Histoire Épistémologie Langage
38/2 (2016), p. 101-113

DOI : 10.1051/hel/2016380208

Disponible en ligne sur
www.hel-journal.org

Comment indexer les corpus oraux ?

Pascal Cordereix

Bibliothèque nationale de France / Laboratoire Ligérien de Linguistique

avec l'aimable relecture de Michel Jacobson, CNRS / LLL

Résumé

La patrimonialisation des corpus oraux fait désormais partie de leur cycle de vie. Le geste de « mettre à part » (Michel de Certeau) qui caractérise toute entrée en archives amène notamment à un ensemble d'actions descriptives (inventaire, catalogage…) normées, qui vont permettre la consultation, la diffusion, l'exploitation et la conservation pérenne, etc. du corpus. Dans cet article, nous présentons certaines problématiques sous-jacentes à la description d'archives sonores dans le cadre d'une institution patrimoniale. Nous replacerons ces questionnements dans une perspective historique, des premières fiches descriptives à la fin du XIX^{e} siècle jusqu'aux modèles conceptuels de données du web sémantique et du web de données aujourd'hui.

Mots-clés

Archives sonores, corpus oraux, corpus de la parole, archive numérique, conservation pérenne, Bibliothèque nationale de France, métadonnées, web sémantique, web de données

Abstract

Becoming part of cultural heritage, patrimonialisation has now become a step in spoken corpuses' life-cycle. The action of 'putting aside, gathering' (Michel de Certeau) which characterizes any archiving leads in particular to a range of standardized descriptive processes (inventorying, cataloging…), which will provide catalogue consulting, dissemination, use and permanent preservation of the body of archives. In this article, we will develop some issues underlying sound archives description in the context of heritage institutions. We will put these issues from a historical perspective, dating back from the XIXth century written descriptive sheets to conceptual data formats in our today semantic web and linked data environment.

Keywords

Sound archives, spoken corpus, digital archive, sound preservation, National Library of France, metadata, semantic web, linked data

Dans cet article, nous présentons certaines problématiques sous-jacentes à la description d'archives sonores dans le cadre d'une institution patrimoniale. Nous replacerons ces questionnements dans une perspective historique, des premières fiches descriptives à la fin du XIX^e siècle jusqu'aux modèles conceptuels de données du web sémantique et du web de données.

1. Les prémisses

La question de la description des archives sonores est consubstantielle à leur création. Fondées en 1899 par Sigmund Exner dans le cadre de l'Académie des Sciences de Vienne, première institution en charge de la production et de la conservation d'archives sonores, les Phonogrammarchiv élaborent une fiche d'enregistrement descriptive qui va s'imposer comme un modèle pour les institutions à venir. Ainsi Ferdinand Brunot s'en inspire-t-il largement lorsqu'il crée les Archives de la Parole à la Sorbonne en 1911. La fiche comprend plusieurs rubriques relatives au « phonographié », aux circonstances de l'enregistrement et à ses caractéristiques techniques.

Dans le même temps, en 1893, le militant socialiste et pacifist belge Paul Otlet met en place un « répertoire bibliographique universel » qui s'appuie sur l'« Office international de bibliographie » qu'il crée avec Henri La Fontaine en 1895. L'idée de vouloir réaliser une bibliographie universelle recensant tous les ouvrages publiés depuis les débuts de l'imprimerie dans le monde repose sur l'utopie de l'accès au savoir pour tous, du savoir partagé par tous, comme fondement de la paix universelle. Pour faciliter cet accès à la connaissance, Otlet crée en 1905 un système de classification du savoir dit « classification décimale universelle », ainsi qu'un modèle standard de fiche bibliographique, au format 125 x 75 mm, un modèle qui va s'imposer dans les bibliothèques du monde entier jusqu'à l'apparition des catalogues informatiques. Et surtout, il va développer la notion de réseau et de coopération internationale entre bibliothèques, qui seront au fondement de tous les mouvements de normalisation au XX^e siècle.

