Compte-rendu de l'Atelier labex n°1

LES ARCHIVES DE L'ETHNOMUSICOLOGIE : METTRE EN COMMUN, METTRE À DISPOSITION

Mercredi 27 mars 2013 de 9h30 à 17h30

Le premier Atelier du labex *Les passés dans le présent* portait sur l'un des projets développés dans son cadre : *Les Sources de l'ethnomusicologie*, projet de numérisation et de diffusion d'archives conservées conjointement à la Bibliothèque nationale de France, au musée du quai Branly et au Centre de Recherches en Ethnomusicologie (CREM-LESC, MAE Nanterre). Cette journée de séminaire proposait à la fois une réflexion sur les enjeux propres au décloisonnement entre des corpus d'archives d'ethnomusicologie complémentaires et sur les enjeux d'une diffusion numérique patrimoniale qui entend s'appuyer sur les pratiques d'usage et les technologies innovantes les plus adaptées. L'objectif double du projet, regrouper des collections éclatées et mettre à disposition du plus grand nombre un patrimoine commun a été confronté à l'analyse de spécialistes, invités à apporter un regard extérieur et distancié au projet

Liste des intervenants

Olivier BAUDE, Laboratoire ligérien de linguistique ; Délégation générale à la langue française et aux langues de France,

Anne-Florence BORNEUF, Cité de la Musique, Paris

Natalie BOURDEAU, Centre national RAMEAU, BnF

Pascal CORDEREIX, département de l'audiovisuel, BnF

Jean-Pierre DALBERA, Musée des intruments, Céret

Françoise DALEX, médiathèqe, musée du quai Branly

Ghislaine GLASSON DESCHAUMES, labex *Les passés dans le présent*

Christine GUILLEBAUD, CREM - LESC

Michel JACOBSON, Archives de France

Aude JULIEN-DA CRUZ LIMA, CREM - LESC

Jean LAMBERT, CREM - LESC

Jean-Luc MINEL, Modyco

Michel MINGAM, Centre national RAMEAU-BnF

Marie-Dominique Mouton, LESC

Nicolas PRÉVÔT, CREM - LESC

Dana RAPPOPORT, CASE

Pierre ROUILLARD, responsable scientifique et technique du labex *Les passés dans le présent*

Agnès SIMON, data.bnf.fr, BnF

Audrey VIAULT, département de l'audiovisuel, BnF

Sommaire

Introduction de l'Atelier labex

Présentation du labex par Pierre Rouillard

Le labex associe un certain nombre de partenaires sur le campus de l'Université Paris Ouest Nanterre La Défense : la Maison Archéologie et Ethnologie, René-Ginouvès (MAE), le Laboratoire d'ethnologie et de sociologie comparative (LESC), Archéologies et sciences de l'antiquité (ArScAn), Préhistoire et technologie, Histoire des arts et représentations (HAR), la Bibliothèque de documentation internationale contemporaine (BDIC), l'Institut des sciences sociales du politique (ISP), Modèles, Dynamiques, Corpus (Modyco). Sont également associés des partenaires extérieurs majeurs tels que la Bibliothèque nationale de France (BnF), le musée du quai Branly et le musée d'archéologie nationale de Saint-Germain-en-Laye. Ce dernier s'excuse de ne pouvoir participer à cet atelier. Le labex est ainsi une structure globale, dont les grands objets de recherche sont les médiations de l'histoire, la vie numérique, les politiques de la mémoire, l'appropriation de l'histoire. Cet atelier sur les archives de l'ethnomusicologie réunit plusieurs de ces items.

Les ateliers sont conçus pour être des ateliers généraux. Nous avions, un temps, songé à les appeler séminaires généraux mais, pour ne pas générer de confusions, nous avons préféré le terme « atelier ». Le mot « atelier », en outre, renvoie aux idées de mise en forme et d'élaboration.

Cet atelier sur les sources de l'ethnomusicologie est donc la première expérience, dont il faudra tirer les conclusions. D'emblée, nous pouvons faire une remarque autocritique, dans la mesure où, même si cet atelier est global, sont essentiellement présents celles et ceux qui sont partenaires d'un projet. Nous aurons à élargir la communication, les échanges au sein du labex, pour que tel ou tel atelier mobilise au-delà de celles et ceux qui en sont acteurs, qu'il y ait davantage d'intervenants extérieurs, davantage d'acteurs d'autres projets, de façon à améliorer ce que nous mettons en œuvre.

Nous remercions les responsables du projet « Sources de l'ethnomusicologie », Pascal Cordereix, Audrey Viault, Françoise Dalex, Claire Schneider, Aude Julien-Da Cruz Lima, Joséphine Simonnot, d'avoir accepté que leur travail fasse l'objet de ce premier Atelier, qui est donc mis en œuvre par eux, avec le comité de pilotage et avec, notamment, Ghislaine Glasson Deschaumes.

Cette journée abordera les questions de la numérisation, de la médiation, des vocabulaires, mais aussi les questions interinstitutionnelles… Au terme de la journée, nous aurons à réfléchir à la manière dont nous pourrions nous projeter vers l'extérieur et, comme nous y a invité le Conseil scientifique, vers le futur. En effet, si l'on a une idée de ce qu'il en sera du labex et de la recherche sur ces questions en 2015 voire en 2020, qu'en sera-t-il en 2050 ? Un deuxième Atelier labex aura lieu au mois de juin et sera consacré à la médiation de l'histoire. Il serait souhaitable qu'un maximum d'acteurs du labex soit présent.

Présentation de la matinée par Françoise Dalex, modératrice de la matinée.

Le projet sur les sources de l'ethnomusicologie s'inscrit dans le second thème du labex : « connaissance active du passé, pratiques et outils de transmissions ». Comme les autres projets de ce thème, il s'appuie sur des fonds documentaires que le labex contribue à numériser pour les valoriser et les mettre à disposition des publics. L'objectif de la journée est donc de conduire une réflexion collective pour approfondir les pistes de recherche, de dégager des solutions, élargir le débat à l'échelle des projets similaires portés par la BDIC, la BnF, le musée d'archéologie nationale ou le LESC.

La journée est organisée en deux parties et divisée en deux demi-journées. La matinée sera consacrée aux questions de mise en commun et l'après-midi aux enjeux de la mise à disposition. Chacune de ces deux demi-journées sera introduite par la présentation d'une thématique du projet *Sources de l'ethnomusicologie.* Ainsi, la matinée débutera par une présentation générale du projet et l'après-midi par un panorama, un état des lieux de toutes les problématiques technologiques liées aux spécificités de ce projet. Ensuite viendront les interventions des invités qui ont accepté de se joindre à nous et que nous remercions vivement. Cette journée sera donc l'occasion d'entendre, d'une part, des présentations de projets et des expériences qui nous semblent contrastives et permettent de mettre en perspective les orientations de tel projet numérique, et, d'autre part, la prise de parole des discutants qui viennent des différents champs et disciplines concernés par le projet : des ethnologues, des historiens, des professionnels de la culture et du patrimoine, des documentalistes et des professionnels des nouvelles technologies. Ces discutants sont impliqués en tant que professionnels, mais également en tant qu'usagers de ces projets et pourront enrichir notre réflexion. Enfin, nous laisserons le plus de temps possible pour des échanges, afin que chacun puisse s'exprimer, approfondir des questions qui auraient été peu abordées pendant les présentations et apporter sa contribution à cette journée.

Trois institutions, trois collections complémentaires, Aude Julien-Da Cruz Lima

Les collections mobilisées pour le programme *les sources de l'ethnomusicologie* dans le cadre du labex sont assez diverses, aussi bien dans leurs formes que dans leurs contenus : enregistrements sonores et audiovisuels, objets (instruments de musique), archives papier (notes de terrain, transcriptions musicales, etc.) et images (photographies)… Le contenu de ces collections est également très diversifié puisqu'elles ont été produites depuis plus d'un siècle et qu'elles couvrent l'ensemble du monde. Elles représentent un patrimoine de l'humanité considérable, étroitement lié à l'histoire des institutions qui les ont produites ou qui les ont reçues en dépôt et qui sont actuellement chargées de leur conservation, de leur gestion et valorisation.

Le projet *Les sources de l'ethnomusicologie* est à l'initiative de trois partenaires : la BnF (service des documents sonores du département de l'Audiovisuel), le musée du quai Branly (service de la médiathèque du département du patrimoine et des collections), et le Centre de recherche en ethnomusicologie (CREM, équipe spécialisée du LESC CNRS Université Paris Ouest), qui gère les archives sonores du CNRS- musée de l'Homme. Si le point central de ces collections est donc constitué de documents sonores et audiovisuels, il est primordial, dans le cadre de ce projet, d'envisager les liens avec tous les autres documents (sous leurs formes originales et numériques) et de considérer la cohérence documentaire entre les instruments de musique collectés sur le terrain, les notes de terrain, photograpgies, etc.

Présentation du fonds de la BnF[1]

Le service des documents sonores du département de l'Audiovisuel de la BnF a une longue histoire. Il trouve son origine dans les Archives de la Parole[2] , -créées par un linguiste, Ferdinand Brunot, avec l'aide de l'industriel Émile Pathé, en 1911 dans le cadre de l'université de Paris. Cette initiative constitue la première collection sonore institutionnelle en France. Ferdinand Brunot s'intéressait notamment à la langue parlée, qu'il considérait comme le véritable génie de la langue. Dans cette perspective, il a utilisé le phonographe pour conserver les manifestations de cette langue parlée. Il a notamment enregistré à l'époque les voix de Guillaume Apollinaire, d'Alfred Dreyfus… Il s'est également intéressé aux patois et aux dialectes, dans le cadre d'enquêtes de terrain qu'il a effectuées dans les Ardennes franco-belges, le Berry et le Limousin entre 1912 et 1913. Il a enregistré non seulement de la langue parlée, mais aussi beaucoup de répertoires musicaux traditionnels chantés et instrumentaux – par exemple, des joueurs de vièle et de cornemuse, notamment les gas du Berry ; des chants de labour, qu'on appelle briolées (décrits notamment par George Sand). Dans le Limousin, il a été à l'origine d'un des premiers travaux comparatifs, en enregistrant plusieurs versions d'un même chant.

Cette orientation musicale est poursuivie par le successeur de Ferdinand Brunot,

---

[1] Cordereix Pascal, « Les fonds sonores du département de l'Audiovisuel de la Bibliothèque nationale de France », Le Temps des médias 2/ 2005 (n° 5), p. 253-264

[1]URL : www.cairn.info/revue-le-temps-des-medias-2005-2-page-253.htm

[2] Voir Gallica : http://gallica.bnf.fr/html/enregistrements-sonores/fonds-sonores

Hubert Pernot, qui prend la direction des Archives de la parole en 1924. Il s'intéresse notamment à ce que l'on appelle à l'époque le « folklore musical ». Lorsque les Archives de la Parole deviennent le Musée de la Parole et du Geste, en 1928, la mission de travailler sur le folklore national et international est officiellement inscrite dans le projet de musée. Cette orientation se concrétise par les missions que Pernot mène en Roumanie, en 1928, en Tchécoslovaquie, en 1929 et en Grèce, en 1930. Les enregistrements réalisés par le Musée de la parole et du geste lors de l'*Exposition coloniale internationale* de Paris en 1931 sont une illustration majeure de ce travail, qui sera au fondement d'autres institutions, comme -la phonothèque du musée d'ethnographie du Trocadéro (futur musée de l'Homme) en 1932, dont nous reparlerons plus tard (c'est la première collection à être entrée dans cette phonothèque) ou encore le musée Guimet, ou le musée des Colonies. Une autre collection majeure à signaler est constituée des enregistrements réalisés lors du Congrès de musique arabe du Caire en 1932. Il n'existe que trois exemplaires de cette collection dans le monde, dont l'un avait été donné à l'époque au Musée de la Parole et du Geste.

Après le départ d'Hubert Pernot, c'est le poète et journaliste Roger Dévigne qui prend la direction du Musée de la Parole et du Geste. Ce dernier va recentrer les activités du Musée sur la collecte du folklore français, avec ce qu'il appelle les « croisières folkloristes », c'est-à-dire des enquêtes de terrain menées notamment dans les Alpes provençales en 1939, dans les Pyrénées en 1941-1942, en Normandie et Vendée en 1946. La phonothèque nationale, chargée du dépôt légal du disque, est créée en 1938. Elle est tout d'abord hébergée au sein du Musée de la Parole et du Geste. Cependant, progressivement, la logique s'inverse et ce sont les collections du musée qui sont intégrées à la phonothèque. La production d'archives sonores entre 1945 et 1975 est moindre. Elle est surtout le fait de dons de collecteurs et de chercheurs..

En 1975, la phonothèque nationale devient un département de la Bibliothèque nationale, avec à sa tête Marie-France Calas. Celle-ci insuffle une nouvelle dynamique à la collecte de l'oralité, aussi bien dans le domaine linguistique qu'ethnomusicologique, avec des dons d'archives, comme celles de Félix Quilici (enquêtes en Corse de 1961 à 1963) ou encore de Geneviève Massignon.

Cette politique s'intensifie avec la création de la Bibliothèque Nationale de France en 1994. La phonothèque nationale devient le département de l'Audiovisuel de la BnF, et la collecte de fonds sonores inédits une priorité. Actuellement, pour ne citer que les fonds les plus emblématiques, ce service a reçu en don et conserve les archives sonores de l'ethnologue Nicole Revel, de l'ethnomusicologue Simha Arom, de Charles Duvelle, de Deben Bhattacharya, etc. Le département de l'Audiovisuel mène une politique de collaboration avec des grandes institutions en ethnomusicologie, dont notamment le Musée national des arts et traditions populaires, devenu MuCEM[3], et l'ex-laboratoire d'ethnomusicologie du musée de l'Homme, aujourd'hui le CREM. Ces deux institutions ont pu déposer leurs fonds sonores à la BnF qui possède des capacités de conservation des supports originaux et d'archivage numérique pérenne. Ce dépôt permet également une ouverture à un public plus large qui dépasse la communauté scientifique.

[3] Musée des civilisations de l'Europe et de la Méditerranée

Présentation du fonds du CREM

Le CREM[4], anciennement Laboratoire d'Ethnomusicologie du CNRS et du musée de l'Homme, gère un vaste fonds d'archives sonores constitué depuis les années 1930. Propriété du CNRS et du musée de l'Homme (rattaché au Museum national d'histoire naturelle), ce fonds d'archives bénéficie d'un important soutien du ministère de la Culture, depuis les années 1990 et actuellement grâce à l'accord-cadre CNRS-MCC, ainsi que de la BnF depuis 2009, pour la conservation et la numérisation des supports (convention juridique BnF- CNRS - MNHN). Depuis son rattachement au LESC en 2006, le CREM est également sous la tutelle de l'université Paris Ouest où il est hébergé depuis 2009.

L'histoire de ces archives débute en 1932[5], avec la création par André Schaeffner de la phonothèque du département d'organologie au musée d'Ethnographie du Trocadéro, devenu en 1937 le musée de l'Homme. Il s'agit de rassembler à la fois des enregistrements sonores édités (publications en série par des firmes commerciales) et inédits (enregistrements dits de terrain). La première collection à entrer dans cette phonothèque est un exemplaire complet de la série de 174 disques des enregistrements effectués lors de l'exposition coloniale de 1931 à Paris par le Musée de la parole et du geste. Dès les années 1930s, la phonothèque est enrichie par les premiers enregistrements de terrain réalisés lors de missions ethnographiques, tout d'abord sur cylindres de cire[6], (André Schaeffner lors de la mission Dakar-Djibouti de 1931, Germaine Tillion et Thérèse Rivière en Algérie en 1936, etc.) et sur des disques dits à gravure directe[7] (Henri Clérisse à Madagascar en 1938-1939, Maurice Leenhard en Nouvelle Calédonie en 1939, Gilbert Rouget lors de la mission Ogooué-Congo en 1946, Pierre Gaisseau lors de la mission Orénoque-Amazone en 1948-1950). En 1943, la Société d'Anthropologie de Paris effectue un important dépôt de plus de 400 cylindres datant de l'Exposition Universelle de Paris de 1900 (enregistrés par Léon Azoulay). Le département entreprend également à cette période une série de publication d'enregistrements de musiques traditionnelles (les éditions du musée de l'Homme devenu CNRS musée de l'Homme jusqu'en 2001[8]). A partir des années 50s sont déposés les premiers enregistrements effectués sur bandes magnétiques (Gilbert Rouget en Afrique occidentale en 1952, Simone Dreyfus au Brésil en 1955, etc.). Depuis sa création, la phonothèque du musée de l'Homme est devenu un lieu de référence pour l'archivage des enregistrements sonores de terrains de nombreux chercheurs en ethnologie et ethnomusicologie parmi lesquels Jacques Soustelle, Monique et Robert Gessain, Constantin Braïloiu, Geneviève Dieterlen, Jean Rouch, Louis Berthe, Mireille Helffer, Monique Brandily, Georges Condominas, Jacques Dournes, Bernard Dupaigne, Marc Gaboriau, Geneviève Dournon, Pierre Sallée, Bernard Lortat-Jacob, Hugo Zemp, etc..

