La diffusion des données de la recherche en Sciences humaines et sociales : l'exemple de la plateforme TELEMETA pour les données sonores

Sara Tandar

CONSERVATOIRE NATIONAL DES ARTS ET METIERS

Ecole Management et Société-Département CITS

INTD

MEMOIRE pour obtenir le

Titre professionnel "Chef de projet en ingénierie documentaire" INTD

RNCP niveau I

Présenté et soutenu par

Sara TANDAR

le 17 décembre 2013

La diffusion des données de la recherche en Sciences humaines et sociales

L'exemple de la plateforme TELEMETA pour les données sonores

Jury : Mme Ghislaine CHARTRON, Professeur titulaire de la chaire d'ingénierie documentaire du CNAM et directrice de l'INTD
Mme Aude JULIEN-DA CRUZ LIMA, Responsable de la gestion et de la valorisation des archives du CREM-LESC (UMR 7186)

Promotion 43

A Aude,

Remerciements

En premier lieu, je tiens à remercier Aude Julien-Da Cruz Lima qui m'a fait découvrir le vaste domaine de l'information et de la documentation, s'est si souvent rendue disponible et s'est montrée d'un appui sans faille. Merci à Joséphine Simonnot, Guillaume Pellerin et aux équipes de chercheurs et d'ingénieurs du CREM et du LAM pour leur confiance.

Je souhaite, également, remercier l'équipe pédagogique et administrative de l'INTD et l'ensemble des intervenants qui m'ont permis de mener à bien la formation et mes projets professionnels.

Un grand merci à tous les « collègues » rencontrés au cours des deux ans de formation avec qui les échanges ont été particulièrement enrichissants. Une grande pensée pour le groupe 1A dont le soutien et la qualité de travail m'ont été très précieux, et un clin d'œil particulier à Odile pour ses encouragements qui m'ont permis d'aller jusqu'au bout…

Merci à Isabelle pour la relecture.

Last but not least, un merci infini à Stéphane sans qui rien n'aurait pu se concrétiser.

Notice

Les références bibliographiques sont ordonnées selon leur ordre d'apparition dans le texte et introduisent des renvois vers la bibliographie analytique sous forme de liens hypertextes.

Résumé :

Dans un environnement numérique omniprésent lié au développement d'Internet et du Web, la communauté politique internationale a pris la mesure des enjeux de l'Open Access dont l'objectif est de soutenir la libre circulation des productions scientifiques. Les premières initiatives ont porté sur la création de plateformes de dépôt des publications d'articles et de travaux universitaires des chercheurs. Depuis environ cinq ans, en France, des projets de diffusion des données collectées par les chercheurs ont été mis en place. Ce mémoire dresse un état des lieux de plateformes de travail collaboratives pour l'exploitation des données dans le domaine des Sciences humaines et sociales, à partir d'une étude de projets dédiés aux données sonores et audiovisuelles ethnographiques. La diffusion de ce type de données est rendue complexe compte tenu des questions juridiques et éthiques et de problématiques techniques spécifiques. Ce contexte particulier met en lumière les difficultés de traitement et de gestion de ces données par le chercheur lui-même. L'auteur propose ainsi des pistes de réflexion aux producteurs de données pour anticiper au mieux les besoins des chercheurs et pour appréhender les contraintes liées à la mise en œuvre d'un projet de diffusion.

Descripteurs : SHS ; document audiovisuel ; document sonore ; accès libre ; professionnel de l'information ; donnée ; recherche scientifique ; plateforme ; TELEMETA ; ethnomusicologie

In an ubiquitous digital context linked to the development of the Internet and the Web, the international political community promotes Open Access and supports the free flow of scientific productions. Initial efforts focused on the development of a range of publishing systems platforms for research outcomes. For about five years, in France, research projects distributing of shared datasets are providing. This study presents an overview of collaborative authoring platforms for audiovisual and ethnographic data mining in the Humanities and Social Sciences. The distribution of these datasets poses many problems of legal and ethical background and technical issues. Study these particular datasets highlight the difficulties of promoting the take up of the workflow for their treatment and management by the user community. The author suggests to data producers how to involve researchers and offer ways for the implementation of software system for distributing data produced by scholarly research projects.

