristiques des données liées telles qu'exprimées par Tim Berners-Lee[17], l'implémentation est assez simple :

1) Utiliser des URI comme noms pour les choses
   Chaque entité utilisée dans la bibliographie nationale doit avoir son propre identifiant unique. Ceci s'applique au minimum aux manifestations, aux vocabulaires contrôlés, aux collectivités et aux personnes, et dans l'idéal à d'autres données d'autorité comme les éditeurs.
2) Utiliser des URI http pour l'on puisse rechercher ces noms
   Si les bibliothèques attribuent des identifiants à leurs entités, ils doivent s'ancrer dans l'infrastructure du WEB.
3) Si quelqu'un cherche une URI, fournir de l'information utile
   Afin de promouvoir la réutilisation des données des bibliothèques, il est impératif de fournir de brèves descriptions du nœud de données identifié par les URIs lorsqu'on a déréférencé un identifiant. Ce faisant, nous incitons d'autres fournisseurs de données à réutiliser les ensembles de données des bibliothèques en se liant aux données d'autorité ou à l'information bibliographique en utilisant nos URIs.
4) Inclure des liens vers d'autres URIs afin de trouver plus de choses
   En réutilisant des données d'autorité ou des données d'autres fournisseurs de contenu dans nos notices bibliographiques, on peut choisir d'afficher certaines propriétés de l'entité liée, par exemple un nom d'auteur ou la forme retenue d'une vedette matière, avec le titre de la manifestation, tout en proposant un lien à une page où l'utilisateur pourra trouver plus d'information, telle que la date de naissance ou la profession.

Que la modélisation des données soit faite selon RDF ou tout autre moyen technique importe peu pour construire cette infrastructure, les données liées ne dépendant pas des formats. L'ensemble des technologies RDF possède vraiment un fort potentiel pour rendre les données et les connexions sémantique lisibles (et donc compréhensibles) par une machine, de façon générique. Cependant, cela mènera dans de nombreux cas à un modèle de données d'une grande complexité, et les problèmes apparus en utilisant RDF pour les applications déployées ont suscité d'âpres débats[18]. A la Deutsche Nationalbibliothek, nous avons aussi pu vérifier qu'il existe des cas d'utilisation, par exemple des relations temporelles ou bien la source d'un graphe, où une expressivité plus grande que ce que RDF peut proposer est nécessaire, à moins d'utiliser des modèles excessivement complexes – au moins pour le moment. La situation pourrait se modifier au fur et à mesure que le modèle RDF évolue et que des concepts comme les graphes nommés se généralisent. Aussi longtemps que nous construirons systématiquement nos modèles de données selon les principes des données liées – en publiant nos données sur le WW sous des identifiants stables et en enrichissant ces représentations avec des renvois à d'autres entités – nous ouvrirons la voie à des langages de description plus expressifs.

Utilisation

Lorsqu'ils recherchent de l'information sur le web, la plupart des utilisateurs ont recours à des moteurs de recherche commerciaux comme Bing, ask.com et Google, ou à des réseaux

[17] Berners-Lee (2006)

[18] Un exemple parmi beaucoup d'autres, voir Miličić (2011) et les arguments qu'il contient.

sociaux comme LinkedIn et Facebook[19]. Pour les bibliothèques (nationales) et les autres institutions gérant le patrimoine culturel, cela signifie que nous devons offrir à ces entreprises la possibilité d'intégrer nos données, de façon créative, aux services qu'elles proposent, de sorte que les utilisateurs puissent bénéficier de cette information au cours de leur recherche sur le web. Par exemple, le portail européen consacré au patrimoine culturel, Europeana, a pu observer que créer une page de destination spécifique pour chaque objet de la collection et donner à cette page un URI unique, améliorait sa visibilité par les moteurs de recherche et faisait augmenter le trafic provenant de ces moteurs[20].