---

[4] http://www.crem-cnrs.fr/

[5] Pour une présentation historique des archives sonores du CNRS – musée de l'Homme, cf. P. Pitoëff, annuario degli archivi di etnomusicologia dell' accademia nazionale di santa cecilia, 1.1993, libreria musicale italiana, p.143-149. URL : http://recherche.ircam.fr/equipes/repmus/marc/ethnomus/archives/histo.html

[6] http://archives.crem-cnrs.fr/archives/fonds/CNRSMH_Cylindres/

[7] http://archives.crem-cnrs.fr/archives/fonds/CNRSMH_DisquesGravureDirecte/

[8] http://archives.crem-cnrs.fr/archives/fonds/CNRSMH_Editions/

Actuellement gérées par le CREM, ces collections forment un fonds patrimonial immatériel historique et prestigieux, à caractère culturel et scientifique. Régulièrement enrichies par le dépôt de documents sonores et audiovisuels (originaux ou copies, anciens ou contemporains), les archives du CNRS musée de l'Homme couvrent l'ensemble du monde et représente plusieurs milliers d'heures ainsi qu'une grande diversité de supports audio et vidéo : cylindres, disques à gravure directe, bandes magnétiques, disques microsillons, DAT, film 8 mm, VHS, DV, etc…– diversité commune aux fonds des 3 institutions, qui nécessite un important travail de documentation (description et indexation) et de numérisation toujours en cours. Depuis 2011, le catalogue et une partie des archives sont consultables sur la plateforme web audio Telemeta du CREM : http://archives.crem-cnrs.fr/ , site qui permet également de mettre en lien les ressources complémentaires conservées au sein d'autres institutions, notamment avec les objets et documents d'archives conservés à la BnF et au musée du quai Branly.

Présentation du fonds du Musée du quai Branly

Un autre aspect très important du projet *Sources de l'ethnomusicologie* concerne en effet les collections d'objets actuellement conservées au musée du quai Branly. Il s'agit de plus de dix mille instruments de musique, pour la plupart conservés auparavant au musée de l'Homme, ainsi qu'au musée national des Arts africains et océaniens, auxquels s'ajoutent des acquisitions récentes. Certains de ces instruments ont été collectés lors de missions ayant également généré des archives sonores et papiers. Il est essentiel de conserver la cohérence de tous ces documents. Madeleine Leclair, responsable de cette collection d'instruments de 2000 à 2012, également ethnomusicologue membre du CREM, a conduit dans le cadre du projet muséographique, un important travail de classification organologique et d'élaboration d'un thesaurus pour permettre l'accès en ligne à ces collections. Nous reviendrons par la suite sur ce thésaurus organologique[9] réalisé en collaboration avec Claire Schneider[10], car il sert actuellement de base à la mise en place d'un vocabulaire commun pour l'indexation des instruments de musique, dans le cadre du programme « Sources de l'ethnomusicologie ». Tous les instruments de la collection ont été numérisés et sont disponibles sur le portail documentaire du site internet du Musée du quai Branly[11].

---

[9] Ce thésaurus s'appuie sur la classification organologique édifiée en 1992 par Geneviève Dournon, conservatrice des collections d'instruments de musique au musée de l'Homme de 1967 à 1995. Cette classification est fondée sur le système de Sachs – Hornbostel, articulé avec celui d'André Schaeffner : « *Cette classification se fonde en partie sur une synthèse des principes énoncés par Sachs et Hornbostel dans leur Systematik des Musikinstrument (1914) et de ceux qu'André Schaeffner proposé à la fin de Origine des instruments de musique (1936). Elle comporte un notable aménagement des données et un développement des rubriques en fonction de critères organologiques. Son élaboration a bénéficié de plusieurs années de travail sur les collections du Musée de l'Homme, de recherches sur le terrain, d'enseignements dispensés aux jeunes chercheurs et de la participation aux travaux collectifs du CIMCIM (Comité international des musées et collections d'instruments de musique de l'ICOM, UNESCO)* » (G. Dournon, 1996, Guide pour la collecte des musiques et instruments traditionnels. Edition revue et augmentée. Paris : UNESCO, p.111).

[10] Chargée des fonds sonores et audiovisuels de la médiathèque du musée du quai Branly

[11] Catalogue des instruments de musique numérisés au musée du quai Branly

L'histoire de cette collection est très ancienne, puisque les premières acquisitions remontent à l'Exposition Universelle de 1878. Madeleine Leclair a repris l'historique de cette collection, en notant, notamment, l'impulsion donnée par Paul Rivet au Musée d'Ethnographie du Trocadéro, grâce aux objets rapportés des missions de recherche effectuées sur le terrain[12]. À ce titre, la mission Dakar-Djibouti a une importance particulière puisqu'André Schaeffner en a rapporté beaucoup d'instruments. Par la suite, des dépôts ont été réalisés par Gilbert Rouget, D. Gaisseau, A. Didier, notamment au retour des missions Ogooué-Congo ou Orénoque-Amazone.

Dès le début du projet, en 2003, le musée du quai Branly a mené une politique d'acquisitions et a également constitué une collection importante de documents sonores et audiovisuels. Principalement tournée vers des documents édités susceptibles de documenter les collections du musée, cette politique d'acquisition a néanmoins eu le souci de constituer des archives inédites. Actuellement, plus de 6000 enregistrements sonores et environ 1500 inédits dans le domaine des musiques de traditions orales sont conservés à la médiathèque du quai Branly. Sur ces 1500 inédits, on trouve des captations audiovisuelles de spectacles réalisés au musée, des enregistrements de terrain issus de collectes scientifiques, parmi lesquelles on peut citer notamment les fonds Gilbert Rouget, Francis Corpataux, Geneviève Dournon, Charles Duvelle, Madeleine Leclair, etc.

La médiathèque du musée du quai Branly conserve aussi un important fonds d'archives muséales et de documentation des collections issus notamment du musée de l'Homme et comprenant des dossiers des collections de fonds publics et privés d'ethnologues et d'ethnomusicologues, telles que les archives de Geneviève Dieterlen ou de Thérèse Rivière. Les archives de cette dernière, par exemple, sont très dispersées : les enregistrements sonores sur cylindre, effectués lors de la mission Algérie-Aurès en 1936 sont conservés au CREM, alors que les archives papier de cette même mission (carnets de terrain, rapports des missions, dessins, transcriptions musicales) sont conservées au musée du quai Branly.

Le musée du quai Branly possède également une importante collection photographique, soit plus de 700 000 pièces, qui concernent les missions scientifiques et l'histoire des collections du musée. Citons parmi les photos celles de la mission Ogooué-Congo, de la mission Orénoque-Amazone, ou des missions de Geneviève Dieterlen et d'André Schaeffner.

[12] Madeleine Leclair, « Les collections d'instruments de musique au futur musée du quai Branly », Cahiers d'ethnomusicologie [En ligne], 16 | 2003, mis en ligne le 16 janvier 2012, consulté le 25 février 2013. URL : http://ethnomusicologie.revues.org/595

Un projet commun, Pascal Cordereix

En France, l'ethnomusicologie a une histoire institutionnelle extrêmement cloisonnante, pour ne pas dire clivante, scientifiquement parlant. Cette discipline s'est structurée, à partir de 1937, entre le musée national des Arts et traditions populaires, d'un côté, et le musée de l'Homme, de l'autre. Une véritable césure s'est opérée à ce moment-là, et pour des nombreuses années. Ce cloisonnement et ce clivage se sont révélés des freins considérables, dès qu'il s'est agi de porter un regard historique sur les collections et, plus particulièrement, sur les collections sonores constitutives de l'ethnomusicologie. C'est dans la prise de conscience de ce blocage qu'il faut trouver l'origine du projet qui nous occupe aujourd'hui.

En effet, en 2003, dans le cadre de ses programmes triennaux de recherche, la Bibliothèque nationale de France (BnF) avait validé un projet de recherche qui consistait en l'élaboration d'un catalogue collectif des collectes sonores à caractère folklorique et ethnographique, communes à la BnF, au musée de l'Homme de l'époque, au musée national des Arts et traditions populaires (aujourd'hui le MuCEM), au musée national des Arts asiatiques Guimet et au musée du quai Branly. Il s'agissait de mettre en place une base documentaire commune et transversale des archives sonores ethnomusicologiques conservées par les cinq institutions citées. La justification du projet, à cette époque, était formulée ainsi : « les cinq institutions concernées par le projet ont en commun, pour des raisons diverses, la particularité d'offrir une lisibilité et une accessibilité des fonds évoqués ci-dessus, [c'est-à-dire les fonds ethnomusicologiques], que l'on peut qualifier comme allant de moyenne à très aléatoire, tant en terme de signalement documentaire qu'en terme d'accès physique. Qui plus est, une vision transversale de ces fonds inter-institutions est totalement hypothétique. En d'autres termes, il est impossible à l'heure actuelle, en 2003, à un chercheur d'embrasser le champ de la collecte sonore institutionnelle. C'est donc à cette lacune que tend de répondre le projet que nous soumettons aujourd'hui [en 2003]». Ce programme a été mené dans le cadre de deux programmes triennaux de recherche de la BnF, entre 2004 et 2009. Il a réussi sur deux points : il a permis un premier inventaire de collections, lequel n'avait jamais été fait, et il a permis aux cinq institutions citées de travailler ensemble sur un projet commun pour la première fois. Ce programme était néanmoins inabouti, dans la mesure où, d'une part, l'inventaire n'a pas été publié en tant que tel et où il reste, d'autre part, assez sommaire, dans un contexte qui a fortement évolué.

Revenons donc sur l'évolution de ce contexte, en mettant à part la situation, un peu compliquée du point de vue des archives sonores, du MuCEM, et la situation, encore plus complexe, des archives sonores au musée Guimet. Pour les trois autres institutions ici présentes, la BnF, le CREM, le musée du quai Branly, la situation a considérablement évolué, de deux points de vue. Tout d'abord, un effort très important de numérisation a été fait, tant quantitativement que qualitativement (utilisation de standards et de normes dans lesquelles se sont inscrites les différentes institutions). L'aboutissement de ce travail est la mise au point de solutions d'archivage numérique pérennes. La numérisation n'est évidemment pas achevée, mais il faut souligner les énormes progrès accomplis. Cette numérisation va de paire avec un traitement documentaire qui, par rapport à 2003, a également considérablement évolué. Les collections du CREM sont accessibles en

Dublin Core[13] sur Telemeta[14], celles du musée du quai Branly le sont en UNIMARC sur le site du musée et dans le Sudoc[15], celles de la BnF, le sont en INTERMARC et en EAD dans les deux catalogues de la BnF. On a donc, aujourd'hui, des bases de données avec des formats différents, mais qui sont toutes interopérables, parce qu'elles respectent toutes des normes très précises de description, d'indexation, de catalogage.

Ce travail en commun engagé depuis 10 ans et les avancées citées sont le terreau du programme discuté aujourd'hui, autour des sources des archives de l'ethnomusicologie. Quel est l'objectif de ce nouveau programme ? Il est précisément de permettre une diffusion la plus large possible de ces fonds patrimoniaux ethnomusicologiques, grâce à des modes d'accès innovants sur le web pouvant faciliter à la diffusion auprès d'un très large public. Il s'agit vraiment du cœur du programme tel qu'on a pu le définir et tel qu'on le porte. Ceci posé, il est intéressant de pointer les différences et les continuités avec le programme de recherche des années 2000.

La première différence de taille, est que, à l'époque, on ne parlait que de la constitution d'une base de données, parce qu'on ne pouvait pas parler d'autre chose. Aujourd'hui, on parle de la diffusion des contenus eux-mêmes et de l'accessibilité des contenus eux-mêmes sur le web. Ceci induit une autre différence : alors que le projet des années 2000 était un projet fermé sur les archives sonores, celui d'aujourd'hui, s'ouvre à la diffusion non seulement des archives sonores, mais aussi des archives papier, photographiques, des objets, c'est-à-dire des instruments eux-mêmes, comme le précisait Aude Julien- Da Cruz Lima. Le projet se fonde sur la restitution du contexte d'un objet. Une autre forme de continuité est celle de la notion centrale de transversalité. Aude Julien- Da Cruz Lima en a cité quelques exemples : la collection d'instruments du quai Branly est inséparable des archives sonores du CREM ; les archives sonores du CREM sont elles-mêmes inséparables de celles du Museum national d'histoire naturelle (MNHN) ; les archives de Gilbert Rouget sont déposées pour partie au CREM, pour partie au quai Branly, pour partie à la BnF ; les archives sonores de Charles Duvelle ont été données à la BnF, alors que la partie film l'a été au musée du quai Branly … On pourrait multiplier les exemples de ce type qui sont peut-être la caractéristique de cette discipline, de ce domaine. Une dernière forme de continuité enfin, est la notion de déplacement du scientifique au patrimonial, déplacement d'un objet constitué scientifiquement à la base et qui devient un objet patrimonial, qui implique un questionnement sur ce que cela induit en termes de description, d'indexation, de présentation et de constitution de cet objet. Il conviendra d'apporter une attention particulière à ce point, qui sera un critère de réussite du projet.

Si on considère que l'objectif du projet *les sources de l'ethnomusicologie* est l'accès du plus grand nombre à ces archives sur le web, les actions à mettre en œuvre pour permettre cet accès sont de trois types :

- Tout d'abord, le projet doit continuer le travail de documentation, voire de re-documentation des archives, ce qui passe par la mise en relation des collections pour créer un réseau de documents. En parallèle, il s'agit aussi de poursuivre la

[13] Pour plus d'informations sur Dublin Core, consulter le site (en anglais) : http://dublincore.org/

[14] Consulter le site Telemeta : http://telemeta.org/

[15] Consulter le site du Sudoc : http://www.sudoc.abes.fr

numérisation des documents.

- La deuxième action, absolument centrale dans le projet, consiste à mettre en œuvre un référentiel commun pour les trois collections, avec l'enrichissement et la mise en cohérence du langage d'indexation RAMEAU (Répertoire d'autorité-matière encyclopédique et alphabétique unifié) dans le domaine de l'ethnomusicologie (instruments de musique, voix, danse, etc). Ce travail est déjà entamé et sera évoqué plus tard par Michel Mingam et Natalie Bourdeau. Ce référentiel a un rôle fondamental dans notre projet, car c'est le pivot sur lequel va reposer, en grande partie, l'exposition des métadonnées et des données liées à nos collections dans le web sémantique, dans le web des données, qu'Agnès Simon expliquera ensuite.
- La troisième action est la mise en place d'outils de diffusion numérique innovants et elle constitue le plus grand défi du projet. Françoise Dalex y reviendra plus tard. Il n'est pas question que le portail reproduise à l'identique Telemeta, ou Gallica, ou le site du quai Branly, etc. Quelque chose d'innovant doit être mis en place. Pour la description, s'appuiera-t-on sur une DTD-EAD? On aura l'illustration d'une réalisation de ce type avec Jean-Pierre Dalbera et Audrey Viault. Est-ce qu'on s'appuiera sur un schéma TEI ? Est-ce qu'on s'appuiera sur un mixte EAD-TEI ? Envisagera-t-on une autre solution ? La question est ouverte, il est encore beaucoup trop tôt pour le dire. C'est l'objet même du projet que de définir le mode de diffusion et le mode de description.

En conclusion, si la BnF, le CREM et le musée du quai Branly constituent le socle de ce programme autour des archives de l'ethnomusicologie, les trois institutions n'ont pas vocation à y rester seules sur ce projet. En soi, le programme appelle à d'autres collaborations nationales, avec le Museum national d'Histoire naturelle par exemple, et d'autres institutions conservant des archives papier des collectionneurs du CREM. Une collaboration internationale est également envisageable avec, par exemple, le musée d'Ethnographie de Genève, les Phonogramm-Archiv de Berlin ou de Vienne, etc. L'avenir du programme passe par ces partenariats à nouer dans un futur assez proche.

Corpus de la parole : retour sur expérience

Présentation du projet « Corpus de la parole » par Olivier Baude

Le programme *Corpus de la parole* est une initiative de la Délégation générale à la langue française et aux langues de France (DGLFLF), une direction du Ministère de la Culture, qui a en charge de proposer la politique linguistique et de la mettre en œuvre. C'est au sein de l'Observatoire des pratiques linguistiques de la DGLFLF, qui a pour vocation de faire un lien entre le savoir scientifique sur les langues et les politiques linguistiques, que s'est concrétisé le programme *Corpus de la parole*, grâce à un partenariat entre le Ministère de la Culture et le CNRS.

Les archives de la parole ont bouleversé la science linguistique car, comme le soulignait déjà Ferdinand Brunot, à partir du moment où l'on peut enregistrer la parole, elle devient un nouvel objet d'étude. C'est un bouleversement similaire auquel on assiste une centaine d'années plus tard, avec le passage au numérique. Le Corpus de la parole tient au fait que les linguistes étaient, pour la plupart, très éloignés de la préoccupation de conserver de l'oral et de le rendre disponible. Deux exemples illustrent ce propos. Dans les années 1960, il y a eu l'expérience du français fondamental, un corpus d'ambition nationale qui visait à la description et l'enseignement du français. Dès le début, les responsables du programme ont souligné qu'il était trop coûteux de conserver les données et que les disques d'enregistrement seraient effacés systématiquement. La plupart des données a donc été perdue. Le deuxième exemple est celui du corpus d'Orléans, avec 350 heures d'enregistrements de français authentique de la fin des années 1960, réalisé par des Anglais aux fins d'enseignement de la langue. Ces données ont donné lieu à beaucoup de recherches en linguistique. Conservées en Angleterre, traitées aux Pays-Bas et en Belgique, elles ont failli être détruites lorsque le département de français de l'université anglaise qui avait conduit cette enquête a fermé. Elles ont été récupérées *in extremis* par le laboratoire d'Orléans.

Le programme Corpus de la Parole a donc été initié pour conserver les données et les rendre disponibles et pour permettre aux linguistes de conduire une réflexion sur la mutation de leurs pratiques. L'accessibilité des corpus supposait de traiter les aspects juridiques, avec l'élaboration d'un guide de bonnes pratiques, d'établir des normes et des pratiques d'interopérabilité, de gérer un réservoir de corpus numérisés, de créer un site de valorisation – avec le passage des opérations de conservation, de numérisation d'archives scientifiques à la constitution d'un objet de patrimoine que cela supposait. Un conseil scientifique associant la DGLFLF, la BnF, l'INA, le ministère de la Culture et plusieurs composantes du CNRS a été formé.