Un autre mouvement de coopération internationale se dessine à partir des années 1920 sous l'égide de l'Institut International de Coopération Intellectuelle de la Société des Nations. On assiste notamment à une légitimation des « arts et traditions populaires » que traduit la tenue du Premier Congrès International des Arts Populaires à Prague, en Tchécoslovaquie, en 1928. Ce congrès consacre une large part de ses travaux à l'enregistrement et à la conservation des « chants et mélodies populaires », et affirme d'une part que la coopération internationale est indispensable et que des phonothèques doivent être mises en place comme lieu de conservation des archives sonores.

C'est précisément l'évolution qu'on observe en France. En effet, si au début des années 1930, le Musée de la Parole et du Geste, successeur depuis 1928 des Archives de la Parole, est encore la seule institution en charge de la production et de la conservation d'un patrimoine phonographique, le paysage de l'archive sonore va être profondément modifié au cours de la décennie à veni :

> En 1932, au Musée d'ethnographie du Trocadéro (futur Musée de l'Homme), André Schaeffner crée une phonothèque au sein du département d'organologie musicale qui devient département d'ethnologie musicale.
> En 1933, Philippe Stern crée une phonothèque dans le cadre de la section musicale du Musée Guimet.
> En 1938, sous l'impulsion de Jean Zay, la Phonothèque nationale voit le jour.
> En 1939, au Musée national des Arts et Traditions Populaires, Claudie Marcel-Dubois crée un service d'ethnographie musicale et une phonothèque attenante.

Les années 1930 sont donc un moment clé de l'archive sonore en France et traduisent une institutionnalisation et une professionnalisation de la collecte phonographique. Dans ce contexte, sous l'impulsion d'André Schaeffner, une tentative d'harmonisation des pratiques de description des enregistrements s'esquisse.

Plus aboutis sont, en 1952, aux Etats-Unis, la publication par la Bibliothèque du Congrès des *Rules for descriptive cataloging in the Library of Congress. Phonorecord* ; alors qu'en 1957, au sein de l'Association Internationale des Bibliothèques, archives et centres de documentation Musicaux (AIBM), la Commission internationale des Phonothèques est mandatée pour élaborer un *Code de catalogage des enregistrements sonores*.

2. Vers une normalisation

Dans l'univers des bibliothèques, la volonté d'unifier les pratiques catalographiques au niveau international apparaît dès 1961 avec la promulgation de « Principes internationaux de catalogage » (plus connus comme « *Principes de Paris* ») sous la double égide de l'UNESCO et de l'International Federation of Library Associations and institutions (IFLA). La Conférence internationale des experts en catalogage réunie en 1969 par l'IFLA à Copenhague adopte une résolution demandant la création de normes pour normaliser la forme et le contenu des descriptions bibliographiques établies dans les différents pays afin de faciliter le partage et l'échange international de l'information bibliographique. De cette résolution naît le concept de la description bibliographique internationale normalisée : ISBD (International Standard Bibliographic Description) et sa publication en 1971. L'ISBD

- précise les éléments requis pour une description bibliographique,
- prescrit leur ordre de présentation en les répartissant entre huit zones cohérentes[1],

1 Zone du titre et des mentions de responsabilité ; zone de l'édition ; zone de l'adresse, etc.

– définit une ponctuation pour les délimiter (ponctuation également utilisée comme codage),
– donne des règles pour leur transcription à partir de source d'informations clairement identifiées.

De 1971 à 2007, ce concept fut traduit en règles déclinées par type de ressource ou mode de publication, publiées en sept ISBD spécialisés[2] largement adoptés dans le monde et qui ont servi de référence pour les règles nationales ou supranationales de catalogage[3].

On notera toutefois que l'ISBD va surtout être mis en œuvre par les pays européens, les pays anglo-saxons développant leur propre modèle de catalogage connu sous le nom de AACR (Anglo American Cataloguing Rules)[4].