Cela démontre que le catalogue en ligne propre à la bibliothèque continuera à jouer un rôle important dans le paysage de l'information. En tant que moteur de recherche hautement spécialisé, c'est l'application de référence pour démontrer ce que l'on peut réaliser avec les données des bibliothèques et comment on peut ajuster ces données à des groupes spécifiques d'utilisateurs. D'autre part, la frontière entre la valeur ajoutée proposée par un catalogue de bibliothèque et un moteur de recherche spécialisé devient de plus en plus floue. On peut penser que, dans le futur, certains fournisseurs de moteurs de recherche se spécialiseront en information bibliographique, par exemple pour des communautés scientifiques particulières. La disponibilité de données bibliographiques de haute qualité, intégrées à l'environnement de recherche d'information des utilisateurs, sera profitable à nos usagers et leur rendra plus facile la recherche, la sélection et l'obtention des ressources appropriées.

Pour que les données des bibliothèques soient facilement réutilisées dans d'autres services, il convient d'indiquer explicitement les conditions de licence. Pour une bibliographie nationale, en particulier, visibilité plus grande et réutilisation des données sont sources de valeur ajoutée, le moyen le plus simple pour en favoriser le développement est de proposer les données sous une licence souple et ouverte, qui en autorise explicitement la réutilisation commerciale. Ceci peut être sujet à controverse, si l'on considère que des acteurs importants comme Google sont financièrement en mesure d'acheter les métadonnées des bibliothèques. On peut cependant douter qu'ils veuillent le faire, et si l'information des bibliothèques n'est pas visible sur les principaux moteurs de recherche, elle sera invisible pour la plupart des personnes faisant des recherches sur le web. Utiliser une licence ouverte permet aussi d'intégrer des plateformes ouvertes, telles que Wikipédia, et de petits et moyens sites comme les portails d'actualité en ligne. De cette façon, la plus grande visibilité l'emportera définitivement sur la perte (potentielle) de revenu.

Le travail courant à la Bibliothèque nationale d'Allemagne et ses implications pour la Bibliographie nationale allemande.

La Bibliothèque nationale d'Allemagne – comme aussi d'autres bibliothèques nationales – travaille à l'implémentation des évolutions nécessaires.

A partir de 2008, la Deutsche Nationalbibliothek a entrepris la publication de sa base de données entière sur le Web de données, actuellement environ 70% des titres sont disponibles. Le premier ensemble de données était les données d'autorité du GND, suivi de la traduction allemande de la Classification décimale Dewey. Au bout de six mois, le projet permettait de

[19] Selon un récent rapport PEW, 92% des internautes adultes utilisent des moteurs de recherche pour trouver de l'information. Cf. Purcell (2011)
[20] Clark et al. (2011) p. 15-17

diffuser la plus grande partie des données titre de la bibliographie nationale en données liées. A part les données de la CDD, qui sont disponibles sous une licence CC BY-NC-ND, toutes les autres données de la Deutsche Nationalbibliothek sont publiées sous CreativeCommons Zero.

De plus, les métadonnées de sujet ou de description sont de plus en plus produites par traitement automatique. Un des axes est l'enrichissement des descriptions bibliographiques du catalogue, par exemple avec des tables des matières, et nous intensifions l'inclusion de métadonnées provenant d'éditeurs et d'universités, y compris des catégories de sujets – par exemple les codes BISAC - ou des vedettes matières. Pour les publications en ligne dans la série O de la bibliographie nationale, nous réutilisons complètement les métadonnées de sujet et de description fournies par les déposants de contenu.[21]

Un projet est en cours pour évaluer la possibilité d'ajouter automatiquement des vedettes matière aux notices bibliographiques. Le processus fonctionne à partir d'un indexeur automatique. Nous analysons comment il traite différents types de documents et la qualité de la recherche pour des publications indexées automatiquement. Les premiers résultats seront disponibles en 2013.

L'indexation automatique des ressources du web sera réalisée avec les mêmes vedettes d'autorité contrôlées que celles utilisées dans une démarche d'indexation manuelle, afin que la recherche d'information demeure aussi homogène que possible. Ainsi, des données d'autorité fiables pourront être proposées. De plus, on pourra utiliser efficacement l'information liée pour les personnes, lieux etc., mais aussi les relations sémantiques à l'intérieur des notices d'autorité.