Le plan de numérisation a permis de relever des difficultés de plusieurs ordres : l'hétérogénéité des objets compris dans les corpus, avec des enregistrements, des transcriptions, de la documentation papier, des corpus enrichis (annotations, TAL) ; les problèmes juridiques ; l'absence de politique des chercheurs, voire des « pratiques sauvages » d'archivage et de mise à disposition (sur ces plans, les choses ont changé en quelques années.) ; l'impossibilité de recourir à des services externes pour les métadonnées, du fait de la méthodologie exclusive des chercheurs (choix théoriques, description scientifique, transcription, traduction) et la nécessité de croiser celles-ci avec

les institutions de conservation (techniques de numérisation, catalogage, indexation). La numérisation a transformé cet objet scientifique ainsi que les pratiques de recherche sur ledit objet. L'impact a également touché toute la chaîne de relation à l'objet : collecte, conservation, traitement scientifique, diffusion publique.

À ce jour, plus de mille heures ont été numérisées. Le volume, qui concerne plus de quarante langues et des dizaines de projets a généré un grand nombre de transcriptions, traductions, annotations, encore accru par l'utilisation des outils de traitement automatique du langage, et il se traduit par un catalogage de plusieurs milliers de documents décrits.

Présentation des aspects techniques du projet par Michel Jacobson

Le projet débute en 2005-2006 avec deux évènements concomitants. D'une part, le CNRS, se préoccupant des sources primaires de la recherche, crée une direction de la formation scientifique ; un appel à propositions est lancé en 2005, qui débouche sur la création du Centre des ressources numériques. Ce centre de compétences est organisé en fonction de la nature des sources de la recherche : corpus oraux, corpus textuels, sources historiques non imprimées, sources spatiales numériques, sources visuelles. Deux propositions sont retenues pour créer un centre de ressources sur les « corpus oraux ». L'une, centrée sur outils de la recherche, est portée par le LPL [16](Aix-en-Provence) ; l'autre, centrée sur l'entrepôt des données, donc la gestion documentaire, est portée par Lacito[17] (Paris).

De manière concomitante, la DGLFLF obtient un financement dans le cadre du plan de numérisation du ministère de la Culture, sur le thème des langues de France et s'engage avec le CNRS dans un partenariat pour organiser la production, la gestion, la diffusion de ressources autour des langues de France. L'apport, côté CNRS, s'appuie sur 3 organismes.

- Les fédérations de linguistique (TUL + ILF) sont chargées, au sein de leurs réseaux respectifs, de faire émerger des corpus, d'évaluer les réponses, et de suivre les projets qui obtiendront les financements.
- L'Inist[18] se voit confier le développement d'un portail d'accès et de présentation des ressources récoltées dans le cadre du projet.
- Enfin, le nouveau centre de ressources CRDO[19] est chargé de la gestion des ressources collectées dans ce cadre, puis, une fois le portail développé, de son hébergement.

Pour la création du portail, un certain nombre de choix technologiques ont été faits. Il fallait un gestionnaire de contenus (CMS) facilitant la gestion collaborative (plusieurs rédacteurs DGLFLF, CNRS, Universités). SPIP a été retenu. Il fallait un mode de relation entre le portail et l'entrepôt de données (protocole standardisé), et l'on a retenu, à cette

16 Laboratoire Parole et Langage

17 Langues et civilisations à tradition orale

18 L'Institut de l'Information Scientifique et Technique du CNRS

19 Le Centre de ressources pour la description de l'oral

époque, l'OAI-PMH (Archives ouvertes). Par souci d'indépendance entre les deux organisations, on a souhaité que les métadonnées soient récupérées par « moissonnage » et qu'elles alimentent une base locale dédiée (sans qu'il soit nécessaire de faire des requêtes d'entrepôt à chaque fois).

Pour le traitement des ressources, on a dû choisir des formats de représentation : le format WAV pour les enregistrements audio, avec une qualité plancher recommandée par IASA[20], les laboratoires engagés dans la numérisation n'ayant pas forcément le même équipement ; le format XML avec encodage DTD pour les annotations ; pour la description de métadonnées, le format XML avec encodage OLAC [21](the Open Language Archives Community).

Le rôle des fédérations de linguistique est la collecte, avec l'identification, la préparation puis le versement (avec dépôt initial, ajouts, mises à jour) de corpus dans l'entrepôt de ces ressources. Le rôle du centre de ressources CRDO est la gestion des archives collectées, leur conservation et stockage, puis l'accès à ces données. (Notons que, dans cet entrepôt, qui comporte des enregistrements audio et des annotations d'enregistrement, il n'y a pas que la collection des Corpus de la parole.) À partir des données gérées dans l'archive ouverte du CRDO, le portail a été monté. Il va moissonner les ressources propres à la collection *Corpus de la parole*. Le portail est organisé autour d'un certain nombre de fonctionnalités visant à présenter, de manière rédactionnelle, le projet (que sont les langues de France ; les décrire une par une, etc.). Ces données sont stockées uniquement sur le portail. Il y a aussi un accès multimédia aux ressources, via une interface qui permet d'aller les piocher dans l'entrepôt. Des moteurs de recherche dans le catalogue (ou base de données) ont été définis pour aller puiser dans les métadonnées (mots-clefs, cartographie, etc.) et l'on a mis en place une recherche d'occurrence de mots pour également interroger la base de données.

En 2008, le CNRS prend conscience qu'il ne faut pas perdre les données numériques stockées, mais il n'existe pas de procédés d'archivage. S'engage alors le processus du TGE-Adonis sur l'archivage pérenne des sources primaires de la recherche en sciences humaines et sociales. Son architecture adopte le modèle fonctionnel de traitement de l'information proposé par la norme OAIS[22]. Le TGE-Adonis s'appuie sur plusieurs centres qui combinent différentes compétences. Le premier, le CINES[23], reprend pratiquement l'ensemble des briques fonctionnelles de l'OAIS. Il élabore la brique « entrée » : comment valider les entrées, les stocker, les pérenniser et comment y donner accès ? La seule limite est que le CINES ne donne accès à ces ressources qu'à leurs seuls producteurs. Donc pour élargir l'accès à d'autres publics, tant les chercheurs que le public plus large, on s'adjoint, pour la brique « accès de l'utilisateur », le Centre de calcul de l'IN2P3[24], qui fabrique des formats de diffusion, affecte des URL pour l'accès aux ressources numériques, et remplit, par ailleurs, un certain nombre d'autres fonctionnalités comme la gestion des droits d'accès ou celle d'un entrepôt OAI. Le troisième acteur est le

---

[20] International Association of Sound and Audiovisual Archives : http://www.iasa-web.org/

[21] Pour consulter le site du OLAC : http://www.language-archives.org/

[22] Open Archival Information System, soit système ouvert d'archivage de l'information

[23] Centre Informatique National de l'enseignement supérieur

[24] Institut national de physique nucléaire et de physique des particules

TGE-Adonis, qui définit le mandat, attribue des ressources et peut faire des arbitrages.

Dans le cadre de ce nouveau programme, le CRDO intervient en tant que projet pilote pour tester cette organisation destinée à conserver sur le long terme des ressources orales. À cette fin, on lance une étude commune aux Archives de France, au CINES et au TGE-Adonis sur les formats manipulés par les personnes travaillant sur l'oral. Il s'agit d'étudier l'ensemble des formats existant autour de l'audio et de la vidéo pour mettre en place une méthodologie : quels sont les formats et les codages à retenir pour pouvoir conserver sur le long terme ce type d'information ? On aboutit à un guide, utilisable en tant que tel, puisque la méthodologie est la même quelque soit le type d'objet que l'on manipule. Mais l'étude aboutit aussi à des choix de formats qui seront utilisés par le CINES pour accepter l'entrée des données. Pour l'audio, on va définir des conteneurs acceptables, puisque reposant sur des spécifications publiques, normalisées, libres, outillées, etc. Ces formats sont le WAVE, le AIFF, le FLAC, mais avec une restriction quant aux codages à utiliser. Le WAVE et l'AIFF ne peuvent être utilisés qu'avec le codage PCM qui est un codage sans compression, FLAC étant en lui-même son propre codage. Pour les formats vidéo (qui comportent d'autres données comme des sous-titres, des ancrages temporels), on retient les formats MP4, MKV, OGG, avec le codage H264. Si les données sont mal normées à l'entrée, l'outil ne saura pas les gérer et il faudra au bout d'un certain temps opérer des migrations, changer de format, etc.

Le CRDO cumulait ces fonctionnalités sans que le CNRS ait vraiment conscience du problème que cela posait de tout mettre au même endroit et d'adosser cela à une organisation potentiellement fragile. Sur ce que le TGE-Adonis appelle sa grille de services, il y a un service de stockage sécurisé. LE CRDO va donc déléguer la responsabilité de la conservation pérenne des ressources au TGE-Adonis. Le CINES affecte les identifiants pérennes aux ressources en utilisant un système de type ARK, assez répandu et utilisé par la BnF, beaucoup de centres d'archives en France, aux États-Unis). Une partie des fonctions d'accès est aussi déléguée au TGE-Adonis : la conversion des formats de conservation (très riches, et donc inadaptés à une utilisation web) aux formats de diffusion, affectation des URL, contrôle des accès, etc. La dernière chose que l'on délègue, c'est l'hébergement des applications web du centre de ressources CRDO (CoCoon) et du portail Corpus de la parole.

Ces choix d'organisation permettent une indépendance entre l'entrepôt (gestion des ressources) et les portails d'accès pour la valorisation, une même ressource va pouvoir être accessible sur plusieurs portails. Une même ressource peut être présentée de différentes manières. Elle pourra être présentée sur le portail d'un projet scientifique de collecte (par exemple, le « corpus de Français Parisien parlé dans les années 2000 » ; l'ANR Epopée Népal), sur le portail d'un laboratoire (par exemple le Lacito) parallèlement à la présentation du laboratoire et de ses travaux, sur un portail thématique (celui sur les langues de France ; celui des langues de la zone tibétaine et himalayenne). Beaucoup de portails peuvent ainsi se constituer en allant piocher les informations dans cet entrepôt. De la même manière, le chercheur peut déposer ses pré-print ou post-print dans HAL[25], les montrer dans sa propre bibliographie et/ou dans les ressources de son laboratoire les archives des différents projets auxquels il participe.

[25] Hyper articles en ligne

Les choix de codage des métadonnées ont également des conséquences. On a voulu aller vers le plus standardisé, le plus normalisé et le plus explicite possibles, avec la normalisation OLAC (basé sur le Dublin-Core qualifié), qui porte sur les dates, les indications spatiales (longitude, latitude, altitude, points de requête), l'identification des langues (référentiels des normes ISO-639-1 et 639-3, qui permettent de parler d'une langue en en partageant la définition), etc. Cette normalisation (utilisation de standards comme l'EAD, le Dublin Core, ARK), qui porte sur les ressources et sur les métadonnées, permet l'interopérabilité, donc l'utilisation de composants existants, le développement de nouvelles fonctionnalités, la réutilisation des ressources par des tiers, l'établissement de liens avec d'autres sources d'information.

Par exemple, si l'on va sur un portail de ressources linguistiques pour la communauté des linguistes et que l'on cherche, sans savoir qui a constitué les ressources, quelque chose sur le picard entre les années 1950 et 2000, on va trouver des dialogues, des récits, un certain nombre de contributeurs, et on trouvera des ressources dans deux entrepôts : COCOON[26] (nouveau nom du CRDO) et WALS[27]. Sans pour autant savoir où se trouvent les ressources, on peut y accéder simplement en étant intéressé par le contenu des métadonnées.

La dernière chose que je voulais préciser, c'est que les descripteurs de ces ressources peuvent être dans des référentiels externes. Toute l'organisation du web de données est basée sur ce principe, selon lequel des référentiels exposés sur le web pourront être utilisés moyennant le partage d'un certain nombre d'identifiants, d'éléments stables souvent basés sur des normes. Par exemple, si l'on a affaire à un enregistrement de Mayotte, il est possible de savoir où est Mayotte, quel en est le nombre d'habitants, quelles langues y sont parlées. S'il s'agit d'une langue, il est aussi possible de voir comment elle est décrite dans un référentiel qui décrit les langues, qui redirigera vers d'autres référentiels fournissant d'autres types d'information. Donc, on trouvera beaucoup de choses dans DBpédia[28], sous forme de fiches, extrêmement volumineuses en termes de documentation.

Deux pistes sont envisagées pour l'évolution du projet. D'une part, l'utilisation d'une couche supplémentaire de description, par-dessus le modèle OLAC qui décrit des ressources, qui permettrait de décrire non plus des ressources, mais des fonds. Le modèle de l'EAD[29] est conçu à cette fin et beaucoup utilisé, à l'origine par les archives et maintenant aussi par les bibliothèques. C'est un modèle en pleine évolution. D'autre part, nous souhaiterions explorer le modèle RDF[30] et l'organisation qu'on appelle aujourd'hui le web de données pour exposer plus facilement nos ressources, nos descripteurs de ressources, et pour les placer dans l'écosystème du Web aujourd'hui en établissant plus facilement des liens avec les autres sources des données.

---

[26] COllections de COrpus Oraux Numériques : http://cocoon.tge-adonis.fr

[27] World Atlas of Language Structures (WALS) : http://wals.info/

[28] Consulter le site de Dbpédia : http://fr.dbpedia.org/

[29] Encoded Archival Description

[30] Resource Description Framework

Sur l'importance des formats :

Les formats sont importants, parce qu'ils permettre la conservation des données. Les informations sont traitées dans un format qui donne la description du contenu et en permet une interprétation correcte. Si on n'a pas la description, il est plus difficile de conserver cette information. Comme il faut être sûr de pouvoir à tout moment interpréter ces données, une attention particulière est portée aux outils d'exploitation. On privilégiera donc plutôt les formats normalisés ou standardisés ou, au moins, pour lesquels les spécifications sont publiées. Cela participe de la méthodologie préconisée par le guide qu'a réalisé le CINES avec le SIAF[31] et le TGE-Adonis, sur le choix de formats pour la conservation. Un certain nombre de critères ont été établis, qui vont du plus propriétaire, à proscrire, au plus normalisé, à promouvoir. Mais cela n'est si simple, parce qu'entre en ligne de compte le système économique. On peut avoir un format libre, publié, mais que personne n'utilise, vouloir le recommander alors qu'il n'y a même pas d'outils pour formater des données et se trouver ainsi à côté du marché. À l'inverse, on peut avoir des formats propriétaires qui sont tellement répandus que l'on prend moins de risque en les utilisant qu'en utilisant certains formats libres. Par exemple, WAVE est un format propriétaire (IBM et Microsoft), mais tout le monde connaît ce format et les spécifications de WAVE sont malgré tout rendues publiques.

Sur le codage des langues :

Pour le codage des caractères, la norme Unicode (ISO 10146) est la seule à avoir une vocation universelle. Elle a mis fin à toute la pratique qui consistait, jusque dans les années 1990, à construire soi-même ses propres polices (codage ascii, iso latin, etc). Quand les besoins n'étaient pas couverts, on définissait les glyphes et il fallait donc disposer de la police pour voir correctement l'objet. Unicode permet de coder correctement tous les caractères existant dans les écritures. Ce langage est utilisé partout, comme pour naviguer sur le web. Et, quel que soit l'outil utilisé, un texte est normalement encodé soit en Unicode, soit dans un codage qui peut être traduit en Unicode. Les problèmes de translitération et de transcription ne sont pas pour autant réglés. Quelqu'un ayant recours à la translitération devra décrire et documenter le système qu'il utilise, sinon il ne sera pas possible de comprendre ce qu'il a voulu noter

[31] Service interministériel des archives de France

Médiation, pédagogie et valorisation des archives audiovisuelles en ethnomusicologie, Anne-Florence Borneuf

Les projets de valorisation des archives audiovisuelles et ethnomusicologiques de la Cité de la Musique courent sur une dizaine d'années. Il s'agira, avec de propos, de faire part des expériences passées et des nouvelles orientations de cette institution, sachant qu'on se situe actuellement dans une période transitoire, où les choses n'ont pas encore abouti.

Un rappel historique

L'ouverture de la médiathèque de la Cité de la Musique en 2005 est concomitante avec la mise sur pied d'un portail documentaire[32] qui puise ses ressources dans diverses archives de la Cité de la musique, notamment celles des concerts enregistrés dans les différentes salles (Cité de la musique et Pleyel). Concernant l'ethnomusicologie à proprement parler, il s'agit de matériaux particuliers, vu qu'il ne s'agit pas d'enregistrements de terrain, mais d'enregistrements de concert – point sur lequel on reviendra plus tard. Ce portail documentaire comporte une version Intranet accessible depuis la médiathèque de la Cité de la Musique et depuis d'autres médiathèques partenaires et une version Internet, disponible pour tous les internautes. La différence se situe dans la mise à disposition en intégralité de certains enregistrements sur l'intranet, ce qui est impossible sur l'internet.