3. La première révolution : informatisation des catalogues et formats MARC

Si l'ISBD donne un cadre normalisé de catalogage, il est plutôt conçu à l'origine dans un contexte de catalogues de bibliothèques sous forme de fiches papier héritières de P. Otlet. Or à la fin des années 1960, et dans les années 1970, la question se pose de savoir comment transposer le cadre normatif de l'ISBD dans le contexte de l'informatisation des catalogues des bibliothèques qui se développe à cette époque. La problématique est ici double :

– comment structurer de manière normée les informations bibliographiques dans un contexte informatique, et éviter que chaque concepteur de logiciel n'« invente » son propre format de description ;
– établir l'indépendance de la description bibliographique par rapport au système informatique, c'est-à-dire pouvoir transférer les données bibliographiques d'un système à un autre.

Et dès 1965, la Bibliothèque du Congrès à Washington met au point le premier format « MARC », pour **MA**chine **R**eadable **C**atalogue, c'est-à-dire un catalogue lisible par une « machine », un ordinateur.

2 ISBD(M) pour les monographies (texte imprimé) ; ISBD(S) puis CR (pour ressources continues) pour les publications en série…

3 En parallèle, pour la France, on rappellera les publications de l'Association Française des détenteurs de documents Audiovisuels et Sonores, l'AFAS, avec notamment L'oral en fiches, en 1985 et 1987, et le Guide d'analyse documentaire, son édité et son inédit : mise en place de bases de données, en 1994 et 2001. En 2014, la Fédération des Associations de Danses et Musiques Traditionnelles, FAMDT a mis en ligne une version remaniée de ce guide, sous le titre : Patrimoine culturel immatériel. Traitement documentaire des archives sonores inédites. Guide des bonnes pratiques, https://halshs.archives-ouvertes.fr/halshs-01065125/document

4 Voir *infra*.

On mesure immédiatement l'enjeu : transposer la norme internationale qu'est l'ISBD dans les systèmes de gestion de l'information des bibliothèques. Schématiquement, MARC est un format informatique de structuration des données qui répond à une norme (ISO 2709) et qui permet la production de catalogues mais aussi et surtout l'échange de données bibliographiques. Aujourd'hui, il existe quasiment autant de variantes de formats MARC qu'il y a de pays mais, au plan international, deux formats MARC dominent : le format UNIMARC, plutôt « européen », le format MARC 21, qui est le format de la Bibliothèque du Congrès.

4. La seconde révolution du numérique

4.1. La « révolution » du numérique

Aujourd'hui, les professionnels de l'information ont à décrire non plus seulement des documents physiques mais aussi de plus en plus des ressources électroniques, « immatérielles », jusqu'à des sites Internet. On assiste ainsi à une diversification et à une complexification de la documentation

Aussi, un deuxième facteur de cette révolution est la place dominante que prend Internet aujourd'hui dans notre vie quotidienne : le bouleversement des usages et des attentes qu'il suscite, et la place qu'occupent les catalogues de bibliothèque dans l'écosystème du Web. Internet conduit à ce qu'on ne parle plus seulement d'échanges de données, mais d'*interopérabilité*[5] entre ces données.

Pourquoi ce besoin d'interopérabilité ? Pour agréger des données, pour constituer des catalogues collectifs, des portails de ressources qui vont être exposées sur Internet, et ceci avec des données qui peuvent être hétérogènes, ce que ne permet pas – ou peu –la rigidité des formats MARC par exemple.

Mais le besoin d'interopérabilité vient aussi et surtout de la nécessité de rompre l'isolement des bibliothèques face aux autres acteurs de l'Internet. L'interopérabilité permet d'échanger avec d'autres communautés pour que l'information bibliographique produite par les bibliothèques soit utilisée dans des applications extérieures à leur univers, un enjeu majeur aujourd'hui.

Cette révolution du numérique s'appuie sur deux constats très forts dans l'univers des bibliothèques.