Les attentes des utilisateurs d'aujourd'hui concernant les données bibliographiques, notamment les accès sujet, sont variées, elles incluent une vue d'ensemble de la documentation disponible, des références bibliographiques ou l'accès direct aux publications. A la Deutsche Nationalbibliothek, nous adhérons à l'opinion de l'IFLA qu'une politique d'indexation cohérente et l'utilisation d'accès contrôlés demeurent importants pour apporter ordre et cohérence des données[22]. Tous nos utilisateurs gagnent à disposer de structures de sujets bien organisées – non dépendantes des formats de données, des réseaux de diffusion ou des formats d'affichage à travers lesquels ils utilisent la bibliographie nationale. Les systèmes de classification et les vedettes matière aident les utilisateurs à rechercher (trouver, identifier, sélectionner ou découvrir) l'information dont ils ont besoin. La majorité des utilisateurs (utilisateurs finaux ou professionnels) s'intéresse à une partie des ensembles de données plutôt qu'à la bibliographie toute entière. La plupart des utilisateurs peuvent même ignorer qu'ils cherchent dans la bibliographie en faisant des recherches dans notre catalogue en ligne. Un nombre croissant d'utilisateurs dispose d'un profil personnel pour faire des recherches dans la bibliographie ou s'abonne à un flux RSS à partir des résultats d'une recherche. Les informations supplémentaires sur le contenu, comme par exemple les tables des matières ou les résumés, qui y sont liées, aident les utilisateurs à découvrir l'information qu'ils recherchent. La mise à disposition de liens à des contenus en ligne, en particulier à partir de livres et de journaux électroniques, répond aux besoins des utilisateurs.

[21] Pour plus d'information sur le traitement des publications en ligne à la Deutsche Nationalbibliothek cf. Gömpel and Svensson (2001)
[22] IFLA (2012)

En raison du volume croissant de l'information publiée, il devient vraiment nécessaire de la classer par catégories, en unités lisibles, sélectionnables, pouvant faire l'objet d'une recherche précise. Dans l'idéal, tous les documents font l'objet d'une indexation complète et détaillée afin de permettre aux utilisateurs de trouver des sujets de recherches pertinents ou de l'information supplémentaire sur le contenu. Devant la quantité énorme de publications imprimées en Allemagne et le nombre croissant de ressources publiées sur le Web sous le domaine *.de*, et aussi en raison des restrictions de budget et de personnel, nous avons adopté une politique d'indexation progressive qui nous permet de gérer l'ensemble de la production nationale avec des niveaux de détails différents. Cette politique[23] de modulation permet de varier les points d'accès selon les différents types de media, mais elle est transparente et soumise à des critères de qualité.

Le niveau minimal d'indexation est un indice de classification très large qui est fourni pour presque toutes les ressources – Sachgruppe / groupe sujet assimilable à un indice de classification sommaire. La structure est basée en grande partie sur les deux niveaux hiérarchiques supérieurs de la CDD (les cent divisions ou 2^{e} niveau d'arborescence Dewey). Quelques aménagements on été introduits en intégrant des niveaux plus profonds, pour répondre aux besoins des utilisateurs locaux. Une recherche séparée est possible pour les ouvrages de fiction et les livres pour enfants, de même que pour les manuels scolaires.

Plusieurs autres projets permettent d'améliorer les liens à d'autres ensembles de données. L'optimisation de nos propres données est un bon point de départ. Il est apparu évident que passer de descriptions textuelles à des données référencées convient particulièrement à une automatisation. D'autres processus de reconnaissance des entités nommées et la transformation de noms d'entités en références sont actuellement au stade expérimental. Pour créer des liens à des ensembles de données externes la Deutsche Nationalbibliothek emploie un mélange de processus coopératifs, manuels et automatiques. En passant par VIAF nous avons pu relier automatiquement des personne du GND à d'autres fichiers d'autorité, alors que MACS[24] utilisait un processus manuel de rédaction de liens entre les vedettes matières de GND, LCSH et RAMEAU. Une coopération avec le Wikipédia allemand ajoute régulièrement des liens entre les articles de Wikipédia et les personnes du GND et nous développons au sein de CONTENTUS une technique pour construire un réseau de noms géographiques provenant du GND, des articles de Wikipédia et de geonames.org.