Le portail documentaire propose à l'internaute de nombreux outils et documents pour l'exploration de la musique : ethnomusicologie, jazz, mais surtout, en ce qu'il s'agit du socle de la cité de la Musique, musiques classique et contemporaine. Tous les documents mis en ligne ne valorisent pas les archives ; on ne s'intéressera ici qu'à ceux qui valorisent les fonds d'archives. Tout d'abord, la rubrique « repères musicologiques » regroupe des textes généraux sur des sujets précis, dont certains passages surlignés peuvent renvoyer à des écoutes illustrant le propos, ces exemples sonores étant tous puisés dans les enregistrements de concerts de la Cité de la Musique. De façon tout à fait complémentaire, le portail propose également des guides d'écoute qui, à l'inverse des présentations thématiques, permettent de plonger directement dans une musique par le son, et qui sont accompagnés de commentaires. On entre véritablement dans une œuvre. Dans ces guides d'écoute, on cherche à restituer au mieux la manière dont cette musique a été conçue. Pour des raisons juridiques, ces guides ne sont accessibles que depuis l'intranet de la médiathèque et d'autres médiathèques partenaires, car ils font appel à la totalité de l'œuvre et de son interprétation.

Revenons plus spécialement sur les guides d'écoute. Le cheminement que je vais adopter à partir de maintenant va essayer de rendre compte de certains enjeux de l'évolution de la médiation au sein d'une institution comme la Cité de la musique et ceux de l'orientation vers des nouveaux publics. Ces deux questions sont traitées de front dans notre cas. Revenons aux guides d'écoute. Il s'agit d'une interface multimédia et interactive qui utilise le logiciel Metascore, commandé spécifiquement par la Cité de la Musique à Olivier Koechlin au moment de l'ouverture de la médiathèque en 2003-2004. Le logiciel comporte une interface éditeur et une interface lecteur, de sorte que l'équipe

[32] Consultable à l'adresse suivante : http://mediatheque.cite-musique.fr

de la Cité peut éditer en interne les guides d'écoute. Il vise à restituer une analyse musicale et à suggérer aux lecteurs une façon d'écouter la musique – à l'aide de commentaires, de schémas…

Pour aborder des sujets un peu plus ethnomusicologiques, prenons l'exemple de l'analyse d'un râga de l'Inde du nord, le *Râga jhinjhoti*. Lorsqu'on lance l'écoute avec le guide d'écoute, on peut voir un curseur qui avance, plusieurs blocs de textes ainsi qu'une représentation graphique de la totalité de la pièce qui permet de toujours savoir à quel moment on se situe. Il y a véritablement une interaction entre le son et ce qui est représenté. Un premier bloc de textes donne une présentation générale de la musique écoutée et des informations nécessaires à sa compréhension. Un deuxième bloc est synchronisé sur la musique et donne en direct des informations sur ce qui est en train de se passer à ce moment-là. Des annotations peuvent également apparaître sur la partition et aider le lecteur à comprendre certaines choses. On nous explique qu'au début la chanteuse chante des syllabes qui n'ont pas vraiment de sens, ce n'est pas véritablement un texte. On voit, également, qu'au début la chanteuse chante dans la partie grave de la tessiture. Si l'on navigue dans le râga, en allant un peu plus loin, on voit que petit à petit on monte dans la tessiture. Ce sont des choses qui sont montrées par le guide d'écoute. Le cycle rythmique est également doté d'un curseur. Les cycles rythmiques ne sont pas toujours simples à identifier et le curseur aide à comprendre où l'on se trouve dans le cycle rythmique. Le guide propose, de plus, la transcription du texte et sa traduction simultanée. Voilà comment fonctionne, globalement, un guide d'écoute : synchronisation, représentation musicale, curseur, lien, interaction, avec une navigation dans la pièce qui est possible à tout moment. Ce type d'objet pédagogique a été conçu à l'origine pour la musique classique. Il s'agissait de restituer une analyse musicale avec une dimension interactive et, dans certains cas, de permettre à l'auditeur de suivre des partitions compliquées. Évidemment, les répertoires de l'ethnomusicologie et du jazz ont dû s'adapter à cet outil, livré avec une sorte de rigidité, avec un pavé destiné à recevoir une image, un autre destiné à recevoir du texte, un autre encore destiné à recevoir du texte synchronisé à la musique, mais qui fonctionne relativement bien et qui peut être utilisée néanmoins.

Le public ciblé au moment de l'ouverture de la médiathèque était composé de mélomanes curieux. Par conséquent, les textes étaient écrits par des musicologues ou des ethnomusicologues spécialistes des musiques en question. Il s'agissait de textes parfois un peu exigeants, un peu techniques, mais qui ont ravi un bon nombre de personnes. Par ailleurs, ces guides étaient utilisés dans le cadre d'activités pédagogiques avec les adultes à la Cité de la Musique. Ils ont également été à disposition des enseignants par l'Éducation nationale, sur des portails dédiés. Par conséquent, beaucoup d'élèves de terminale ont préparé leur baccalauréat avec les guides d'écoute proposés par la Cité de la Musique. En 2010, il y a eu un tournant. La Cité de la Musique a voulu diffuser beaucoup plus massivement ses concerts qui étaient filmés, les diffuser en direct sur un portail dédié qui est citedelamusiquelive.tv[33]. Là, il est à tout moment possible de savoir quel sera le prochain concert diffusé et d'avoir accès aux archives des concerts. Dans la perspective de la mise en ligne de ce nouveau portail en 2010, juridiquement, des droits

[33] Pour consulter le site http://www.citedelamusiquelive.tv/

ont été négociés avec la SPEDIDAM[34] pour permettre la diffusion de concerts en intégralité. Le type de public visé via ce portail était beaucoup plus large que celui des précédents guides d'écoute. Il a donc fallu adapter les guides existants pour les rendre plus accessibles. Prenons l'exemple d'un guide d'écoute sur un calypso de Trinidad. La nouvelle version du guide (citedelamusiquelive.tv) fait une part bien plus importante à la vidéo et à l'image que la version médiathèque. Le texte est plus réduit, avec mention des interprètes et un texte beaucoup plus léger que dans l'autre version, qui vient expliquer le déroulement de la pièce musicale au fur et à mesure. L'interaction est moindre par rapport à la version précédente. Si certains guides ont été adaptés, d'autres ont été conçus directement pour le portail vidéo. C'est le cas du guide sur la *Symphonie pastorale* de Beethoven, dans lequel on a pris l'option de ne pas proposer de partitions de musique, mais de représenter graphiquement la musique. Une maîtresse de CE2, qui travaillait sur le thème de l'eau, a raconté qu'elle avait plusieurs séances avec ce guide, et que cela avait été très fructueux.

En ce moment, on est dans une phase de transition. On expérimente, on s'interroge, on teste. L'objectif, maintenant, est de véritablement ouvrir le portail à un public beaucoup plus large : enfants, familles, notamment via des partenariats avec des médiathèques en région, dont les adhérents pourront bénéficier d'un accès direct à l'intégralité du portail. Du coup, les personnels de ces médiathèques sont en demande de matériel pédagogique et de médiation beaucoup plus accessible, tant pour eux que pour le public. On s'oriente véritablement vers des produits grand public. Pour l'équipe de la médiathèque de la Cité de la Musique, cette étape est assez complexe, parce qu'elle se double d'une mutation technologique. La technologie « Director »[35] des anciens guides d'écoute conçus en 2003-2004 est obsolète, et cela pose des problèmes de diffusion de ces guides d'écoute à l'extérieur du portail de la Cité de la Musique (non-compatibilité avec certains navigateurs, etc.) Un système a donc été mis en place pour transférer les anciens guides d'écoute en HTML5.

La phase actuelle d'expérimentation et de maquettes porte surtout sur la musique classique. Un étudiant en psychologie cognitive, en stage à la médiathèque, a par exemple conçu une maquette de jeu interactif à destination des enfants, pour découvrir une œuvre, avec une mascotte guidant l'enfant pour lui apprendre à reconnaître un thème, à en définir le caractère, à lui faire comprendre que ce thème pouvait être joué par d'autres instruments, etc. Par ailleurs, une réflexion est menée par Delphine Anquetil pour casser la structure même du guide d'écoute et pour en faire quelque chose de plus attrayant pour les enfants. De même, et toujours à partir des outils HTML5, Delphine Anquetil a proposé, à l'occasion de l'exposition *Musique et Cinéma*, une nouvelle activité, toujours basée sur les archives sonores de la Cité de la Musique. Il s'agit d'un quizz destiné aux adultes, à la fois ludique et pédagogique, sur la thématique de musique et cinéma. Il s'agit d'explorer comment la musique est utilisée dans les films. Le quizz est structuré en trois étapes : 1) quelle musique pour quel effet ? 2) un hit parade des musiques de films ; 3) la musique classique au cinéma. Pour chaque quizz, on entend plusieurs extraits musicaux avec des questions associées, comme par exemple (parcours n°1) : « à quel genre de film pourrait se rapporter cette musique ? » La réponse, qui explique comment la musique est

[34] Société de Perception et de Distribution des Droits des Artistes-Interprètes

[35] Director est un logiciel de création d'applications vidéo (cd-rom, jeux, démos, simulations, tutoriels,...)

utilisée dans le film, permet de sortir un petit peu du jeu pour entrer dans un aspect plus pédagogique. Grâce au HTLM5, on peut désormais éclater les guides d'écoute, on peut entrer dans les codes et prendre seulement les éléments qui nous intéressent. Prenons un autre exemple : dans le troisième quizz « la musique classique au cinéma », il s'agit d'associer les morceaux de musique aux films. Ainsi, pour l'utilisation de la *Symphonie n°7* de Beethoven dans le film *Le discours d'un roi,* l'explication suivante est donnée : « la scène précédant celle du discours se déroule sans musique, ce qui augmente l'impression de tension. Le spectateur sait que ce que doit accomplir le roi est un défi. Le deuxième mouvement de la *Symphonie n°7*, commence dès ses premiers mots. Il accompagne particulièrement bien le discours, son grand crescendo est le reflux musical du gain progressif de confiance du monarque, et à la fin de l'explosion du sentiment patriotique. » On a pu intégrer à la réponse un mini guide d'écoute, mais pas d'image (impossible pour des questions de droits). Dans ce quizz, on trouve également la *Danse hongroise n°5* de Brahms utilisée pour *le Dictateur* de Charlot. Un guide d'écoute, intégré à la réponse, proposera une petite animation reprenant la scène où Charlie Chaplin rase son client en rythme sur cette musique. Ce guide d'écoute ne ressemble plus du tout aux guides d'écoute antérieurs.

Au terme de cette phase de réflexion, il s'agira de décider quelle(s) piste(s) retenir, tant du point de vue de la forme que du contenu. Le travail porte aujourd'hui surtout sur la musique classique, et il n'y a pas encore, pour l'ethnomusicologie, de piste nouvelle.

Discussion de la matinée

Christine Guillebaud :

Je voudrais réagir au plus proche de l'activité de chercheur en ethnomusicologie. Le projet *Sources de l'ethnomusicologie* pose, dès le départ, la question de savoir comment nous, chercheurs, allons déposer nos propres corpus, comment nous allons les utiliser dans la recherche et dans l'enseignement. Le projet soulève également la question de la diffusion de ces corpus.

Une idée essentielle à rappeler est que l'ethnomusicologie est une discipline qui, plus que d'autres, s'est développée en étroite imbrication avec l'évolution des technologies. La discipline a pu naître avec l'enregistrement. Pour les musiques n'ayant pas de traces écrites, c'est l'enregistrement lui-même qui permet d'avoir une distance et un regard analytiques.

Dans le cadre du projet, la conservation des enregistrements est donc un premier point à discuter. Comment conserver toutes les données, comment archiver, en sachant qu'en ethnomusicologie la notion de pièce n'existe pas. Depuis les années 1970, des enregistrements de performances permettent, par exemple, d'étudier le rapport entre musique et danse. L'ethnomusicologue fait, bien sûr, de l'annotation musicale, de l'analyse des pièces, mais il s'attache aussi à voir ce que cela crée en termes de mouvements, dans la durée de tout un rituel, par exemple. Les questions de cognition musicale sont également traitées avec l'utilisation notamment de cameras 3D. Mais ces sources importantes pour l'ethnomusicologie ne sont pas archivées, parce qu'il n'y a pas encore les outils, ni les moyens de trouver les formes d'indexation pour croiser les données de terrain, les supports sonores, audiovisuels…

Un autre aspect doit être souligné. Le passage de la recherche scientifique à l'aspect patrimonial a été mentionné. C'est un phénomène qui se retrouve dans tous les pays où l'on travaille. Lorsqu'on fait du terrain en Inde, un pays qui porte des politiques culturelles autour de la musique de manière très forte, depuis les années 1930, on ne collecte plus seulement des données pour les archiver ensuite au retour, on collecte aussi des travaux qui sont déjà des formes de patrimonialisation. Concrètement, mon travail sur les politiques culturelles en Inde m'amène à avoir des corpus composés, par exemple, des vidéos sur CD, des enregistrements sonores sur CD et des plateformes vidéo qui sont financées à la fois par le privé, comme la Ford Fondation, et par l'Unesco. Ce sont des données d'analyse qui demandent une prise de distance supplémentaire, parce qu'elles sont imbriquées dans une histoire, supposent des données historiques, et, par conséquent, sont difficiles à intégrer dans nos bases d'archives.

Pour ce qui concerne le multimédia, il faut souligner qu'il existe aussi en ethnomusicologie des guides d'écoute, développés par les chercheurs, pour un public de chercheurs et le grand public. On les appelle, au CREM, les « clefs d'écoute »[36]. Elles ont été portées, au départ, par Marc Chemillier qui était associé à l'équipe, et l'IRCAM, qui est bien équipé pour venir en appui sur les aspects cognitifs, sur la manière de présenter et de synchroniser le son et l'image. Là aussi, un problème de conservation se pose, parce

[36] Pour consulter ces clefs d'écoute : http://www.crem-cnrs.fr/realisations-multimedia

que ces clés d'écoute constituent des sources, au même titre que nos articles et nos publications. Anne-Florence Borneuf a évoqué la question du passage en HTML5. La majorité des ouvrages qui sortent en France sont accompagnés de DVD écrits en Director, qui, par conséquent, pourront difficilement être lus dans 5 ans. Comment peut-on archiver ces animations là ? Pourrait-il y avoir une réflexion sur ces questions au sein du projet sur les *Sources de l'ethnomusicologie* ?

La clef du lien entre les travaux des chercheurs et le grand public, c'est l'aspect multimodal de la diffusion de la musique, à la fois dans sa représentation graphique, mais aussi ses aspects multi sensoriels, par le mouvement, la danse, le visuel. Il y a matière à s'appuyer sur les outils développés dans la pédagogie pour toucher le grand public.

Un autre point encore est à souligner. Les travaux classiques, de type articles, ouvrages, sont de plus en plus souvent accompagnés de documents audio et audiovisuels. À chaque fois, la question se pose de savoir où stocker ces données. Les met-on sur le site du CREM ? Les dépose-t-on à la Société française d'ethnomusicologie, association qui ne peut pas garantir une conservation à moyen terme (environ 10 ans)? Pour le projet *Musique et politiques mémorielles*, en préparation, la question se pose de savoir comment intégrer, au moment de la publication, la logique d'un travail, avec tous les corpus récoltés sur le terrain, en intégrant autour d'un numéro de revue, par exemple, une série de documents sonores et audiovisuels qui seraient analysés et mis directement en lien avec ces corpus d'archives? Il ne s'agit pas uniquement de questions techniques, mais de choix de type éditorial. Et puis, il y a un aspect épistémologique. On réfléchit beaucoup, avec plusieurs collègues du CREM, à la manière dont mettre en lien les textes, imprimés ou diffusés en ligne sur revues.org[37] avec les archives sonores. Le choix actuel est de mettre des pastilles dans le texte, « exemple 1 », « exemple 2 », « exemple 3 », et de mettre en ligne, en parallèle, des séries d'exemples numérotés. D'autres modèles ont été explorés : la revue *Ateliers d'anthoropologie*[38], du LESC, incruste directement les documents sonores dans le texte, puisqu'il s'agit d'une revue en ligne. C'est notamment le cas avec le numéro sur la virtuosité, coordonné par les ethnomusicologues. La revue *Musimédiane*[39] utilise également le même procédé. À partir du moment où l'on a l'accès directement à la source audiovisuelle, cela implique des modes d'écriture scientifique un peu différents. Ce sont des questions qui devraient être aussi prises en compte au moment même de l'archivage des corpus. D'autres choix encore ont été faits, notamment pour les thèses qui ne sont pas encore publiées. Les doctorants, par exemple, qui ont déposé tous leur corpus, sonore et audiovisuel, renvoient directement par des liens à la base Telemeta et à leur collection et n'ont pas besoin de modéliser, à côté, un DVD dont la durée de vie n'excédera pas 5 ans. Tous ces enjeux de mise en lien entre analyse et support sont essentiels.

Le dernier point que je souhaiterais aborder est la question des langues, et les problèmes, essentiels, de typologie et de translitération. Sur les terrains de l'ethnomusicologie, on travaille majoritairement avec des gens qui ne sont pas de langue française. Penser au multilinguisme, ce n'est pas seulement réfléchir à la diffusion

[37] www.revues.org

[38] Consulter la revue en ligne : ateliers.revues.org

[39] Consulter la revue en ligne : www.musimediane.com

internationale de nos travaux, mais c'est aussi politique. Lorsqu'on travaille sur les questions de patrimoine musical, en Inde par exemple, on n'est plus dans la relation d'observateur à musicien observé, on travaille avec des gens qui se situent aussi dans un processus de mise en patrimoine de leur savoir. Ce processus est assuré, évidemment, par des chercheurs, par des médiateurs culturels, mais aussi par les musiciens eux-mêmes qui sont leurs propres producteurs de disques, leurs propres producteurs de vidéo CD. Ils travaillent d'ailleurs avec tout un réseau d'éditeurs locaux, souvent privés, qui diffusent largement dans le monde, parce qu'il y a des communautés qui ont émigré dans le monde entier. Pour l'Inde, ces communautés se trouvent principalement dans les pays du Golfe. Il s'agit d'un public de dizaines de milliers de personnes. Et il serait dommage que le travail que l'on peut faire sur ces matériaux soit seulement diffusé en français. Grâce à Telemeta[40], des chercheurs peuvent annoter sur place, avec des collaborateurs locaux, et ajouter des informations historiques, analytiques, le nom des pièces, la terminologie, etc. Il est donc important, d'avoir en termes ergonomiques, en termes de langues, des outils facilement utilisables par les collaborateurs, qui ne sont pas forcément des spécialistes des archives, ou des ingénieurs.