Tout d'abord, il convient de replacer l'usager au centre du catalogue. « L'utilisateur » n'existe pas : il y a non seulement des utilisateurs, mais des utilisations, et différents contextes d'usage des notices : catalogue de bibliothèque, certes, mais aussi édition et diffusion, gestion des droits etc. L'usager est replacé

5 Nous portons en italiques un certain nombre de notions-clés que nous ne pouvons développer ici.

au centre du catalogue dans un contexte qui a radicalement évolué (multiplication des ressources documentaires hors bibliothèques, comme Internet), et où l'utilisateur cherche non plus un « document », mais de l'« information ».

Deuxième constat, celui des limites des ISBD et des formats MARC qui, précisément, ne sont plus adaptés à ce nouveau contexte. L'ISBD est trop rigide et ses huit déclinaisons par type de ressource ont amené à des incohérences. Les formats MARC posent des problèmes de complexité, de lourdeur de gestion, de rigidité et leur arborescence est très limitée (deux niveaux). De plus la multiplication des formats MARC vernaculaires fait que l'interopérabilité est restée un vœu pieux.

4.2. De XML à l'EAD

Enfin, ISBD et MARC isolent les bibliothèques dans un environnement web où le codage en XML : eXtensible Markup Language (langage à balise extensible) est devenu la règle. Un premier pas de l'immersion des bibliothèques dans l'écosystème du web passe donc par l'utilisation de schémas de métadonnées en XML, c'est-à-dire des langages à balises.

On prendra comme exemple de ces langages l'EAD, format intéressant aujourd'hui pour la description de fonds d'archives sonores. L'EAD (Encoded Archival Description) est une DTD (Document Type Definition) XML développée à l'initiative de la bibliothèque de l'Université de Californie, à Berkeley en 1993 et réactualisée par la Bibliothèque du Congrès aux Etats-Unis depuis 1995. Ses développements sont assurés par un groupe de travail placé sous la responsabilité de la Société des Archivistes Américains (SAA).

L'EAD repose sur la norme ISAD (G) (International Standard Archival Description-General : Norme générale et internationale de description archivistique). L'objectif est de créer un standard d'encodage des descriptions de documents d'archives qui contienne des informations beaucoup plus riches que les notices traditionnellement décrites en format MARC. Il s'agit de transposer informatiquement la richesse des « Instruments de recherche », c'est-à-dire les inventaires de fonds d'archives. Par rapport aux formats MARC, la caractéristique de l'EAD est d'offrir une très riche arborescence, là où les formats MARC sont limités à deux niveaux d'arborescence : une notice et une sous notice. En d'autres termes, au-delà de la description générale d'un fonds, l'EAD permet de construire une description hiérarchisée restituant précisément l'imbrication des composants et sous-composants, ce qui n'est pas possible avec les formats MARC. Le corolaire est évidemment de circuler facilement dans cette arborescence : du plus général au plus particulier et inversement. Cette richesse / hiérarchisation / arborescence / navigation explique notamment qu'aujourd'hui tous les fonds d'archives sonores de la Bibliothèque nationale de France (entre autres) sont décrits en EAD.

Mais avec l'EAD, on reste dans la transposition, le codage informatique de modèles antérieurs. Et avec les développements du web sémantique et du web de données (voir *infra*), une orientation majeure a entrepris de dépasser ce stade de la transposition avec le développement de *modèles conceptuels de données*.