Table des matières

Table des matières

Liste des tableaux

Introduction

Depuis une dizaine d'années, les productions scientifiques au sens large font l'objet d'une réflexion et de collaborations à l'échelle internationale. En effet, la recherche scientifique est un vecteur de développement industriel et socioéconomique garanti par l'exploitation des résultats des projets de recherche menés à leur terme par la communauté scientifique. L'enjeu actuel est donc de mettre en place des systèmes servant de support à la diffusion des productions scientifiques. Avec l'utilisation massive d'Internet et les avancées technologiques qui lui sont associées, le secteur de la recherche a à sa disposition des moyens et des ressources qui lui offrent de nouveaux champs à investir pour développer des réseaux de compétences et mettre en place des projets offrant un environnement au sein duquel il devient possible de partager les savoirs sur le principe de l'Open Access. C'est pourquoi une réflexion a été mise en place dans le cadre du projet de création d'un Espace européen de la recherche (ERA) afin d'élaborer une politique commune qui soutienne la diffusion des productions scientifiques. Les projets d'e-infrastructures de recherche ont pour vocation de garantir l'interopérabilité des initiatives nationales et de fournir des espaces de collaboration et de mutualisation des compétences. Par ailleurs, les systèmes dédiés aux données de la recherche sont confrontés à des défis majeurs que représentent le stockage et l'archivage de données dont la croissance est exponentielle. Le développement d'outils et d'applications destinés à l'exploitation des données constitue un autre enjeu à relever. L'objectif est d'inciter les chercheurs à s'impliquer dans un nouvel espace de travail tourné vers le numérique. Cet environnement numérique impacte les méthodes de recherche et les pratiques des chercheurs. La recherche dans le domaine des Technologies de l'information et de la communication (TIC) représente un cadre technique porteur d'innovations qui bénéficie d'un soutien politique et financier important avec pour objectif de concevoir des solutions innovantes de partage et de diffusion des résultats scientifiques. Les productions scientifiques regroupent tout un ensemble de contenus variés. D'une part, il peut s'agir des résultats scientifiques portés à la connaissance du public, soit par le biais d'une publication dans des revues scientifiques en ligne[1], soit dans le cadre d'un dépôt sur une plateforme d'archives ouvertes. Dans ses recommandations pour l'exploitation des données numériques de la recherche financées par des fonds publics [1, OCDE, p. 18], l'OCDE définit les données de la recherche :

> *« Comme des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche. »*

De fait, à la base de tout projet scientifique, la collecte des données permet au chercheur de procéder à des expérimentations et d'en tirer des conclusions qui alimenteront ses études théoriques. Ces données doivent pouvoir être identifiées comme valides et fiables et donc provenir de sources

[1] En 2013, le journaliste scientifique John Bohannon a soumis un « faux » article, rédigé par un pseudo-scientifique dont le profil avait été créé de toutes pièces, à plus de 300 revues scientifiques. La moitié de ces revues ont accepté de le publier. Une polémique s'est alors développée autour de la crédibilité de certains titres et de la fiabilité de leur système de validation scientifique.

reconnues par la communauté. Les données ainsi collectées sont alors réutilisées et exploitées à des fins d'analyse dont les résultats, présentés sous forme de rapports ou d'articles, seront validés par les pairs avant leur publication et leur diffusion. Une donnée est dite primaire ou brute quand *« elle présente un caractère original, c'est-à-dire lue ou vue par le lecteur dans l'état où l'auteur l'a écrite ou conçue »* [2, ADBS]. Une donnée brute nécessite donc d'être contextualisée pour en extraire une information pertinente et produire un savoir et par conséquent, elle peut être réutilisée dans différents projets de recherche. Ainsi, les données brutes se révèlent essentielles à la compréhension des résultats scientifiques. Elles se composent aussi bien de données techniques, « *résultats d'expériences ou d'observations fournis par différents instruments, des magnitudes, […] des rayonnements, des températures, des précipitations, des altitudes* » [3, ROUGE-DUCOS] que de données collectées par le chercheur (carnets de notes, images, enregistrements sonores, vidéos, etc.). Ce dernier type de jeux de données est régulièrement produit dans le cadre des recherches en Sciences humaines et sociales (SHS) et leur diffusion pose des problématiques spécifiques : variété des supports et des formats, contraintes juridiques complexes, etc. Depuis quelques années, dans le domaine des SHS, des projets de diffusion de corpus scientifiques ont été réalisés. Les projets qui seront étudiés dans le cadre de ce mémoire ont été développés au sein d'unités de recherche. Leur point commun est de fournir aux chercheurs une plateforme technologique où déposer et mettre en ligne leurs données audiovisuelles. Cependant, les systèmes qui sous-tendent chaque projet proposent des services différents dont les atouts et les inconvénients sont à relever afin de proposer des pistes de réflexion à mettre en œuvre lors de la conception d'un projet de diffusion de fonds d'archives à destination de toute institution détentrice de données audiovisuelles.