Pascal Cordereix

Sur les questions de conservation, rien de ne me paraît insurmontable, en l'état actuel et compte-tenu de ce que l'on sait faire. Quelques soit l'institution patrimoniale, des choses sont mises en place. Ces questions montrent donc l'importance de se rapprocher des institutions patrimoniales – et le labex est un cadre pour cela – et inversement, pour comprendre les besoins à la source. Que cela soit par le biais de migration ou par le biais d'émulation, dont on ne parle pas suffisamment, je pense qu'il y a des réponses à plusieurs questions posées ici. Ce que met en exergue le projet lui-même, c'est ce besoin de rencontre entre les chercheurs et les institutions ou les structures patrimoniales.

Dana Rappoport

Je suis chercheuse en ethnomusicologie, je travaille en Indonésie orientale, depuis 20 ans, et j'ai un très vaste fonds d'archive, que j'ai pu mettre en ligne grâce au projet Telemeta depuis trois ans. En dehors de mes archives sonores, j'ai une quantité de photos, de vidéos. Le premier problème rencontré est celui du stockage. Pour les archives sonores, nous avons Telemeta, mais que fait-on de nos archives vidéo, de nos archives photos ? Aujourd'hui, on bricole. On va dans un laboratoire, à Villejuif, qui paie d'autres laboratoires pour la numérisation. On conserve les données sur des disques durs, chez soi, dans les laboratoires et ailleurs. Pour le moment, il n'y a pas d'autres solutions. Faire un pas en avant vers la mise en commun de toutes les données paraît, en regard de cette situation, presque prématuré. Les normes de description que j'ai utilisées pour mes propres archives sur le net ne sont pas conformes. On décrit comme on peut, mais les instruments de musique qu'on a enregistrés ne sont pas bien normés. On aimerait, par exemple, que qu'un enregistrement puisse être renvoyé à une photo du musée du quai Branly. Si l'on décrit telle cithare tubulaire, on aimerait pouvoir cliquer sur un lien et que

[40] Plateforme Telemeta du CREM : http://archives.crem-cnrs.fr/

l'on voit ce que c'est. Mais pour l'instant, je n'en suis pas là. Ce sera peut-être le cas dans 5 ans, je ne sais pas. Ce travail nécessite beaucoup de main d'œuvre.

Un deuxième problème tient à la distinction entre archive brute et archive pensée. J'ai beaucoup d'archives brutes, sur lesquelles je n'ai pas réfléchi, et j'ai également beaucoup d'archives sur lesquelles j'ai construit des interprétations, que j'ai documentées très finement, en plusieurs langues. En écoutant les interventions de ce matin, je pense qu'il y a une confusion entre archives brutes et archives pensées. En tant que chercheuse, je distingue les archives brutes, les données que je ne modifie pas et que je place sur Telemeta, en les coupant simplement. Mais même dans ce cas, il y a des choix à faire, notamment les choix dans la segmentation. Où va-t-on couper dans un rituel de 7 jours 7 nuits ? Est-ce que je segmente au niveau du rituel, est ce que je segmente au niveau du chant ?

Je distingue ces archives brutes des corpus qui donnent lieu à des publications. En 2003, j'ai voulu publier mes corpus et mon interprétation. J'ai placé tout cela dans un DVD, dans lequel j'ai concilié scientifique et patrimonial. J'ai publié un gros coffret avec 60 heures de sons, chaque son étant relié aux photos, aux vidéos et aux textes. Cette publication est appelée à disparaître si elle n'est pas mis en ligne. Cela m'amène à demander si, dans l'avenir, il ne faudrait pas distinguer, pour les archives de l'ethnomusicologie, entre la mise à disposition des données et la mise à disposition des publications. Car le corpus pensé est différent de l'archive brute, parce que j'ordonne les musiques. D'un côté, je mets les musiques de funérailles, de l'autre coté, les musiques pour les transes, etc. Ne faudrait-il donc pas, par conséquent, intégrer au portail sur les sources de l'ethnologie les publications multimédias des chercheurs ?

Marie-Dominique Mouton

En tant que bibliothécaires, dernièrement, nous avons été associés à un certain nombre de colloques et de conférences. Il est question, maintenant, d'un nouveau secteur des archives qu'on appelle « les données de la recherche ». Il se distingue de l'archive stockée dans un tiroir et que l'on n'utilise pratiquement plus pour la recherche, sauf pour des questions patrimoniales (histoire, historique d'un terrain, etc.) Ces données de la recherche sont les données brutes, issues du terrain ; par exemple, les astronomes qui ont maintenant des télescopes avec des capteurs qui reçoivent des données numériques en quantités astronomiques (Big Data). Ce secteur devient gigantesque, il y a un énorme problème de gestion, d'autant qu'il s'agit de données vivantes, récentes. Il semblerait que l'on ait maintenant deux secteurs différenciés. Faut-il les traiter de la même façon ? Et, lorsqu'on emploie le terme « archive », parle-t-on bien de la même chose ?

Dana Rappoport

En quoi le chant agraire que je rapporte est-il une donnée ou une archive Il reste en fichier wave. Pour moi il devient une archive, quand il est mis sur Telemeta. Jusque-là, il reste une donnée sur mon disque dur.

Marie-Dominique Mouton

Une donnée devient archive à partir du moment où on lui donne une côte, une catégorie et que cette catégorie permet de retrouver l'objet.

Dana Rappoport

En tout cas, cette initiative du portail « sources de l'ethnomusicologie » est extrêmement favorable pour nous, ethnomusicologues. On rêve depuis longtemps de pouvoir déposer des données et de les relier entre elles. C'est un pas en avant. Telemeta est comme mon laboratoire, j'y circule dans mes fichiers. Sans cela, tous mes vieux enregistrements auraient été perdus. Telemeta constitue une véritable avancée dans la recherche. Mais, maintenant, il y a de grands pas à faire de mise en relation des corpus entre eux.

Joséphine Simonnot

Cet outil a vu le jour grâce à un travail pluridisciplinaire, et notamment grâce à la collaboration du Laboratoire d'acoustique musicale[41] qui a permis de trouver des outils adaptés à nos besoins de recherche en ethnomusicologie. C'est justement cette mise en concordance de compétences qui a permis de créer cet outil.

Jean Pierre Dalbera

Telemeta a été soutenu par le TGE Adonis via un projet auquel le Mucem[42] avait participé à l'époque, avec le Musée de l'Homme, et qui est le projet Anthroponet[43], qui posait exactement les problèmes soulevés à l'instant par Dana. Ces questions se posent dans tous les domaines des sciences humaines, à chaque fois qu'un chercheur va sur le terrain et rapporte des données qu'il faut conserver. On est encore moins bien outillé dans un musée national que dans un laboratoire, sauf peut-être au musée du quai Branly. En tout cas, au Mucem, on n'était pas équipés pour conserver les données de la recherche. Or c'est une mission de service public. Cela suppose de donner aux chercheurs des outils pour indexer les données, et il n'est pas toujours facile de convaincre une hiérarchie qu'il faut des informaticiens et des documentalistes derrière les chercheurs, parce que cela coûte cher ! Les chercheurs doivent, néanmoins, rester impliqués dans le processus.

Nicolas Prévôt

Je voudrais revenir sur la question de l'opposition archives/données de recherche. J'ai l'impression que ces catégories sont en train de changer et que cette opposition est amenée à disparaître complètement, et ce, pour deux raisons. D'une part, les possibilités techniques qu'on a, des outils comme Telemeta, rendent vivantes des archives qu'on considérait auparavant comme classées. On est de plus en plus amenés à retravailler ces archives dites classées et qui, par conséquent, ne le sont plus. Les outils technologiques permettent d'agir sur cette classification préalable, non pas en modifiant les données

---

41 Pour consulter le site du laboratoire : www.lam.jussieu.fr

42 Musée des civilisations de l'Europe et de la Méditerranée

43 Plus d'informations sur le projet à cette adresse : www.iri.centrepompidou.fr/projets/anthroponet

entrées par des chercheurs, mais en ajoutant de nouvelles couches destinées à de nouveaux chercheurs qui retravaillent les anciennes données et apportent de nouvelles strates d'information. La deuxième raison est une raison politique, quasi éthique. Dans le domaine de l'ethnologie, nous ne travaillons plus comme à l'époque coloniale. À cette époque, on récoltait des données que l'on archivait chez soi. Aujourd'hui, ainsi que l'a dit Christine Guillebaud, il y a un dialogue permanent avec des chercheurs des pays dans lesquels on travaille. Nous sommes amenés à ouvrir nos archives à ces personnes là. Par conséquent, elles ont aussi un droit de regard sur ce que l'on a fait de ces données, sur les points de vue que l'on a émis sur leur culture et elles peuvent elles-mêmes donner leur point de vue, c'est-à-dire ajouter des strates d'informations nouvelles. Elles peuvent aussi remettre en question les choses qui étaient fixées auparavant. Je crois qu'un outil comme Telemeta, pour prendre cet exemple, permet cela, puisque l'idée est de donner accès au plus grand nombre à des données sonores, et que chacun, pour autant qu'il ait un accès (un code), puisse, là où il se trouve, ajouter des informations avec une traçabilité. Toutes les personnes qui ont modifié ou ajouté des informations sont identifiables, et l'on peut les retrouver, non pour des raisons juridiques, mais pour des questions scientifiques.

Marie-Dominique Mouton

Entre les archives de gens disparus et leur traitement comme données vivantes dès lors qu'on les retravaille, il y a un *continuum*. Il faut avoir le souci de préserver les données aux différentes étapes, depuis leur création. Il n'est pas fondamental de passer du temps à s'interroger sur ce qui est archive et ce qui ne l'est pas. En complément de Telemeta, on pourrait mentionner une démarche proche, le portail ODSAS[44] de Laurent Dousset, qui ne comporte pas seulement des données sur la musique, mais aussi sur tout ce qui est films, images, textes. Ces projets montrent à quel point on a besoin, à l'heure actuelle, maintenant que les gens travaillent avec le numérique, de lieux pour stocker ; et l'on a également besoin de faire évoluer ce stockage pour continuer à travailler.

Michel Jacobson

Je pense qu'il faut distinguer les outils servant à décrire, à référencer, à enrichir, à gérer les données, et donc à gérer des informations, des outils servant à la conservation, et qui auront d'autres fonctionnalités (gestion de l'intégrité, de la visibilité, de l'authenticité). Quant à la question de savoir ce qui est archive ou ce qui ne l'est pas, les archivistes considèrent que toute donnée est une archive, et ils distinguent des âges différents : quand on la crée, quand l'on acquiert, quand on s'en sert, ainsi que l'âge intermédiaire où on la garde pour d'autres raisons que celle pour laquelle elle a été constituée, puis l'âge définitif. Pour les archivistes, tout est archive. Le modèle utilisé aujourd'hui est une norme ISO qui ne parle plus d'archive, mais d'information.

Françoise Dalex

Ce concept d' « information » plutôt que d'archive permet d'éviter que les corpus des sources de la recherche ne soient jetés, comme dans l'exemple que vous donniez plus

[44] Online digital sources and annotation system, consultable à l'adresse www.odsas.net

tôt, où il y a dix ans encore, des chercheurs ne prenaient pas conscience de l'importance de leurs travaux et étaient prêts à les jeter.

Prospectives des technologies innovantes, Françoise Dalex

Sur les contenus et leur accès

Pour entamer cette cession consacrée aux enjeux de mise à disposition, la présentation concerne d'emblée toutes les problématiques qui sont en gestation pour ce projet. Elles concernent les aspects de traitement des contenus et les choix de valorisation numériques, encadrés par les questions juridiques et éthiques de diffusion des collections. Le mot « réflexion » va régulièrement revenir dans cette présentation puisque nous sommes en début de projet, en phase de construction de projet.

Les sources de la réflexion sont multiples. Tout d'abord, les pratiques numériques des institutions culturelles connaissent une évolution majeure depuis plusieurs années. Les sites internet en particulier sont aujourd'hui moins des espaces de communication et deviennent davantage des sites de contenus, comme le nouveau site de contenus Centre Pompidou Virtuel. On peut aussi citer les mises en ligne de contenus des bases de données professionnelles. C'est le cas des catalogues de collections de la BnF, des 4 catalogues de collections du musée du quai Branly sur son portail documentaire. C'est aussi le cas de Mélémeta, la base de données du CREM.

En filigrane de cette présentation ne seront pas abordés les contraintes juridiques et éthiques, déjà identifiés, ni l'utilisation possible des données personnelles ou la question de l'archivage pérenne des contenus, qui relève davantage d'une politique nationale des tutelles que des initiatives des institutions, même d'un Labex.

En filigrane de cette présentation ne seront pas abordés les contraintes juridiques et éthiques[45], déjà identifiés, ni l'utilisation possible des données personnelles ou la question de l'archivage pérenne des contenus, qui relève davantage d'une politique nationale des tutelles que des initiatives des institutions, même d'un Labex.

La première mise à disposition concerne les contenus, c'est à dire les documents primaires numérisés et les données produites pour les décrire et les identifier. Ces deux éléments forment un patrimoine numérique complémentaire.

Les contenus et leurs accès reposent sur des chantiers de traitements documentaires communs et de pratiques numériques. L'accès aux contenus passe par l'utilisation d'un site internet, existant ou à développer. Une piste à explorer concerne les propositions de mutualisation des données numériques de *Gallica* et *Marque blanche* qui pourraient être une solution, à la fois pérenne et économique, et laisseraient la priorité au développement d'outils spécifiques au projet.
L'expertise en numérisation des documents audiovisuels des trois institutions a permis d'élaborer rapidement un cahier des charges des standards de formats de numérisation, de nommage des fichiers.

[45] Voir par exemple pour le musée du quai Branly, les conclusions du comité de mise en ligne et conditions de mise ne ligne des collections http://www.quaibranly.fr/fr/documentation/les-conditions-de-mise-en-ligne-des-collections.html
45

La pratique documentaire des trois institutions a imposé un chantier dans Rameau, pour enrichir et harmoniser les mots clés des instruments de musique et qui constitue un langage commun d'indexation. Michel Mingam présentera Rameau dans quelques instants. (RAMEAU (Répertoire d'autorité-matière encyclopédique et alphabétique unifié) est un langage d'indexation matière. Les impacts de ces travaux sont nombreux et constructifs. Concernant les documents primaires, la numérisation et la volonté de donner accès à une consultation de qualité ont orienté l'équipe vers le développement d'un player spécifique et le plus adapté à une consultation de tous les publics : par une utilisation simple et la mise à disposition des outils de consultation pointus comme des marqueurs.

Trois atouts caractérisent ces référentiels communs :

Il y a tout d'abord la possibilité d'afficher les référentiels, les mots-clefs communs dans une navigation et une ergonomie simplifiée, de les'utiliserpour alléger la navigation, l'ergonomie du futur site, par des nuages de tags, par exemple, qui guideront les internautes, ou pour mettre en place des rebonds qui suggèreront à l'internaute de poursuivre ou d'approfondir une recherche. On sort ainsi de la présentation quasi professionnelle des recherches et des bases de données sur internet, pour aller piocher dans nos outils professionnels statiques, ce qui permettra à l'internaute de naviguer facilement et de comprendre les contenus de notre site.

Deuxième atout, la possibilité de fédérer une recherche sur les différentes collections renseignées, par ces mots-clefs communs, jusqu'aux instruments de musique du musée du quai Branly qui sont, eux aussi, indexés avec les mêmes référentiels, avec le même thésaurus.

Et puis, enfin, il y a la possibilité de constituer des sets de métadonnées. Les métadonnées sont les outils qui permettent l'interopérabilité. Elles enrichissent les possibilités de recherches et améliorent le référencement du site. Elles sont mises à disposition sur internet pour communiquer avec des sites extérieurs, dans un système numérique, où l'on sait aujourd'hui que la dissémination, la mise à disposition des contenus est indispensable à leur valorisation et permet de les rendre accessibles au plus grand nombre, au-delà de nos propres sites. Dans le cadre de la constitution de ces métadonnées, les questions posées par le projet « Sources de l'ethnomusicologie » sont pas les formats d'échange, comme l'OAI/PMH ou le RDF. Elles portent aussi sur les modes opératoires pour garantir la pérennité des partages et le moissonnage, sur la sélection des partenaires, sur les outils à mobiliser pour atteindre ces données. On peut faire le choix de disséminer les informations vers des partenaires choisis (autres institutions publiques, par exemple) ou bien, au contraire, faire le choix d'ouvrir complètement nos données à tous les partenaires y compris les plateformes de contenus de type Youtube, en utilisant leurs outils de recherche, style API, pour être aspiré automatiquement vers leurs sites. Dans ce cas de figure, on ne fait pas le travail de

sélection pour aller poser des contenus sur leur site, mais on les laisser venir chercher nos contenus.

La dernière question, celle des contenus et de leur accès, concerne la mobilité des applications. Doit-on prévoir des applications de mobilité, des développements vers des outils nomades pour diffuser des contenus qui sont avant tout scientifiques et appellent des outils de lecture spécifiques qui risquent de ne pas passer sur ces applications ? Et, si la réponse est oui, avec quels partenaires les développer ?