5. La modélisation des données

5.1. Ontologies et référentiels

Aujourd'hui, toute conception d'un système d'information suppose une phase préalable de *modélisation*. Il s'agit d'élaborer une vision abstraite et synthétique du réel saisi dans un contexte déterminé et dont la réalisation va s'appuyer sur la définition d'*ontologies*. On peut concevoir l'ontologie comme un vocabulaire partagé par une communauté astreinte à partager de l'information dans un domaine donné. L'ontologie contient des définitions lisibles en machine des concepts de base du domaine concerné et de leurs relations. Au-delà, par le biais de « *mappings* » (des tableaux d'équivalence), la modélisation permet de mettre en relation et de faire *converger* des bases de données hétérogènes vers un modèle conceptuel commun (et non plus de format à format, comme c'était le cas avec les formats MARC). Afin d'éviter tout cloisonnement, et permettre la circulation et la récupération de l'information conformément aux principes du web de données (cf. *infra*) d'une communauté à une autre, d'un modèle de données à un autre, d'une ontologie à une autre, des passerelles doivent être construites, réalisées sous la forme d'*alignements* d'ontologies ou de *référentiels*.

On comprend donc qu'il n'y a pas un, mais *des* modèles de données, et qu'un modèle de données n'a de sens que s'il est partagé par une communauté d'utilisateurs d'une part, et d'autre part que s'il est capable de dialoguer avec – d'être ouvert à – d'autres modèles de données. On ne peut détailler ici l'ensemble des modèles de données existants, nous en citerons trois intéressant notre propos.

5.2. CIDOC CRM (Conceptual Reference Model)

CIDOC CRM[6] tout d'abord, est le modèle conceptuel de référence pour l'information muséographique. C'est un modèle sémantique qui constitue une « ontologie » de l'information relative au patrimoine culturel qui

> [...] vise fondamentalement à fournir un langage commun à des gisements d'information hétérogènes et à permettre leur intégration, par-delà leurs éventuelles incompatibilités tant sémantiques que structurelles. Il s'agit donc de faciliter l'échange et la recherche d'informations dans le domaine du patrimoine culturel et de permettre aux musées de rendre compatibles leurs documentations

6 Conceptual Reference Model / Modèle conceptuel de référence.

> sans rien perdre de leurs spécifici és ni du niveau de précision de leurs données actuelles. Depuis 2003, un rapprochement s'est opéré entre le modèle CIDOC CRM des musées et le modèle FRBR des bibliothèques, lequel a été reformulé selon le formalisme orienté objet pour devenir une extension du CIDOC CRM. La combinaison du CIDOC CRM et de FRBRoo constitue désormais un modèle conceptuel commun à l'information muséographique et à l'information bibliographique.[7]

5.3. FRBR (Functional Requirements for Bibliographic Records)

Le modèle FRBR[8] est largement implanté dans le monde des bibliothèques. Ce n'est pas une norme de catalogage mais une modélisation conceptuelle de l'information contenue dans les notices bibliographiques des catalogues de bibliothèques. Les FRBR sont le résultat d'une analyse qui part des besoins de l'utilisateur (lecteur ou personnel des bibliothèques, chercheur, mais aussi éditeur, distributeur, etc., en fait tout utilisateur potentiel d'un service d'information). Cette analyse a permis de dégager les entités fondamentales pour l'utilisateur, les attributs de ces entités et les relations entre entités.

Conçus originellement selon le formalisme « entité-relation », les FRBR organisent les différentes composantes de la description bibliographique (les autorités[9], les accès sujet et les informations sur le document proprement dites) en trois groupes d'entités reliées ensemble par des relations. Le premier groupe d'entités regroupe tout ce qui concerne les contenus et leurs différentes versions (œuvre / expression de l'œuvre / manifestation de l'expression ; item ou document).

La grande originalité du modèle FRBR est la notion d'« œuvre » qui permet, par exemple, de rapprocher un roman de ses traductions ou adaptations (quelle que soit leur forme), ce dont sont incapables les catalogues de bibliothèques, sauf par rebond à partir du titre ou de l'auteur.

Le second groupe d'entités correspond à la modélisation des personnes physiques ou morales qui ont une responsabilité dans le contenu intellectuel ou artistique, la production matérielle et la distribution, ou la gestion juridique des entités du premier groupe.

Le troisième groupe regroupe des entités qui sont le sujet des œuvres : concept, objet, événement, lieu.