Pour poursuivre ce développement, la DNB a fondé avec hbz, le réseau des bibliothèques de Rhénanie du Nord–Westphalie, la plateforme culturegraph.org. L'idée qui sous-tend ce projet est de fournir une infrastructure technique et organisationnelle pour prendre en charge l'interconnexion des données bibliographiques, des thésaurus, des systèmes de classification et d'autres données d'autorité. Le réseau d'information ainsi produit sera aussi publié sous forme de données liées. Un des attributs centraux de la plateforme est de permettre la « citabilité ». Actuellement, la plateforme contient les descriptions bibliographiques de toutes les publications depuis 1945, provenant des réseaux allemands de bibliothèques. Un premier objectif, à moyen terme, est de créer des grappes de données pour déterminer quelles notices bibliographiques décrivent la même manifestation, et de nommer et publier ces grappes sous

[23] Pour une vue d'ensemble des différents canaux de distribution des données bibliographiques à la Deutsche Nationalbibliothek, voir Svensson and Jahns (2010)

[24] MACS (Multilingual access to subjects) : Accès multilingue par sujet, est un projet dont l'objectif est de développer un système permettant un accès matière multilingue aux catalogues de bibliothèques utilisant des langages d'indexation existants. Il utilise actuellement RAMEAU, LCSH et les vedettes matiière du GND. Cf. http://www.nb.admin.ch/nb_professionnel/projektarbeit/00729/00733/index.html?lang=fr

un unique identifiant commun. Une évaluation plus approfondie permettra de savoir s'il est possible de créer des grappes pour les œuvres (au sens FRBR) et si ces données peuvent servir de point de départ à un fichier d'autorité des œuvres intellectuelles.

Toutes ces activités de niveau national doivent s'intégrer au contexte international, la tâche la plus importante sera peut-être de continuer à participer aux projets internationaux et aux travaux des organisations de normalisation sur l'avenir du catalogage et l'échange des données.

Conclusions et travaux futurs

Dans cet article, nous défendons l'idée d'un écosystème des données de bibliothèques construit sur le principe des données liées, publiées dans le World Wide Web sous licence ouverte. Pour diffuser ces données aussi largement que possible, nous devrions avoir pour objectif de coopérer avec les principaux fournisseurs de moteurs de recherche et de leur faire intégrer des données bibliographiques dans leurs bases de données.

Afin de fournir à nos usagers une grande variété de services, - à la fois en termes de choix de données et de formats – la base de données sous-jacente devra être aussi souple que possible. Pour utiliser une analogie avec l'indexation, nous devons passer de données coordonnées *a priori* à une architecture de l'information où les morceaux d'information seront coordonnés *a posteriori* pour répondre à des demandes en constante évolution.

Les débats actuels sur les règles de catalogage devront se concentrer sur les besoins révélés par l'inclusion de données bibliographiques dans le WWW. L'objectif de RDA est d' « être une nouvelle norme de description des ressources et des accès destinée au monde numérique ». [25] Bien qu'on puisse légitimement lui reprocher de favoriser trop fortement les catégories traditionnelles de catalogage et de ne pas pousser sa réflexion suffisamment loin, c'est un pas dans la bonne direction étant donné la souplesse accrue qu'il propose. Il ne fait aucun doute, cependant, que de nouvelles révisions seront nécessaires.

Peut-être encore plus important que les règles de catalogage des bibliothèques : la possibilité d'interagir avec des données n'appartenant pas à l'écosystème des bibliothèques. Des données tout à fait utiles et disponibles – par exemple l'annonce d'un éditeur pour une nouvelle publication, qui pourrait contenir des mots répondant à la recherche d'un utilisateur – ne sont pas utilisées pour rechercher et récupérer de l'information. Pour ce faire, les bibliothèques doivent disposer des moyens permettant de conserver la provenance des données de parties tiers et de rendre cette information transparente pour l'utilisateur final.

En outre, la « citabilité » doit devenir un critère de qualité. Une description bibliographique marquée comme « citable » ne devra jamais être détruite, et il devra être possible de retracer toutes les modifications qui y ont été apportées. Pour permettre cela, il nous faut améliorer les processus de création et de maintenance des données et ajouter une information de provenance à toutes les données émanant de traitements automatisés.