Au regard de mon parcours universitaire et de mon expérience professionnelle, cette étude se focalisera essentiellement sur l'exploitation des données de la recherche appliquée au domaine de l'ethnologie et de l'ethnomusicologie. Ce périmètre d'étude met l'accent sur une discipline qui génère des jeux de données hétérogènes issues « du terrain » qui nécessitent de mettre en place des processus particuliers : numérisation des supports papier et analogiques, conservation des supports d'origine, diversité et obsolescence des formats numériques, gestion des droits de diffusion, accessibilité et diffusion auprès des populations d'origine. Par ailleurs, au-delà de leur valeur scientifique, les données ethnographiques représentent des enjeux patrimoniaux et culturels au sein des sociétés où elles ont été collectées, ce qui en fait un champ d'analyse singulier qui met en lumière un ensemble de contraintes auxquelles il faut être attentif lors de la mise en place d'un projet de diffusion des données. Il m'est donc apparu pertinent d'engager une réflexion sur l'exploitation des données ethnographiques mise en perspective dans le contexte plus global des SHS.

La première partie de ce mémoire présente le contexte politique actuel qui, en réponse au mouvement de l'Open Access, cherche à promouvoir et à favoriser la libre circulation des productions scientifiques. Le développement d'Internet et du Web et la recherche dans le domaine des TIC ont servi de support aux projets de plateforme de diffusion de la recherche dans le cadre de coopérations politiques internationales. Les moyens mis à contribution ont créé une synergie entre les équipes de

recherche et les équipes techniques qui a eu pour résultat le développement d'infrastructures dédiées aux données de la recherche. Ces infrastructures répondent-elles aux enjeux de l'Open Access ? Comment les technologies du web sémantique sont-elles intégrées à ces infrastructures ?

En deuxième partie, une sélection de projets de diffusion des données audiovisuelles ethnographiques permettra de dresser un panorama des services développés au sein des plateformes technologiques dans le domaine de la recherche en SHS en France et d'évaluer les processus à mettre en œuvre pour mettre à disposition les données scientifiques des chercheurs. Une attention particulière sera portée à un projet pilote du CNRS axé sur la diffusion d'archives sonores conçu par le laboratoire du CREM[2] : la plateforme collaborative TELEMETA. En quoi ces réalisations ont-elles ouvert de nouvelles perspectives pour la recherche en SHS ? Quels sont les enjeux à relever pour assurer leur exploitation auprès des chercheurs ?

En dernier lieu, nous proposerons aux producteurs de données sonores et audiovisuelles un ensemble de recommandations dont l'objectif est de fournir des pistes de réflexion qui permettent d'appréhender les principaux besoins techniques et documentaires à envisager dès la conception d'un projet de valorisation et de diffusion de fonds scientifiques.