Chaque entité peut être décrite par des attributs (le titre est par exemple un attribut de l'entité « œuvre »). Les entités font l'objet de relations entre elles (une œuvre est réalisée à travers une expression ; une expression est matérialisée dans

7 http://www.bnf.fr/fr/professionnels/modelisation_ontologies/a.modele_cidoc_crm.html consulté le 05 août 2015

8 Functional Requirements for Bibliographic Records / Fonctionnalités requises des notices bibliographiques.

9 Forme retenue d'un nom (auteur, sujet…) à laquelle on va rattacher toutes les variantes de ce nom.

une manifestation, etc. ; une œuvre est créée par une personne ou une collectivité ; une expression est réalisée par une personne ou une collectivité, etc.). Là aussi, ce type de relations (toutes les œuvres d'un auteur, tous les items qui appartiennent à une bibliothèque…) existent aujourd'hui dans les catalogues de bibliothèques, mais dans une forme bridée, guère exploitable sauf par les « rebonds » permis par les notices d'autorité.

Un des intérêts majeurs d'un modèle de données et donc du modèle FRBR est d'ouvrir les données bibliographiques sur un nouvel environnement. La notice bibliographique est analysée comme une superposition de niveaux dont chacun peut être *récupérable* dans un contexte donné ; le modèle énumère pour chacun de ces niveaux les éléments de données qui peuvent faire l'objet d'une recherche, enfin, il traite l'information bibliographique comme un *réseau* d'éléments de données liés entre eux.

5.4. RDA (Resource Description & Access)

Sur ce modèle FRBR s'appuie ce qui se dessine comme étant le nouveau code international de catalogage, RDA (Resource Description & Access). RDA est issu des règles de catalogage anglo-saxonnes : AACR. Comme pour l'ISBD, le besoin de faire évoluer les AACR de manière très profonde s'est fait jour dès les années 1990 pour prendre en compte l'informatisation des catalogues et l'explosion des ressources numériques, d'où ce nouveau code RDA.

RDA a pour caractéristiques essentielles :

- de reposer sur le modèle FRBR. Si l'organisation de l'information bibliographique selon les entités du modèle FRBR permet d'améliorer la présentation des catalogues, elle ouvre aussi la voie à une interopérabilité accrue des données des catalogues de bibliothèque par la médiation d'ontologies de référence, et vers une ouverture des catalogues sur le web de données ou web sémantique.
- d'être une norme de contenu (et non d'encodage ou de présentation) qui fait porter l'accent sur l'information requise pour décrire une ressource, qu'elle soit numérique ou traditionnelle, et non sur la manière d'encoder (format MARC ou autre) ou de présenter (ISBD ou autre) cette information.
- d'être conçu pour le monde numérique : RDA permet de décrire les ressources numériques, avec leurs spécificités et de récupérer automatiquement des métadonnées fournies par les créateurs ou éditeurs de ressources.
- de s'adresser à d'autres communautés que celles des bibliothèques. Un des objectifs est de faciliter l'intégration des notices bibliographiques à celles produites par d'autres communautés créatrices de métadonnées sur le web afin d'en permettre de nouveaux usages.

Ainsi, RDA, tout en conservant une structuration très forte de l'information, permet d'éclater le schéma traditionnel, « monolithique », d'une notice de bibliothèque et d'autonomiser chaque élément de la description pour en faire une information indexable et utilisable en soi.

On rappellera la nécessaire adéquation entre le modèle et les données à structurer en précisant à nouveau qu'il n'y a pas *un*, mais *des* modèles de données, adaptés à des usages et à des contextes spécifiques. Si le modèle FRBR est particulièrement efficace là où la notion d'œuvre – et ses déclinaisons en expression(s), manifestation(s) – est opérante, c'est probablement moins vrai dans d'autres contextes où, précisément, la qualification d'« œuvre » perd de son sens. C'est le cas notamment des corpus de la parole qui, s'ils se déclinent parfaitement dans les composants et sous-composants de l'EAD (cf. *supra*), ne peuvent être qualifiés d'« œuvres » au sens où FRBR définit ce terme. D'autres modèles de données, inscrits eux aussi dans le web sémantique et le web de données, peuvent constituer des alternatives ; ainsi du modèle de données développé par la bibliothèque numérique européenne Européana[10], EDM (Europeana Data Model) et plus particulièrement son extension aux documents sonores (EDM-S) dans le cadre du projet Europeana Sounds[11], EDM-S.