La valorisation, les modes d'éditorialisation et un renouvellement de la médiation

La valorisation et l'accès aux contenus passent par une question complémentaire qui est celle de par leur enrichissement sur internet, encadré par une politique éditoriale, et par la production d'outils de médiation. Elle implique donc, de nouveaux acteurs et des outils de diffusion à définir.

Elle apparait comme une étape parallèle ou une seconde étape à la mise en ligne pour donner davantage à voir et à comprendre.

Le premier but de cette éditorialisation, de cette médiation est de donner à voir davantage, à comprendre davantage que ce que donne à voir une diffusion « primaire » des contenus et de leurs outils de renseignement. Pour optimiser la compréhension des données disponibles, on peut faire appel à différents dispositifs, allant de la transcription des documents – Christine Guillebaud a déjà soulevé cette question ce matin – à la traduction des contenus, dans un contexte de diffusion internationale qu'est le web, avec une identification des langues, mais surtout, avec une identification de la profondeur de la traduction. Est-ce qu'on se contente de traduire la première page et les éléments de navigation ? Est-ce qu'on fait un site complètement multilingue, y compris dans les référentiels de recherche, etc. ? Et puis, pour optimiser la compréhension des données disponibles, il semble intéressant de récupérer nos données d'indexation, nos mots-clefs, de les détourner et d'en faire des outils de compréhension pour tous les publics, par une modélisation, par une visualisation différente. Nous travaillons sur des corpus qui ont une forte dimension géographique et historique, et on ne peut pas s'empêcher de penser à des cartes qui seraient être réalisées à partir de nos listes de mots-clefs des nom de lieux, et une ou des frises du temps faites à partir des données de datation, que ce soit pour de grandes périodes ou des dates précises. Pour le projet *Sources de l'ethnomusicologie*, il est possible d'envisager le développement de dispositifs qui rendent les données accessibles au plus grand nombre : la transcription des documents audiovisuels, la traduction des contenus dans un contexte de diffusion internationale avec une identification des langues à traduire et surtout de la profondeur de la traduction : la première page et les éléments de navigation ou l'ensemble des contenus ? A introduire également, le « détournement » des outils d'indexation, réutilisés pour s'afficher sous des formes intuitives : Un thésaurus des toponymes devient support de géolocalisation et permet l'affichage de cartes, un index de dates se fait frise chronologique.

Deuxième élément d'éditorialisation : produire des contenus et des modes d'affichage complémentaires. Bien entendu, on peut envisager un enrichissement éditorial fait par des spécialistes – ethnologues, historiens, ethnomusicologues – pour

ajouter un appareil scientifique à nos documents. On pourrait aussi produire des programmes scénarisés sous la forme de visites virtuelles qui seraient plus axés vers le grand public, qui permettraient de valoriser une sélection de documents, et d'en faire des productions – exportables sans doute - pour un plus grand public.

Enfin, il est aussi possible d'associer des contributions au partage des contenus, qui peuvent à nouveau être des contributions de scientifiques. On peut également envisager, pour accompagner notre dissémination des contenus, que des spécialistes du projet collaborent à l'enrichissement de sites de nouveaux partenaires, tels que des encyclopédies collaboratives, qui seront, elles, tentées « d'aspirer » nos contenus.

La question des publics

Cette politique d'éditorialisation pose toute la question des publics, des internautes eux-mêmes, qu'ils soient spécialistes ou grand public. On a vu avec la Cité de la Musique, il y a une dizaine d'années, qu'on parlait des publics curieux et amateurs. Aujourd'hui, on sait qu'il faut aller vers de nouveaux publics. Nous travaillons sur des corpus du monde entier et il y a un devoir de restitution, de partage avec les civilisations et les peuples d'origine de ces archives ethnomusicologiques. Donc la réflexion porte également sur les dispositifs de réappropriation des contenus.

L'enrichissement des contenus peut certes venir des scientifiques qui seraient identifiés par le projet, ou, *a posteriori*, identifiés au fil de leurs contributions. Mais il peut également venir des publics, si l'on met à disposition des outils qui permettant d'annoter, tagguer les documents, d'enrichir les référentiels, les mots-clefs ou de déposer des documents. On peut envisager que, dans le cadre de ce projet d'ethnomusicologie, des ethnomusicologues ou des artistes, s'inspirent de ce qui est en ligne pour enrichir les contenussur un espace de sauvegarde et de partage des créations musicales liées aux contenus officiels.

À cet enrichissement éditorial vient s'ajouter une réutilisation gratuite et de qualité des contenus. Aujourd'hui, tous les internautes ont l'habitude de télécharger et d'imprimer. La question est de savoir si l'on donne des possibilités de téléchargement et d'impression en haute définition, par exemple. On pense au développement d'un « player » *ad hoc*, doté de toute la boîte à outils dont a besoin un musicologue pour travailler. Ne pourrait-on pas envisager que ce player soit exportable, tout comme le feuilletoir de Gallica qui peut être exporté sur les blogs et d'autres sites? Enfin, la possibilité pour les internautes de personnaliser les contenus qu'ils ont identifié comme des références, avec l'usage des liens pérennes, c'est-à-dire la possibilité de sélectionner un contenu qui garde le même lien, trois mois plus tard, six mois, deux ans plus tard, huit ans plus tard et qui sera toujours accessible pour les internautes, ou avec la possibilité de citer des documents en installant des outils comme des logiciels de bibliographie de type Zotero ou Endnote.

Enfin une dissémination sur les réseaux sociaux et les plateformes de partage dédiées aux documents sonores et audiovisuels semble aujourd'hui indispensable, comme YouTube

ou Dailymotion. Ils opèrent à la fois comme plateforme de sauvegarde et espace de partage des documents. Il reste à savoir lesquels seront sélectionnés. Ils sont indispensables, car c'est la première porte des internautes pour accéder à un patrimoine qu'ils ne connaissent pas.

Une question reste en suspens concernant ce dispositif de réappropriation des contenus par des internautes. Quelle place donner à ces productions dans le projet ? Elles semblent légitimes aujourd'hui, mais n'ont pas de statut bien défini qui permettrait de les mesurer quantitativement ou qualitativement, de les valider, de les conserver voire de les valoriser.

En guise de conclusion, nous pourrions évoquer deux pistes de réflexion. La première est de se demander si le projet doit envisager d'importer des contenus de sites similaires. On peut, par exemple, citer, au sein du labex *Les passés dans le présent*, le projet de portail « naissance de l'ethnologie française », qui pourrait donner lieu à des enrichissements et des partages de contenus réciproques. La deuxième porte sur les fonctionnalités de la diffusion. Nous avons à trouver notre périmètre pour les choix éditoriaux, les choix de valorisation, les modèles de données (web de données, plateformes). Peut-être pourrions-nous, dans les mois qui viennent, développer un prototype autour d'une collection significative, emblématique des fonds concernés, ou autour d'outils innovants qu'on est sûr de vouloir développer et qui nous serviraient à lancer le développement.

Question sur l'importance du respect de normes éthiques et juridiques au moment de la diffusion

Il faut toujours tenir compte des contraintes juridiques et éthiques. Nos trois institutions ont l'habitude d'encadrer tout leur travail de valorisation par une réflexion en amont sur les questions juridiques et éthiques pour voir quelles sont les possibilités en termes de diffusion et de mise à disposition des contenus. Il n'empêche que cela pose la question de la qualité de la mise à disposition pour une réutilisation.

Question sur la volonté des porteurs du projet de s'engager dans une politique de réutilisation des données de type Open Data, par exemple

La question est ouverte. Cela paraît aller dans la logique de ce qui a été exposé. Néanmoins cela dépend des politiques mises en place par les institutions.

Question sur la possibilité laissée aux chercheurs ou artistes de déposer des contenus sur la plateforme numérique

C'est une piste. Il est possible que, dans un processus d'interactivité, des chercheurs ou des artistes reviennent vers nous, inspirés par ce qui aura été mis en ligne ou par des contenus complémentaires. Par exemple, des archives des années 1930 vont être prochainement mises en ligne, il se peut qu'un musicologue ait fait un travail sur le même terrain et souhaite le faire partager, ou bien qu'un artiste s'inspire de ce qu'il entend d'une archive sonore pour faire quelque chose de contemporain et qu'il le propose pour le portail.

Sur la question des publics

La question des publics est centrale, comme on l'a vu avec l'intervention de la Cité de la Musique. On voit bien que les usages déterminent le mode de présentation des collections. L'historique des documents de la Cité de la Musique sur le portail est très révélateur de ce point de vue. Il permet de voir comment, en partant de quelque chose de très pédagogique, on aboutit à quelque chose de plus modulable, que le public peut s'approprier. Avec notre projet, nous sommes dans le même questionnement : quels sont les usages des publics et comment peut-on y apporter des réponses sans fermer les présentations ?

Les archives sonores de la mission en Basse-Bretagne du MNATP en 1939[46] : l'exemple d'une valorisation innovante

Jean-Pierre Dalbera

J'étais au MuCEM, à l'époque où a été lancée l'opération des archives sonores de la mission en Basse-Bretagne avec Marie-Barbara Le Gonidec. Nous avions à gérer dans les musées, au musée de l'Homme et au MuCEM, des fonds auxquels s'ajoutait la collection Europe du Musée de l'Homme, qui n'était pas encore au musée du quai Branly. Que faire de tout ce savoir accumulé ? Comment utiliser au mieux les données de la recherche, les stocker, les conserver et les valoriser, et réaliser des expositions, ce qui est le rôle premier d'un musée ? Comment le numérique pouvait-il améliorer la circulation des données tout en satisfaisant les chercheurs, notamment en termes de conservation des données et de valorisation de leurs productions par les médiateurs du musée, les documentalistes, etc. Il fallait pouvoir retrouver aisément les données, et, par conséquent, les indexer.

Le projet Anthroponet s'est posé ces questions. Les formats ont beaucoup été débattus, la question du web sémantique (à ses débuts) a été évoquée, tout comme celle de la documentation RDF. Il y a plus de 20 ans, ce n'était pas simple à organiser parce que les pays du sud de la Méditerranée n'en étaient qu'au début de l'Internet, tandis que les Canadiens étaient déjà très avancés, bien plus que les Français. On avait conscience des besoins d'indexation sous des formats interopérables, transposables, mais il fallait avoir les moyens de le faire.

Le projet Anthroponet a également pensé non seulmeent le problème des archives scientifiques, mais celui des archives administratives au sens large – par exemple, dans un musée, toutes les archives des expositions.

Audrey Viault

La plateforme mission Basse-Bretagne 1939 est en EAD, un standard de description archivistique normalisé, selon une méthode qu'on appelle DTD ISAD(G), très utilisée dans le monde archivistique et dans les bibliothèques. Cet instrument est un bon exemple de valorisation réussie, tant pour le traitement scientifique fait par Marie-Barbara Le Gonidec que pour le choix de traitement documentaire offrant des possiblités de recherche variées. Surtout, et c'est ce qui nous intéresse dans le cadre de cette journée, cette base de données est aisément accessible en ligne.

La mission, qui s'est déroulée en 1939, sur quelques mois, en Bretagne, était menée par Claudie Marcel-Dubois (1913-1989) assistée par l'abbé François Falc'hun (1909-1991) et Jeannine Auboyer (1912-1990). Il s'agissait d'une enquête ethnomusicologique, qui s'est également intéressée aux danses, aux aspects linguistiques, aux scènes de la vie courante, au cadre de vie des Bretons d'avant-guerre. La grande caractéristique de ce fonds est qu'il inclut les archives du travail de préparation (repérages, préparation des questionnaires, etc). À cela s'ajoute la collecte en elle-même ; des enregistrements, des films, des photographies, des analyses, des carnets de route, des

[46] Consulter le site de la mission Basse-Bretagne : http://bassebretagne-mnatp1939.com

transcriptions, des notations, etc. De retour au musée national des Arts et Traditions Populaires de Paris (MNATP), tout cela a été inventorié, identifié...S'est ensuivi, également, un travail de « publicité », notamment avec la presse, qui constitue un troisième ensemble d'archives papier. Ce fonds nécessitait donc d'être décrit dans sa globalité ; les enregistrements, surtout, les films prenant tout leur sens avec la documentation réalisée simultanément ou après la collecte. De plus, beaucoup d'enregistrements étant très abimés, voire presque inaudibles, il était important de les mettre en relation avec les documents d'analyse produits par la mission.

Réalisé sur le plan scientifique par Maria-Barbara Le Gonidec, et édité sous la forme d'un livre exposant la méthode scientifique d'analyse des documents et accompagné d'un DVD. Or, non seulement le DVD n'est pas un support pérenne, mais il impose un parcours déterminé. De plus, celui-ci ne proposait ni images, ni photos. Marie-Barbara Le Gonidec a donc recherché un autre mode d'expression documentaire, qui combinerait une description précise et une structuration des contenus, leur indexation normée, une possibilité d'interroger les contenus à partir de parcours adaptés aux différents publics, le tout accessible en ligne. L'outil de l'EAD a été choisi. L'EAD est basé sur un langage de balisage très connu, l'XML, lui-même issu du langage SGML qui a donné le HTML et XML. C'est un langage de balisage extensible. On structure un texte de données en l'encadrant dans des balises qui ont un sens précis, normalisé par une DTD, un mode d'emploi (l'EAC[47], par exemple, est un dictionnaire de balises). Il s'agit d'un emboîtement de balises, avec une balise principale appelée « élément racine » qui donne un sens global à ce qui va être décrit et qui réunit des informations de même nature. Il y aura, par conséquent, des balises sur le matériel, sur la localisation, sur les titres, etc. Par cet emboîtement de balises, on crée un phénomène de hiérarchie. Le XML permet de décrire le contenu mais pas la mise en forme et la structure, contrairement au HTML. L'avantage du format EAD est qu'il transporte les données et la structure. L'XML EAD permet de transporter données et structure et permet aussi de relier les instruments de recherche EAD à d'autres documents en formats XML. On peut ainsi conjuguer beaucoup de ressources.

Revenons au site mission Basse-Bretagne 1939. Les parcours de recherche sont transposés sur le site via les onglets. Le deuxième onglet « consulter le fonds » donne accès à la description organique du fonds, à partir de sa structure (avant, pendant, après la mission). En cliquant sur l'une des 3 grandes sous-parties de l'arborescence, on trouve des dossiers, divisés eux-mêmes en sous-dossiers, dans lesquels les documents sont décrits de façon typologique (par exemple dans la partie « pendant la mission » les documents sont classés entre « journaux de route », « collectage », « cahiers de terrain »…). Les sous-dossiers proposent des descripteurs renvoyant aux niveaux supérieurs (titres, cotes, contenus), et des notions qui, par un clic, renvoient soit à des liens, soit à des indexations. Ce sont là deux aspects principaux de la valorisation d'un outil de recherche, puisque, à partir d'une description organique, il est possible d'établir des liens entre différents composants, différentes pièces d'un même dossier, sans pour autant casser la composition organique du fonds et la temporalité de la collecte. Vous pouvez naviguer d'un document à l'autre grâce à ces liens, et enrichir votre recherche grâce à ces indexations. Ainsi, par exemple, la documentation du journal de la mission

[47] Plus d'informations sur l'EAC, sur la page Wikipédia dédiée

tenue par Jeannine Auboyer est accessible dans son intégralité, elle peut être imprimée, sélectionnée dans un panier document, etc. Le deuxième aspect de valorisation de ce fonds est l'indexation. L'indexation demande un certain recul sur le fonds, une bonne analyse documentaire, et implique de réfléchir au type d'indexation que l'on souhaite. Cette indexation, avec Pléade[48], se concrétise par des index thématiques (onglet à gauche de la page et mots en surbrillance). En cliquant sur un terme d'indexation, on accède à tous les dossiers qui s'y rapportent.

En revenant à la page d'accueil, un troisième onglet « faire une recherche permet de faire une recherche par mots ou une recherche transversale au moyen d'un formulaire, qui peut également être une recherche thématique sur tel ou tel aspect de l'enquête ethnomusicologique. Ce sont des outils de recherche destinés à une communauté de connaisseurs.

Mais l'intérêt du site est qu'il propose également une autre approche de la mission, pour d'autres personnes – amateurs de folklore, passionnés de vieilles photos, etc. On y accède par le quatrième onglet « visite guidée », plus visuel et de structuration moins complexe. Cette visite guidée correspond au contenu du DVD, avec un aspect plus ludique, plus abordable, avec même un mode d'emploi qui indique comment naviguer dans cette visite guidée.

L'intérêt de ces documents est que, en étant bien pensés, bien classés au préalable, et avec un format normalisé, standardisé avec un langage XML, ils sont exportables avec un protocole que propose Pleade, OAI, vers des portails documentaires externes, par exemple, et combinables avec d'autres instruments de recherche. Faire des instruments de recherche normalisés, finement décrits, exportables, indexés, interrogeables et publiables, permet d'envisager aussi un mode de description qui serait valable pour d'autres enquêtes de nature assez différente. Dans l'objectif de ce labex, cela pourrait ouvrir une voie pour que des fonds de nature similaire communiquent entre eux, aient un langage commun. On pourrait, dès lors piocher dans ces fonds pour créer un nouvel instrument, et ce, dans le contexte du web de données qui fonctionne sur l'interopérabilité, et l'interconnexion de fonds. Mais cela suppose une norme, cela suppose que l'on parle tous de la même façon, qu'on ait des outils référentiels communs. Ceci pose, bien sûr, la question du web sémantique et la question de l'Open Source. Dans le cadre du projet labex, pour les fonds ethnomusicologique, il y a un besoin de descripteurs musicaux. Cependant, si le référentiel RAMEAU international, utilisé en France par de nombreuses institutions, est exhaustif sur les instruments occidentaux, autant il est lacunaire pour les instruments non occidentaux. D'où l'idée de créer un référentiel sur les instruments non occidentaux dans le référentiel RAMEAU et de mettre en commun les connaissances des instruments. Un thesaurus a ainsi été créé au musée du quai Branly sur les instruments de musique, la BnF apportant une certaine connaissance de l'utilisation de RAMEAU.