[2] CREM, Centre de recherche en ethnomusicologie – LESC (Laboratoire d'ethnologie et de sociologie comparée – UMR 7186)

Première partie

1 Pour un partage des données de la recherche : un contexte politique propice et un environnement technologique porteur d'innovations

1.1 Internet, au cœur d'une stratégie politique internationale pour la recherche scientifique

1.1.1 L'Open Access, un mouvement pour « une société de la connaissance »

1.1.1.1 Internet et le Web : de nouvelles perspectives de diffusion de la recherche

La création d'Internet et du Web sont intrinsèquement liés à la recherche scientifique. C'est au cours de projets de recherche financés par l'armée américaine que le réseau Internet a été conçu. Le World Wide Web a, quant à lui, été créé pour favoriser les échanges internes entre chercheurs du CERN[3]. Ces développements technologiques ont ouvert de nouveaux horizons au secteur de la recherche scientifique. C'est ainsi que les usages d'Internet et les nouvelles pratiques initiées par le Web ont donné naissance au mouvement du Libre Accès, ou Open Access, qui repose sur le projet d'Open Archives Initiative (OAI). Ce projet consiste à faciliter l'échange des contenus numériques en ligne[4]. Dans la Déclaration Internationale de Budapest ou Budapest Open Archive Initiative (BOAI) [5, COLLECTIF] du 14 février 2002, les chercheurs signataires rappellent le devoir des scientifiques de publier leurs recherches dans une démarche d'ouverture et de partage de la connaissance. Cet appel revendique un accès libre, c'est-à-dire gratuit et sans restriction, aux résultats scientifiques et invite les institutions à y prendre part. Deux stratégies sont proposées à l'initiative des chercheurs : l'auto-archivage dans des archives électroniques ouvertes, qui sera désigné comme « la voie verte », et la création de nouvelles revues de publication axées sur le Libre Accès, qui deviendra « la voie dorée ». En 2003, « La Déclaration de Berlin sur le Libre Accès à la Connaissance à la Science exacte, Sciences de la vie, Sciences humaines et sociales », ou « Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities », marque un tournant pour l'Open Access. Plusieurs grandes institutions ont en effet pris conscience du rôle que pouvait jouer Internet dans la diffusion des savoirs et décident d'y participer. Elles répondent à l'appel de Budapest et adhèrent à ce mouvement. La Déclaration de Berlin réaffirme le droit à un accès libre aux résultats scientifiques ainsi qu'à l'ensemble des productions scientifiques :

[3] Dans les années 90, Tim Berners-Lee, informaticien au CERN (Centre européen de la recherche), conçut un système de liens hypertextes sur le réseau informatique interne afin de permettre l'échange et le partage d'informations entre les chercheurs et créa le premier navigateur et éditeur web qu'il désigna comme le World Wide Web et le premier serveur http [4, BERNERS-LEE].

[4] < http://www.openarchives.org/ > [consulté le 12 novembre 2013].

> *« Les contributions au libre accès se composent de résultats originaux de recherches scientifiques, de données brutes et de métadonnées, de documents sources, de représentations numériques de documents picturaux et graphiques, de documents scientifiques multimédia. » [6, COLLECTIF]*

L'élargissement du périmètre de diffusion à l'ensemble des données produites dans le cadre de la recherche fait évoluer le statut des données brutes et de leurs métadonnées. Elles deviennent des productions à valoriser au même titre que les publications, puisqu'elles sont au cœur du travail de recherche des scientifiques. C'est à partir de l'exploitation et de l'analyse de données brutes que les chercheurs formulent des hypothèses de recherche et publient leurs résultats d'étude qui, après validation par les pairs, seront publiés. D'autre part, la déclaration de Berlin désigne un ensemble d'acteurs dont le devoir est de répondre aux exigences de partage et de diffusion prônées par l'Open Access :

> *« Les responsables politiques en charge de la science, les institutions de recherche, les agences de financement, les bibliothèques, les archives et les musées »*

Ce mouvement, initialement porté par une petite communauté de chercheurs, fait donc l'objet d'une reconnaissance institutionnelle à partir de 2003. Cela contribue à l'émergence d'une dynamique qui a créé un contexte propice à l'élaboration d'une réflexion commune et à la mutualisation des savoir-faire et des compétences des différents acteurs impliqués dans l'Open Access. L'objectif final est de concevoir des projets innovants qui soutiennent et facilitent l'échange des données de la communauté scientifique.