6. Web sémantique et web de données

Quel est l'enjeu ? On l'a suggéré à plusieurs reprises dans ce qui précède : c'est l'inscription des données de catalogue dans le web de données.

Par web sémantique, on entend ici un ensemble de technologies et de standards développés par le W3C[12] pour construire le web de données entendu comme une extension du Web permettant de relier non pas des documents (pages HTML) mais les données elles-mêmes, et de les rendre exploitables par les machines. Le web de données consiste à exposer des données structurées sur le web et à les relier entre elles, ce qui accroît leur visibilité et leurs possibilités de réutilisation.

Or aujourd'hui, avec les catalogues « traditionnels », d'une part les données sont entreposées dans des silos inaccessibles, et notamment inaccessibles par le web. On se situe dans le web profond, qui n'est pas « attaquable » par les moteurs de recherche. D'autre part, ces silos ne communiquent pas entre eux. Si j'interroge le catalogue de la BnF, les notices de la Bibliothèque du Congrès, par exemple, n'apparaissent pas. Ou plus exactement, si je veux faire communiquer entre eux

10 http://www.europeana.eu/portal/
11 http://www.europeanasounds.eu/fr/
12 World Wide Web Consortium : consortium international à but non lucratif regroupant plus de 380 entreprises partenaires, chargé de promouvoir la compatibilité des technologies du Web (HTML, XML, RDF, SPARQL…).

ces catalogues, je vais mettre en place des protocoles tels que Z 3950 ou OAI-PMH, c'est-à-dire des protocoles d'interopérabilité.

Un double problème subsiste : d'une part mes données ne sont toujours pas complètement dans le web de surface. Je crée une interface qui leur donne une visibilité, mais pour autant ces données sont toujours dans des entrepôts. D'autre part et surtout, je ne suis absolument pas dans l'écosystème du web. OAI-PMH, par exemple, est reconnu par la communauté des bibliothèques-archives-documentation et de la recherche mais c'est à peu près tout. Notamment, OAI-PMH ne fait pas partie des protocoles des moteurs de recherche (Google et autres).

On aura donc compris qu'en faisant « éclater » la notice comme on l'a vu plus haut, l'enjeu est de sortir les données du web caché pour les exposer sur le web en les inscrivant directement dans les principes de fonctionnement mêmes du web et notamment des moteurs de recherche. Ainsi pour être visible sur le Web, la bibliothèque doit :

- mettre à disposition des données ;
- dans l'écosystème du web ;
    - c'est-à-dire indexables par les moteurs de recherche ;
    - reliées avec d'autres données existantes ;
    - sur les chemins empruntés par les utilisateurs.

En faisant « éclater » ces notices, ce qui est relié, ce ne sont plus des documents, c'est-à-dire des pages html, mais les données qu'on rend exploitables par les machines. Cela correspond à la définition même du web de données (voir plus haut). FRBR, RDA, les modèles conceptuels de données d'une manière générale, permettent de segmenter la notice et d'exploiter séparément chaque élément d'information – chaque donnée bibliographique si l'on veut – indépendamment les unes des autres. Et le Web de données va s'appuyer sur les outils du web sémantique pour exploiter ces données, c'est-à-dire les rendre indexables par les moteurs de recherche et les relier les unes aux autres. Pour ce faire, il recourt à plusieurs technologies :

- des URI (Uniform Ressource Identifier / Identifiant uniforme d ressource) qui, comme leur nom l'indique, identifient une ressource sur un réseau (le Web par exemple) ;
- RDF (Ressource Description Framework) qui est un modèle de description des données dans lequel toute ressource est identifiée par une URI. Schématiquement, un document structuré en RDF est représenté sous la forme de triplets : {Sujet, Prédicat, Objet} où :
    - le sujet représente la ressource à décrire ;
    - le prédicat le type de propriété applicable à cette ressource ;
    - l'objet la valeur de la propriété.