Pour revenir à la réalisation sur les archives de la mission en Basse-Bretagne, cette réalisation, dans la catégorie des archives ethnomusicologiques, est une très belle réussite. Il s'agit d'un document scientifique, traité de façon scientifique qui passe dans le domaine archivistique, clairement classé, hiérarchisé, indexé, et c'est ainsi qu'un corpus qui était strictement scientifique devient quelque chose de plus patrimonial, destiné à un

[48] Consulter le site de Pleade : http://pleade.com/

plus grand public. On sort d'une emprise strictement scientifique, quelque chose qui mérite, aussi, d'avoir un statut plus connu, plus diffusé avec des approches différentes.

Question sur les besoin de financement pour réaliser un tel projet

Il n'y a pas eu suffisamment d'argent pour aller au-delà du travail de Marie-Barbara le Gonidec et créer une plateforme qui soit réellement grand public. Aussi, il est essentiel, dans le cadre du labex, que les chercheurs aillent chercher des outils, en s'adressant à des professionnels des archives et du traitement de ressources, que ce soit dans de grandes institutions ou de petites sociétés. Par exemple, sur les pages de la BnF, il y a une mise à jour régulière des actualités concernant les formats, le web sémantique… Il est donc important que les chercheurs visent aussi ce but d'élargir leur public et d'aller chercher des collaborations professionnelles afin de mettre en valeur et d'exposer leur travail.

Question sur l'hébergement des contenus

Les contenus sont hébergés par AJLSM. Pour l'instant, le site est un peu en électron libre. Le MUCEM ne souhaite plus l'héberger, il nous faut trouver un autre endroit. Le Ministère de la Culture ne le prendra pas, parce qu'il faut du matériel dédié, c'est assez complexe. En revanche, le site a l'avantage d'être en XML.

Question sur la finesse de l'instrument

Avec ces instruments, quand on a beaucoup de documents, on peut choisir le niveau de description. On peut, dans un premier temps, signaler les fonds et les décrire plus brièvement, ce qui n'empêche pas de donner une information de qualité. Cela permet de signaler, de communiquer sur ces fonds et pour cela, ce genre d'instrument est très efficace. Parfois, le fond ne peut pas être traité intégralement, faute de temps, par exemple, mais rien que l'enquête EAD permet de signaler de façon très complète un fonds et de dire aux chercheurs et au grand public, où est ce fonds, comment le consulter, à qui s'adresser, et de pouvoir les renvoyer vers des sites en relation. Cela permet, déjà, de communiquer et valoriser ces fonds, même si la profondeur des descriptions n'est pas infinie, c'est déjà un travail qui rentre dans le thème de cette journée.

Le langage d'indexation RAMEAU, Michel Mingam, Natalie Bourdeau

Michel Mingam, Centre national RAMEAU-BnF

Avant d'en venir au langage RAMEAU lui-même, quelques mots sur le contexte. RAMEAU est né en 1980, au moment où les bibliothèques s'informatisaient. On avait la possibilité de constituer des catalogues collectifs, de mettre en commun des données bibliographiques et cela supposait d'avoir un langage commun y compris pour l'indexation-sujet. Plutôt que de créer un répertoire *ex nihilo*, on a choisi d'adapter la liste d'autorités établie par la Bibliothèque de l'Université Laval, au Québec, elle-même dérivée des Library of Congress Subject Headings.

Aujourd'hui, après 33 ans, RAMEAU est le langage majoritairement utilisé dans les centres documentaires en France, Belgique, Suisse, au Maghreb, en Afrique sub-saharienne francophone, au Liban et par une nébuleuse d'utilisateurs qu'il est impossible de chiffrer. Près 280 établissements autorisés participent concrètement à l'enrichissement du langage RAMEAU par leurs propositions terminologiques : les départements de la BnF, tout le réseau SUDOC, des bibliothèques publiques, municipales, des musées, des établissements privés. Ce réseau d'établissements est régi par une convention interministérielle, renouvelée en 2011, qui associe la BnF, la BES (bibliothèque de l'enseignement supérieur), le ministère de la Culture et le ministère de l'Enseignement supérieur. Il revient à la BnF d'animer le réseau, de gérer le langage et de fournir les outils. Le site internet rameau.bnf permet l'accès au vocabulaire ainsi qu'à un module de propositions terminologiques en ligne que l'on traite et auxquelles on répond également en ligne. RAMEAU est donc une structure assez lourde, mais dynamique et en expansion.

RAMEAU représente aujourd'hui environ 170 000 descripteurs, mots-clefs, ou vedettes, dont un peu plus de 100 000 noms communs, 60 000 noms géographiques et d'autres types que nous n'avons pas le temps d'évoquer. Il s'agit d'un langage encyclopédique, qui porte sur tous les domaines. Il n'est pas constitué *a priori* comme un thesaurus, et il s'enrichit au fur et à mesure des besoins et des propositions des utilisateurs. En termes de chiffres, depuis le lancement du fichier national, on dénombre plus de 17 000 propositions faites par des établissements de toutes sortes. Pour autant, évidemment, il n'est pas exhaustif et reflète la documentation existante. En effet, à chaque fois qu'un document nécessite un descripteur qui n'existe pas, la notion est créée dans le langage Rameau.

Le fait qu'il s'agisse d'un langage encyclopédique en développement constant implique nécessairement des problèmes de mise en cohérence, notamment entre les besoins parfois contradictoires des établissements (il faut à la fois servir la bibliothèque municipale de quartier et la bibliothèque extrêmement spécialisée).

Il faut imaginer RAMEAU comment une sorte de dégradé depuis les termes les plus généraux, qui sont aussi les plus communs (autrement dit, *le Petit Larousse*), jusqu'aux termes les plus spécialisés. Depuis 2007, toute une équipe reprend systématiquement le langage RAMEAU, domaine par domaine, afin de l'améliorer. À ce jour, 30 000 termes ont été revus et des domaines entiers retravaillés. Nous progressons dans l'amélioration de la qualité des données et de la cohérence du système. RAMEAU était initialement conçu pour les catalogues de bibliothèques, mais, si l'on admet que le

web de données est un lieu dans lequel les recherches se font par concepts, il faut des données structurées comme le RAMEAU, qui trouve une nouvelle carrière dans ce contexte. Cela implique donc que les données soient de qualité, car un système technique performant sans données de qualité ne tient pas. Et, inversement, de bonnes données sans un système technique capable des décrypter n'ont pas d'intérêt.

Démonstration du site RAMEAU

Le site est un outil de dialogue avec le réseau des établissements qui participent à sa mise à jour. Il propose des informations sur RAMEAU, des aides à la formation et à l'autoformation, et donne accès aux outils pour utiliser RAMEAU. Les « autorités RAMEAU », sont le vocabulaire. Le web sémantique ouvre des nouvelles possibilités, et d'autant plus que RAMEAU est accessible en format bibliothèque numérique, comme dans cette application, mais aussi en format RDF. RAMEAU ne se contente pas de rassembler des mots-clefs (ou descripteurs ou vedettes), par exemple « instruments de musique », mais il les accompagne de tout un ensemble de données. Comme il s'agit d'un langage contrôlé, un terme est accompagné de variantes, du langage courant et du langage savant, qui permettent l'utilisation de RAMEAU à plusieurs niveaux. Le premier niveau de RAMEAU permet des recherches par mots clefs.

Le deuxième niveau est sémantique. Il s'agit de notions reliées entre elles, comme dans un thésaurus, avec des termes génériques qui permettent d'élargir la recherche, des termes associés et des termes spécifiques qui sont des termes plus spécialisés (par exemple, les types d'instruments de musique et dans chacun des types, les instruments eux-mêmes, etc.). Ces catégories peuvent être enrichies par les contributions. Les sources consultées sont indiquées. Pour les termes très courants, cela peut-être le Petit Larousse par exemple, mais pour des termes plus spécialisées, les sources sont plus variées

On a, également, l'équivalent anglais, ici dans le répertoire de la Bibliothèque du Congrès, ce qui permet d'organiser les programmes, au niveau européen, de recherches multilingues. C'est-à-dire qu'on pourra interroger à partir du descripteur français des fonds qui ont été décrits avec des descripteurs anglais et allemands.

Donc, nous avons un niveau terminologique, un langage contrôlé, un niveau sémantique, avec une sorte de thesaurus dans lequel on peut naviguer, mais RAMEAUest aussi un langage pré-coordonné. On crée des vedettes matières avec des subdivisions, qui peuvent être complétées au moyen de catégories (géographie, dates, etc). Cela permet un affichage sous forme d'index, ce qui donne une lisibilité aux résultats. La coordination se fait au moment de la description, et non au moment de l'interrogation.

Évidemment, si l'on cherche un terme générique, il est possible d'affiner très vite la recherche. Si l'on veut associer musique et langage, on passe sur « langage » et l'on arrive très vite dans un autre domaine, du côté de la linguistique. RAMEAU étant encyclopédique, on peut passer d'un domaine à l'autre, ils sont tous inter-reliés.

Sur les évolutions et adaptations de RAMEAU

Quand il y a des évolutions au sein d'une discipline ou de nouveaux concepts, on crée de nouvelles vedettes matières. Le langage courant, le langage commun sera toujours privilégié. Par exemple, il y a eu une requête sur les Lapons. Eux se désignent eux-

mêmes comme les Samis et en Suède, le terme Lapon est dévalorisant. Mais en France, ça ne l'est pas. Et tant que ce terme sera dans le Petit Larousse, on laissera le terme Lapon sur RAMEAU, car tout le monde sait à quoi renvoie le terme Lapon, à la différence du terme Samis que seuls les spécialistes connaissent. Il ne faut pas oublier que Rameau est un langage universel. Donc, plus on est dans la spécialité, plus on peut arriver dans les nuances. Si c'était un peuple absolument inconnu en France, cela ne poserait pas de problème de changer la désignation. Pour ajouter des concepts, de nouvelles autorités, il suffit de faire des propositions, appuyées sur des documents, évidemment. Pour le choix de la vedette, il y a des règles qui veulent que lorsqu'il s'agit de notions courantes, on essaye de prendre le terme courant. RAMEAU est le plus scientifique possible, mais il doit rester pratique.

Du catalogue au web de données : l'exemple de data.bnf.fr

Agnès Simon

Le projet Data.BnF.fr, développé par le département de l'information bibliographique et numérique de la BnF, a fait l'objet d'une première mise en ligne en juillet 2011. C'est donc un projet récent, avec une dimension recherche et développement. Il montre que des référentiels comme RAMEAU, par exemple, ont une vraie valeur sur le web, car ce sont des données fiables, structurées, contrôlées, avec des identifiants stables et pérennes. L'intérêt de les intégrer est important. Dans l'univers concurrentiel du web, la BnF a plusieurs atouts. Elle propose des millions de documents (12 millions de notices bibliographiques qui viennent du dépôt légal de l'édition française, des dizaines de milliers d'archives et manuscrits notamment), et un accès direct aux documents numériques grâce à Gallica, ce qui est important dans l'univers du web où la pratique des internautes est de vouloir aller directement au document final.

Sur le web, les données de bibliothèques et, ici, de la BnF ont une valeur ajoutée qui suit un peu ce phénomène de longue traine, étudié par Chris Andersondata.bnf. Dans data.bnf, le choix a été fait de mettre progressivement les données des catalogues sur le web. Pour l'instant, nous n'avons que 1% des auteurs cités dans les catalogues de la BnF, mais ces 1% représentent 20% des notices du catalogue. Il faut dire que ces 1%, ce sont des auteurs comme Victor Hugo, comme Alexandre Dumas, avec beaucoup de notices bibliographiques associées qui, sur le web, sont attendues. Mais il y a également un certain nombre d'auteurs ayant moins de notices bibliographiques associées, mais qui sont intéressants pour des niches de marché pertinentes sur le web et qui, du coup, remontent dans les moteurs de recherche.

Cependant, les bibliothèques, et plus particulièrement la BnF, rencontrent des difficultés dues à 3 types de problèmes.

Le premier est de savoir comment un moteur de recherche peut traiter des millions de documents d'un coup, sans tri préalable. (La logique de Google est d'avoir un référencement hiérarchisé. Ce qui est au delà des 10 premiers résultats ne compte pas.)Le deuxième problème est celui de la diversité des sources. On a, dans data.bnf, par exemple, une base construite en EAD-XML, avec une logique hiérarchique, celle de « BnF Archives et manuscrits ». À côté, un autre silo de données, celui des notices bibliographiques, correspondant aux documents du catalogue général. À côté encore, les données numériques de Gallica, que l'on ne veut pas isoler. La jonction est très intéressante pour un utilisateur, car elle permet d'avoir le document numérique à côté de sa description complète dans le catalogue général. Il s'agit de rattacher ces silos. C'est en fait une opération d'interopérabilité.

Le troisième problème est que ces données sont dissimulées dans ce qu'on appelle le web profond. Elles sont certes structurées, mais ne sont pas forcément compréhensibles par des machines sur le web, et ne sont donc pas référencées par les moteurs de recherches ou pas nécessairement réutilisables par d'autres machines ne connaissant pas les formats spécialisés de bibliothèques. Sur le web, en effet, il y a de moins en moins de passage par les pages d'accueil des portails et, d'autre part, les recherches se font davantage par mots clefs. Pour aller cherche *les Misérables* de Victor Hugo, on ne

choisira pas d'y accéder via les *œuvres complètes* de Victor Hugo, mais on cherchera « Misérables, Victor Hugo » tout simplement. En outre, la recherche se fait de plus en plus par suivi de liens.

Trois enjeux sont liés à ces atouts et faiblesses des bibliothèques sur le web. Le premier est la visibilité. Le deuxième est l'interopérabilité – servir de pivot (c'était le premier nom du Web Data, « pivot documentaire ») en fédérant les ressources de la BnF en interne, en les liant à des ressources d'autres bibliothèques, voire à d'autres organisations qui n'ont rien à voir avec les bibliothèques. Le troisième enjeu est l'utilisation de ces données par des tiers qui ne maitrisent pas forcément les habitudes des bibliothèques en matière de données.

Prenons l'exemple « peuple bakossi ». En proposant la recherche « peuple bakossi » sur Google, les résultats proposent Wikipédia – aucune surprise – et, ensuite, on trouve data.bnf. Revenons à la page « peuple bakossi ». Il s'agit de traiter automatiquement les données de la BnF, en les regroupant sur des pages auteur, œuvre et thème qui sont les trois principales entités de ce site. Les données d'autorité constituent le socle de ces pages. On voit donc apparaître dans un encadré la donnée d'autorité RAMEAU avec toute la richesse de ses informations, les autres formes du nom Bakossi, le lien vers le catalogue général de la BnF ou le lien vers des données extérieures, en l'occurrence la notice équivalente dans la bibliothèque du Congrès américain. Il y a aussi la richesse des liens hiérarchiques, vers les pages « Bantous », « Ethnologie Nigéria », par exemple, puis, en dessous, les ouvrages de la Bibliothèque nationale avec le lien vers le catalogue, liés ici, de manière structurée, à la notice sujet.

On a aussi des pages auteur. Pour la recherche « Christine de Pisan », on accède à une notice d'autorité qui traite d'une personne avec les autres formes de son nom, ses sources, ses dates, les œuvres rattachées et d'autres activités, qui viennent à nouveau de la structure des données d'origine. On a des liens, dans les données d'origine et les notices bibliographiques, qui sont typés par des codes de rôle, en l'occurrence, des codes qui disent « Christine de Pisan est traducteur de la *Divine Comédie* de Dante » par exemple. On a des liens vers Gallica et les documents numérisés, des liens vers le catalogue BnF archives et manuscrits, des liens vers le catalogue général.

Pour des pages œuvres, l'utilisation des algorithmes d'alignement est important, à la fois pour que l'utilisateur puisse facilement retrouver des documents d'archives et des éditions papier de l'œuvre, mais aussi pour faciliter le référencement sur le moteur de recherche. L'algorithme d'alignement, qui commence à être assez classique dans le web, permet de regrouper une notice « œuvre » à une page « rôle » par exemple. Ce qui est assez récent en matière de bibliothéconomie de la BnF, c'est que ces traitements de liens entre un œuvre, une notice bibliographique, son édition, seront reversés dans les catalogues sources de la BnF. Le lien pourra donc être retrouvé directement dans la notice bibliographique.

Il existe d'autres exemples de traitement automatique. Le lien entre la Dewey, qui est le cadre de classement pour retrouver les livres dans les bibliothèques, et les thèmes RAMEAU, est fait de manière très sommaire, autour de grandes catégories.

On peut aussi créer des pages de dates, ou encore déduire tous les auteurs en relation. Par exemple, on trouvera autour d'Erasme le nom de Hans Holbein, en tant

qu'illustrateur de documents auxquels ce penseur a contribué. Il s'agit de traitements automatiques, et un affinage des règles pourra être nécessaire, mais c'est un premier travail intéressant.

Ce travail de traitement des données en aval du travail de production dans les catalogues, puis pour la diffusion sur le web de données s'appuie sur des principes fondamentaux : l'importance des identifiants (les URI ou identifiants web pérennes et http), qui, à la BnF, suivent le protocole ARC et un cadre de description qui s'appelle RDF. Ainsi peut-on lier non plus seulement des notices, mais des données (auteur, œuvre, etc.) entre elles. Le principe voulu par data.BnF est celui de l'ouverture : technique (accès aux données brutes par téléchargement RDF) et juridique (licence ouverte de l'État et logiciel libre cubicweb).