Enfin, nous devons nous impliquer davantage dans les liens des données de bibliothèques à d'autres ensembles de données. Nous devons autant que possible utiliser des références à des

[25] Cf. le plan stratégique pour RDA 2005-2009. Disponible en ligne à http://rda-jsc.org/stratplan.html

données d’autorité au lieu de contenus textuels (« libellés »). Une coopération internationale autour d’un fichier commun de données d’autorité – par exemple VIAF - sera un progrès majeur dans la bonne direction et évitera de refaire inutilement le travail déjà fait ailleurs. Un entrepôt centralisé des titres d’œuvres apportera une aide immense à la connexion des descriptions bibliographiques, par delà les frontières nationales et linguistiques. Une base de données d’éditeurs (un sous-ensemble de collectivités) ne fournit pas seulement des points d’accès, mais favorisera la coopération entre les bibliothèques et les éditeurs.

La bibliographie nationale devra être un graphe dans le web de graphes. L’ère numérique a fait changer le secteur de l’information. Au lieu de s’accrocher aux structures et aux pratiques existantes, les bibliothèques devraient accompagner le changement et essayer au contraire d’en devenir les acteurs. Davantage d’ouverture et de souplesse est un bon point de départ.

Bibliographie

Anderson (1974)
Anderson, Dorothy: Universal bibliographic control : a long term policy; a plan for action. Pullach 1974.

Berners-Lee (2006)
Berners-Lee, Tim: Linked Data. Disponible à : http://www.w3.org/DesignIssues/LinkedData.html

Clark et. al (2011)
Clark, D. J.; Nicholas, D.; Rowlands, I.: D3.1.3 – Publishable report on best practice and how users are using the Europeana service. Disponible à : http://www.europeanaconnect.eu/documents/D3.1.3_eConnect_LogAnalysisReport_v1.0.pdf

Gömpel and Svensson (2011)
Gömpel, Renate; Svensson, Lars G.: Managing Legal Deposit for Online Publications in Germany. 2011.
urn:nbn:de:101-2011061609.
Disponible à : http://nbn-resolving.de/urn:nbn:de:101-2011061609

Heath and Bizer (2011)
Heath, Tom; Bizer, Christian: Linked Data: Evolving the Web into a Global Data Space. San Rafael (Calif.) 2011. (Aussi disponible à : http://linkeddatabook.com/editions/1.0/)

IFLA (1998)
Functional requirements for bibliographic records : final report. München 1998.
All references in this paper are to the 2009 revised online version at http://www.ifla.org/files/cataloguing/frbr/frbr_2008.pdf

> Traduction française établie par la Bibliothèque nationale de France : Groupe de travail IFLA sur les Spécifications fonctionnelles des notices bibliographiques. *Spécifications fonctionnelles des notices bibliographiques : rapport final*, disponible en ligne : http://www.bnf.fr/documents/frbr_rapport_final.pdf

IFLA (2009)
IFLA Cataloguing Principles: Statement of International Cataloguing Principles (ICP) and its Glossary. Ed. Barbara Tillett and Ana Lupe Christián. München 2009.
Pour une liste des versions (y compris des traductions) voir : http://www.ifla.org/publications/ifla-series-on-bibliographic-control-37

> Disponible en français : *Principes internationaux de catalogage* http://www.ifla.org/VII/s13/icp/ICP-2009_fr.pdf

IFLA (2012)
Guidelines for Subject Access in National Bibliographies. Ed. Yvonne Jahns. Berlin: 2012.

Miličić (2011) Miličić, Vuk: The Ultimate Problem of RDF and the Semantic Web. Blogue disponible à : http://milicicvuk.com/blog/2011/07/19/ultimate-problem-of-rdf-and-semantic-web/

Purcell (2011)
Purcell, Kristen: Search and email still top the list of most popular online activities: Two activities nearly universal among adult internet users. 2011. Disponible à :
http://pewinternet.org/~/media//Files/Reports/2011/PIP_Search-and-Email.pdf

Schuster and Rappold (2006)
Schuster, Michael; Rappold, Dieter: Social Semantic Software – was soziale Dynamik im Semantic Web auslöst. In: Semantic Web: Wege zur vernetzten Wissensgesellschaft. Ed. Tassilo Peregrini.
Berlin 2006.

Svensson and Jahns (2010)
PDF, CSV, RSS et autres acronymes : redéfinition des services bibliographiques à la Bibliothèque nationale d'Allemagne (traduction française : disponible à :
http://www.ifla.org/files/hq/papers/ifla76/91-svensson-fr.pdf

urn:nbn:de:101-2012052306