1.1.1.2 Une reconnaissance politique internationale des TIC pour répondre aux enjeux de l'Open Access

En 2003, à Genève, s'est tenu le Sommet Mondial sur la Société de l'Information (SMSI) organisé par les Nations Unies. Un ensemble d'acteurs politiques et d'organisations internationales du secteur privé et de la société civile se sont réunis afin de poser les bases d'une société de l'information accessible à tous. La déclaration issue de cette première phase du SMSI reconnaît la nécessité de valoriser et de diffuser la science, à la source d'innovations au service de la société de l'information :

> *« Nous reconnaissons que la science joue un rôle capital dans le développement de la société de l'information. Bon nombre des éléments constitutifs de la société de l'information sont la conséquence des progrès scientifiques et techniques rendus possibles par la mise en commun des résultats de la recherche. » [7, COLLECTIF]*

Internet est reconnu comme l'instrument répondant aux objectifs de développement de la communication du savoir puisqu'il fournit un réseau mondialisé sur lequel s'appuyer. Un des enjeux

relevés lors de ce sommet est de fournir des moyens et des ressources afin de soutenir la recherche dans le domaine des Technologies de l'Information et de la Communication (TIC). En effet, les TIC sont à même d'apporter des solutions innovantes pour enrichir la valeur des contenus mis en ligne. Dans les engagements de Tunis, lors de la deuxième phase du SMSI qui s'est déroulée en 2005, l'apport des TIC est réaffirmé et redéfini dans le cadre d'une stratégie commune de gouvernance d'Internet [8, COLLECTIF]. Connecter les centres scientifiques et les centres de recherche aux TIC se situe d'ailleurs en troisième position (sur dix cibles à atteindre) des projets mis en place[5]. Les TIC au service de la recherche scientifique sont donc au cœur des préoccupations des Nations Unies pour atteindre les objectifs de développement énoncés dans la Déclaration du Millénaire d'ici à 2015.
En publiant les « Principes directeurs pour le développement et la promotion du Libre Accès » en 2013, l'UNESCO encourage les décideurs et responsables politiques à engager une politique nationale de Libre Accès. L'UNESCO met en valeur l'impact de la recherche sur l'économie nationale, le développement humain, les activités commerciales, le dialogue entre les cultures, et préconise de soutenir le développement des TIC :

> *« L'UNESCO s'attache donc, dans ses domaines de compétence, à améliorer l'accès à l'information et au savoir pour le bénéfice de ses États membres en mettant à profit les technologies de l'information et de la communication. » [9, SWAN]*

Internet porte l'espoir de créer une interconnexion entre les peuples et les sociétés et les TIC semblent offrir les moyens de rendre le monde plus équitable et solidaire à travers l'élaboration d'outils facilitant les échanges et la diffusion de la connaissance.

1.1.2 Un Espace européen de la recherche à l'initiative de projets d'infrastructures pour l'exploitation des données

1.1.2.1 Le projet d'Espace européen de la recherche

A partir de l'an 2000, le lancement du projet de création d'un Espace européen de la recherche (ERA – European Research Area)[6] s'inscrit dans une action politique s'engageant pour la libre circulation des chercheurs, des savoirs et des technologies. Le développement de la recherche dans le domaine des TIC bénéficie ainsi du plus important budget de l'ensemble des thématiques soutenues dans le cadre du programme de l'ERA :

[5] Un résumé analytique du rapport sur le développement des télécommunications/TIC dans le monde datant de 2010 est disponible en français à l'adresse suivante : < http://www.itu.int/dms_pub/itu-d/opb/ind/D-IND-WTDR-2010-SUM-PDF-F.pdf >. Source : < http://www.itu.int/pub/D-IND-WTDR-2010/fr > [consulté le 5 novembre 2013].
[6] < http://ec.europa.eu/research/era/index_en.htm > [consulté le 5 novembre 2013].