Pour les référentiels, le web de données a des potentialités intéressantes en termes d'économies de moyens, car il permet de lier les données plutôt que de les copier. En l'occurrence, dans data.BnF, on utilise des vocabulaires existants, compris par les acteurs du web qui ne sont pas nécessairement habitués à nos formats : FOAF[49] qui correspond surtout au vocabulaire des réseaux sociaux, mais qui s'est généralisé pour la description des personnes (dates de naissance, etc.), SKOS[50] pour la description des concepts, DUBLIN CORE notamment pour les éditions des œuvres.

Quelles perspectives pour ces référentiels ? Créer des liens internes, comme avec l'exemple des liens Dewey-Rameau, et créer des liens extérieurs. Sur data.bnf, on trouve des liens vers Wikipédia et DBpédia (version structurée pour le web de données de Wikipédia), vers VIAF, une base d'autorités internationale qui regroupent toutes les formes des noms, en France, en Russie, dans toutes les formes littérales possibles des noms, vers la bibliothèque du Congrès, la bibliothèque nationale allemande, geonames et versle thesaurus des Archives de France.

Ghislaine Glasson Deschaumes

La cote Dewey reflète les modes d'organisation des savoirs en Occident et elle fait débat dans le contexte d'une réflexion postcoloniale sur les modes de production et de circulation des savoirs en temps de mondialisation. Elle pourrait être révisée. Au fond, pensez-vous que data.BnF pourrait faire bouger la cote Dewey comme l'indexation RAMEAU commence de le faire ?

Agnès Simon

Comme le Data s'appuie davantage sur la notion de sujet que sur celle de classement, c'est donc RAMEAU qui constitue le socle des pages œuvres et qui est diffusé sur le web de données. Les cotes Dewey sont utilisées à une granularité très grossière, par exemple pour classer Viollet-le-Duc dans la classe « architecture ». Mais on n'a pas du tout d'axe de travail là dessus.

---

[49] Pour plus d'informations sur FOAF, consulter le site (en anglais) : www.foaf-project.org/

[50] Pour plus d'informations sur SKOS, consulter le site (en anglais) : www.w3.org/2004/02/skos/

Question sur les possibilités d'adapter data.bnf au local

Le produit data.BnF, lui-même, va conserver son identité BnF pour garantir la lisibilité des données. Actuellement, un prototype qui répond au cahier des charges d'un appel à service innovant lancé par le ministère de la Culture est en cours d'élaboration, avec la bibliothèque municipale de Fresnes et avec la bibliothèque départementale de Saône-et-Loire. Il s'agit de prendre toutes les données de data.bnf, de les croiser, au moyen de ce prototype de logiciel, avec les données d'une bibliothèque locale, par exemple la cote d'un document, de les croiser avec d'autres données extérieures également, et d'en ressortir des pages œuvres où l'on pourra retrouver le document de la bibliothèque locale – puisque tel est l'enjeu. Dans le cadre d'une bibliothèque départementale de prêt, qui fonctionne en réseau, cela ouvrira la voie à des possibilités de fédération.

Françoise Dalex

Vous avez expliqué qu'actuellement il y a 1% des auteurs représentant 20% des œuvres dans data.bnf. Comment avez-vous procédé aux choix pour data.bnf ? N'avez-vous pas pu tout mettre votre catalogue ?

Agnès Simon

Le choix a été d'avancer progressivement, à partir de données propres, simples aussi, et attendues (Victor Hugo, les classiques). Ensuite, comme il n'était pas possible de tout « donner à manger » d'un coup au moteur de recherche, la mise en ligne devait être progressive et régulière. Donc, les 1% correspondent au manuel Lagarde et Michard[51], auquel ont été ajoutés des corpus spécialisés considérés comme des niches, comme par exemple les juristes anciens, les auteurs antiques ou des auteurs étudiés dans les programmes.

Nous avons adapté, plus ou moins, le modèle FRBR de manière pratique, qui conçoit trois entités : œuvre, auteur, sujet, principalement. Schématiquement, on peut dire que, pour l'entité œuvre, les niveaux suivants sont distingués: *expressions*, par exemple la version traduite d'une œuvre, *manifestation* (édition, publication de l'œuvre), *exemplaire* de l'œuvre (par exemple l'œuvre numérisée).

En réponse à une question sur les liens possibles entre data.bnf et le projet du labex *Les passés dans le présent*.

Comment ce projet collaboratif de la BnF pourrait-il s'insérer dans le projet du labex ?

Agnès Simon

Il faudrait réfléchir autour de l'aspect de récupération de nos données. Qu'est-ce que les partenaires du Labex souhaiteraient récupérer ? Les données RAMEAU en RDF ? Un logiciel du type de celui qu'on a utilisé (CubicWeb) ? Est-ce que cet aspect logiciel, qui permet de regrouper des sources différentes, d'en sortir à la fois des pages avec une vue HTML pour le public et une vue en données brutes, intéresse le Labex ? Un

[51] Ancien manuel scolaire de littérature française.

dernier aspect, plus compliqué, serait de lier les données du Labex à celles de la BnF (lien entre deux silos).

Françoise Dalex

Comment data.bnf va-t-il s'inscrire dans le paysage de la BnF ? A-t-il vocation à devenir le moteur de recherche du site de la BnF ?

Agnès Simon

data.bf n'est pas un catalogue et ne remplit pas du tout ces fonctions en termes de localisation, de gestion des documents. Il ne peut pas remplacer Gallica. Son but est de conduire les internautes sur le site BnF. Pour l'instant, c'est le cas, puisque 80% des internautes de data.bnf viennent des moteurs de recherche et rebondissent sur les applications sources, c'est-à-dire Gallica, BnF archives et manuscrits, etc. Le taux de conversion est d'environ 70%, selon les mois. La position est de de différencier la partie production, avec ses outils et ses formats adaptés, et la partie diffusion. La question de savoir comment Data va vivre sur le long terme n'est pas encore tranchée. C'est au terme du marché de développement en cours que l'on s'interrogera sur ce qu'il y a lieu de faire. L'application doit-elle être conservée en tant que tel ou bien ses algorithmes et ses outils peuvent-ils être réutilisés dans les catalogues sources ou ailleurs ?

Françoise Dalex

Une question un peu technique, par rapport aux auteurs, parce que le projet *Sources de l'ethnomusicologie* se demande un peu par quels moyens relier les différents réservoirs. Il y a les mots-clefs, mais il y a aussi les autorités auteurs qui sont très importantes pour faire le lien dans les archives. Comment articulez-vous FOAF et les autorités auteurs dans VIAF ? Pour faire des liens par rapport à des autorités auteurs, que faut-il utiliser ?

Agnès Simon

Reprenons l'exemple de Christine de Pisan. On utilise 3 vocabulaires :

- le SKOS qui exprime les données liées à l'autorité, au niveau concept, par exemple, nom et autres noms de Christine de Pisan. On utilise FOAF, « friend of a friend », qui est l'expression de Christine de Pisan comme personne.
- Et VIAF fait aussi effectivement cette distinction de SKOS et foaf. Par ailleurs, les données sont alignées vers VIA qui regroupe beaucoup d'informations de bibliothèques et de différentes organisations, et permet ainsi de récupérer des liens qui sont déjà faits vers DBpédia ou vers la Bibliothèque allemande ou d'autres.

Françoise Dalex

Justement vous avez également parlé, tout à l'heure, de récupérer des liens qui étaient déjà générés grâce au système *LinkedData*. Vous réintégrez donc ces liens dans votre catalogue ?

Agnès Simon

Les liens VIAF, DBpédia ne sont pas réintégrés dans le catalogue, ils sont sur data.bnf. Par exemple, pour la page Houellebecq, on n'avait pas d'image Houellebecq, on est allés sur Wikipédia pour récupérer l'image de Houellebecq via ces liens. Ce qui est

répercuté dans le catalogue, par contre, ce sont les travaux d'alignement, entre autres. Le fait de retrouver cette édition « extension du domaine de la lutte » sur la page œuvre, la notice autorité titre « extension du domaine de la lutte », a exigé un travail automatique de traitement automatique parce que les liens ne suffisaient pas, on avait besoin d'un petit traitement automatique supplémentaire. Ce lien va être progressivement reversé dans une perspective de FRBérisation du catalogue, ce qui demande un contrôle intermédiaire. C'est un reversement semi-automatique.

Restitution de la journée et conclusion, Jean-Luc Minel, Ghislaine Glasson Deschaumes

Jean-Luc Minel

Data.bnf est pour moi le bon exemple du modèle que l'on peut chercher à avoir au labex et j'expliquerai pourquoi. Au point de départ, il y a des producteurs qui sont très différents, très hétérogènes, Gallica, le grand catalogue, etc. Le projet consiste donc à chercher comment lier ce type de productions, en s'appuyant, d'abord sur un standard et l'interopérabilité, donc RDF, et en réussissant à ne pas imposer un langage commun. Sur ce dernier point, il y a peut-être des désaccords et, en tout cas, un point méthodologique à préciser : Foaf, SKOS, ce sont des propriétés proposées, prédéfinies, par le langage du web sémantique, mais ce ne sont pas des listes d'autorités. Foaf est une grammaire, mais ce n'est pas une valeur que l'on attribue aux syntaxes. Dans les listes d'autorités, au contraire, on pourra avoir différentes formes de graphies. Ces listes sont indispensables, parce que les interrogations doivent pouvoir être faites par des publics scientifiques comme par des publics peu érudits. Il nous faut une forme qui ne soit pas une forme canonique, mais qui permette, quand même, de retrouver l'information. C'est l'un des choix, que l'on discute dans le groupe « Modélisation et référentiels ». Il faut utiliser des vocabulaires spécialisés existants, sans chercher à refaire le monde, et c'est ce qu'a fait data.bnf.

Un autre point très intéressant dans data.bnf, et qui peut être conflictuel, c'est que toute l'information n'est pas dans data.bnf, et que l'on peut aller la chercher ailleurs. C'est une question que l'on va se poser dans le labex. Estimons-nous que nous avonstoute la connaissance et qu'il nous appartient de la diffuser et de la maîtriser, ou bien acceptons-nous la diversité, le fait d'être un élément du puzzle de toute la connaissance ? Dans le projet Isidore[52], les producteurs ne sont pas maitrisés par le TGE Adonis, ce sont les producteurs qui signalent leur accord pour que l'on vienne chercher leur production. Pour le labex, c'est plutôt ce modèle là que j'aurais tendance à promouvoir, en disant « il faut accepter de ne pas tous être d'accord, de ne pas tous avoir les mêmes vocabulaires, de ne pas tous avoir les mêmes catégories. Mais il faut trouver lesmoyens pour ce faire, et le web sémantique est là pour ça ».

Un autre élément de réflexion, plus complexe, a trait à ce qui nous a été présenté en termes d'accès et d'usages par rapport aux différents publics. Ces différents niveaux sont présents dans data.bnf, mais de manière moins élaborée qu'à la Cité de la musique, par exemple. La question est de savoir quels outils utiliser et qui doit s'en occuper. Car cela suppose beaucoup d'heures de développement et de description.

Enfin, il y a la question de l'établissement de données structurées et extrêmement qualifiées, qui supposent beaucoup de temps passé à décrire, à affiner. Le choix fait sur data.bnf est de ne pas tout qualifier, car des algorithmes automatiques ou semi-automatiques font les traitements. Cela peut entraîner des erreurs, mais que l'utilisateur pourra accepter si on lui donne un lien pour trouver l'information. Pour les scientifiques, c'est plus difficile à accepter…

[52] Pour plus d'informations sur Isidore, consulter le site : www.rechercheisidore.fr/

Réponse à une question sur la validité des informations fournies sur le web, notamment du point de vue du CNRS…

Jean-Luc Minel

De mon point de vue, le web étant interactif, on peut s'attendre à des réactions lorsqu'il y a une donnée extrêmement conflictuelle ou insuffisamment qualifiée. Les sites peuvent prévoir d'intégrer ces retours en vue d'une modification. C'est le choix fait par Wikipédia.. Mais sommes-nous prêts à accepter cette logique ? La question se pose d'autant que Wikipédia est, un vecteur de la diffusion de la connaissance extrêmement important, qui a un impact extraordinaire : 90% des étudiants consultent Wikipédia ! Comment agir là-dessus ? C'est une question qui doit nous mobiliser.

Ghislaine Glasson-Deschaumes

En me resituant par rapport au début de la matinée, je voudrais relever quelques points.

Des choses importantes ont été soulignées ce matin en ce qui concerne le projet « Sources de l'ethnomusicologie », qui doivent nourrir la réflexion du labex Les passés dans le présent d'une manière plus générale. Ce projet s'insère dans une tradition qui est celle de la collecte des fonds sonores inédits. Cet intérêt pour l'oralité n'est pas dissociable de l'histoire des institutions culturelles et, donc, d'une histoire des processus et des modes de patrimonialisation dans les différentes institutions. Cet aspect a été beaucoup discuté au sujet des ATP, du musée de l'Homme, mais nous pourrions peut-être systématiser la réflexion sur le rapport entre les différents projets du labex qui portent sur des fonds et collections et l'histoire des institutions culturelles auxquelles ils sont liés.

Cette journée a par ailleurs permis d'engager une réflexion fructueuse sur la patrimonialisation des archives de chercheurs, qui mérite d'être étendue et discutée plus avant en associant d'autres projets du labex, comme par exemple celui sur les archives des sites et chantiers de fouilles archéologiques. Nous devons être encore plus volontaristes dans la mise en commun.

Une autre idée forte de cette journée, c'est que l'on passe de savoirs accumulés (les musées, les laboratoires sont des lieux d'accumulation des savoirs) à une logique d'architecture de cette accumulation, et doncà une logique d'articulation des savoirs. Cela déplace nécessairement les points de vue, tant dans la perspective de la description des fonds et collections (on l'a vu avec l'Index Rameau) que dans celle du chercheur. Et il nous faudra étudier plus avant de quelles manières cela déplace les points de vue.

Cela renvoie au statut et aux finalités de la numérisation. Le geste de numérisation est un préalable, sur ce chantier comme sur les autres dans le labex. Mais les nombreuses questions posées dans la matinée se rejoignent toutes pour demander : pour qui numérise-t-on ? Le large public ? Le public moins large ? Pourquoi numérise-t-on ? À qui s'adresse-t-on et comment ? On constate un certain flottement. Cette notion de large public, qui est quelque chose de très diffus, très insaisissable, et on ne peut la laisser flotter comme cela. Elle nous appelle à la vigilance. Le grand public de la Cité de la Musique n'est pas le même que celui de Gallica, de data.bnf, des futurs utilisateurs des Albums Valois en ligne, par exemple. Peut-être qu'une aide consisterait à réarticuler la

question des publics avec celle des pratiques culturelles. Concernant les sources de l'ethnomusicologie, nous savons bien, par exemple, que le rapport aux musiques orales passe d'abord par la scène, par la performance, par le spectacle vivant. Donc, comment un projet comme celui-ci peut entrer en résonance avec des pratiques culturelles actuelles, dont on n'a pas nécessairement identifié les acteurs et qui renvoient elles-mêmes à des publics extrêmement diversifiés ? Il ne faut pas perdre de vue ce rapport à la création. On pourrait imaginer un atelier du labex fédérant différents projets sur ce rapport à la création contemporaine et aux pratiques contemporaines de la création artistique...

Le dernier point à relever porte sur le partage des données. Les intervenants ont traité de questions hautement techniques et qui engagent une réflexion sur le numérique d'une grande profondeur et d'une grande ampleur de champ. Dans le même temps, on ne peut pas séparer ces questions du rapport entre la position des sachants et celle des enseignés. Comment déplacer cela ? Le labex a-t-il les moyens de déplacer ce rapport-là ? Au fond, il me semble que le partage des données est non seulement une question de politique scientifique, mais aussi une question de politique culturelle ou, comme Christine Guillebaud l'a dit ce matin, de politique tout simplement. Il serait bien de poursuivre cette réflexion au sein du labex, et de nourrir collectivement une réflexion sur les politiques culturelles à l'épreuve de la question du partage des données.

Une question, pour l'équipe[53] qui a préparé l'atelier, et que je voudrais remercier au nom du comité de pilotage du labex : de quelle manière cette journée vous a-t-elle permis d'avancer sur certains aspects de votre projet, et de quelle manière ? A-t-elle ouvert des perspectives nouvelles ?

Pascal Cordereix

La journée a ouvert des perspectives et elle a mis au jour une tension entre la question des publics et la question des modes de présentation des corpus sur lesquels nous travaillons. Tension dans la mesure où l'on est face à des archives qui représentent à peu près 10 000 heures d'enregistrement, ce qui est considérable. Un travail dans la finesse, comme celui de la Cité de la musique, n'est clairement pas possible pour un corpus de cette dimension là. Cela signifie qu'il y a des priorités, des choix, des choses à déterminer qui sont pour moi un peu le fruit de la journée.

Jean Lambert

Pour répondre à Ghislaine Glasson Deschaumes sur le rapprochement entre les archives et les politiques culturelles, nous aurons l'occasion, avec le projet de Christine Guillebaud, de nous interroger sur la manière dont ces musiques, souvent archivées chez nous, mais aussi ailleurs, sont valorisées dans des processus de patrimonialisation qui sont souvent liés à des problématiques d'identité nationale ou d'identité locale. Je pense par exemple au Congrès du Caire dont les archives sonores se trouvent à la BnF et qui sera au cœur de notre participation au programme. Récemment, j'ai reçu deux sollicitations, d'artistes ou de médiateurs, qui veulent produire quelque chose là dessus. C'est dans l'air du temps.

---

[53] Pascal Cordereix, Françoise Dalex, Aude Da Cruz-Lima, Joséphine Simonnot, Audrey Viault, Claire Schneider.