Un ensemble de triplets RDF qui décrivent une ressource ou un ensemble de ressources composent un graphe.

– SPARQL (SPARQL Protocol and RDF Query Language) : élaboré par le W3C pour la construction de requêtes sur les données en RDF, SPARQL est à la fois un protocole, un langage de requêtes et un formalisme pour l'expression des résultats. Les requêtes SPARQL permettent d'interroger dynamiquement les données en RDF, sans télécharger l'ensemble des données brutes.

À l'heure actuelle, une des réalisations parmi les plus abouties s'appuyant sur le modèle FRBR et mettant en œuvre les technologies du web sémantique et du web de données est data.bnf.fr (http://data.bnf.fr/). L'objectif est d'exposer les données de la BnF dans le Web de données. Il s'agit de mettre à disposition les données du catalogue via des formats structurés permettant leur interopérabilité et leur mise en relation, soit avec d'autres ressources de la BnF, comme pour la bibliothèque numérique en ligne Galllica (http://gallica.bnf.fr/), soit avec des partenaires extérieurs possédant des données complémentaires comme *Wikipedia* par exemple. Concrètement, des pages Web relatives aux auteurs et aux œuvres, aux sujets, aux lieux, etc. sont créées à partir des données de la BnF, en reliant les contenus. Data.bnf.fr permet de réunir sur une même page toutes les informations relatives à un auteur ou à une œuvre, etc. Sont recherchées ici l'exposition et l'intégration de ces données structurées dans le web sémantique. L'objectif est évidemment de renforcer la visibilité des ressources de la BnF. Au-delà, le projet s'inscrit dans une démarche internationale d'ouverture des données publiques, institutionnelles, culturelles ou administratives (*linked open data*)

EN CONCLUSION

Aujourd'hui, pour une bibliothèque, un centre d'archives, un laboratoire, mettre à disposition un catalogue, offrir une bibliothèque numérique, ou être présent sur les réseaux sociaux n'est plus suffisant. Il est nécessaire de structurer les données suivant les règles du web sémantique afin de les exposer dans le web de données, de les placer sur le chemin des utilisateurs et de créer une interopérabilité qui leur permette de dialoguer avec les données produites par d'autres sites. L'enjeu est d'importance puisqu'il s'agit de rendre possibles l'appropriation, la réutilisation, la dissémination de ces données par d'autres communautés que le microcosme (bibliothèque ou archive ou laboratoire…) à l'origine de la production.

C'est à ce prix que la circulation de l'information deviendra effective : des instances de recherche aux institutions patrimoniales et à un public le plus large et le plus varié possible. Et c'est à cette condition que le rêve prémonitoire de Paul Otlet, de l'universalisme du savoir, se réalisera.

N.B. : cette présentation doit beaucoup aux pages « Pour les professionnels » du site Web de la BnF, auxquelles il conviendra de se reporter pour des développements plus longs que ce qui peut être exposé ici, notamment :

« Innovation numérique », rubrique « Web de données » :
http://www.bnf.fr/fr/professionnels/innov_num_web_donnees.html

« Catalogage et indexation », rubriques « Formats, encodage » ; « Modélisation et ontologies », « Protocoles d'échanges de données » ; « Autorités » :
http://www.bnf.fr/fr/professionnels/catalogage_indexation.html

Cette présentation doit également beaucoup aux travaux d'Emmanuelle Bermès, Françoise Leresche, Anila Angjeli, Patrick Le Bœuf ; qu'ils en soient remerciés.