> *« Les TIC sont identifiées comme un des moteurs de l'amélioration de la compétitivité européenne et doivent avoir un impact sur la productivité et l'innovation (en facilitant la créativité et le management) ainsi que sur la modernisation des services publics (santé, éducation, transport). Elles doivent par ailleurs soutenir les avancées scientifiques et technologiques en facilitant la coopération et l'accès à l'information. » [10, CNRS]*

La Communauté européenne prend donc la mesure des enjeux liés au Libre Accès qu'elle considère comme un vecteur de développement socioéconomique porteur d'innovation et favorisant la compétitivité [11, SIREN]. Internet, en tant que réseau mondialisé, permet une diffusion immédiate et illimitée à l'Information Scientifique et Technique (IST) et les TIC fournissent les moyens d'enrichir les contenus en proposant des outils de recherche et de travail innovants ainsi que la possibilité de relier des informations disséminées sur la Toile. Le développement d'infrastructures, notamment pour mettre à disposition les données de la recherche, est encouragé. En 2007, la Commission européenne rappelle dans sa communication sur l'information scientifique à l'ère numérique que :

> *« Les données issues d'activités de recherche entièrement financées par les pouvoirs publics doivent en principe être accessibles à tous, en accord avec la Déclaration ministérielle de l'OCDE de 2004 sur l'accès aux données de la recherche financée par des fonds publics. » [12, COMMISSION DES COMMUNAUTES EUROPEENNES]*

Comme le souligne Thérèse Hameau, de l'INIST-CNRS[7], dans son article « Panorama des projets européens en faveur du libre accès à l'IST », les données de la recherche représentent donc un nouvel enjeu pour les institutions [13, HAMEAU] :

> *« Si au départ, ce sont les articles de revues scientifiques qui sont visés, on assiste rapidement à un élargissement à d'autres types de documents : les ouvrages, les données de la recherche, ainsi que tout ce qui a trait à l'enseignement, comme les cours et les thèses. Cette évolution est liée au fait que ce sont des archives institutionnelles qui se développent en lien avec les mandats des institutions. »*

1.1.2.2 e-Science, des infrastructures technologiques au service des données des chercheurs

Les problématiques liées à l'exploitation des données de la recherche ont concrètement été abordées dans le projet UK e-Science. En 2000, John Taylor, alors directeur général des conseils de recherche au sein du Department of Trade and Industry britannique, veut rendre accessible, à des fins d'exploitation, la profusion et la variété des données collectées par les chercheurs dans le cadre de leurs activités. Il lance le projet UK e-Science qu'il définit ainsi :

[7] Institut de l'information scientifique et technique

> *« e-Science traite de la collaboration mondiale dans des domaines clés de la science, et de la prochaine génération d'infrastructures qui permettra cette collaboration. » [14, DESRUELLES]*

Des équipes réunissant des chercheurs, des informaticiens et des ingénieurs mettent en commun leurs compétences. Ces collaborations créent une synergie entre les scientifiques et les experts techniques avec comme objectifs de fournir des moyens innovants à l'exploitation des données de la recherche, de leur collecte à leur traitement : grille de calcul[8], espace de stockage, services (gestion, enrichissement des données, etc.), applications (visualisation et modélisation), développement de surface de réseaux. UK e-Science, qui a bénéficié de moyens financiers conséquents, a été l'initiateur de nombreux autres projets nationaux et européens de constitution de grilles de calcul et d'infrastructures d'exploitation des données de la recherche.

Afin d'intégrer les initiatives nationales dans un cadre plus global qui permette de les mettre à profit de l'ensemble de la communauté scientifique, la Commission européenne se prononce sur sa volonté d'harmoniser les e-infrastructures (ou cyberinfrastructures) réalisées et d'assurer leur interopérabilité afin de répondre aux défis de l'e-Science et de mutualiser les expériences internationales en ce domaine. C'est dans ce cadre que l'ESFRI, European Strategy Forum on Research Infrastructure, a été mandaté en 2002 pour coordonner les réalisations et les stratégies nationales afin de garantir les relations entre les répertoires numériques et l'optimisation des usages des infrastructures. Le projet CHAIN-REDS[9] est un projet orienté vers la recherche dont l'objectif est de :

> *« Développer un écosystème d'infrastructure électronique mondial qui permettra à des communautés virtuelles de recherche (VRC), des groupes de recherche ou des chercheurs, d'utiliser efficacement les ressources réparties à travers le monde. »*[10]

En France, le ministère de l'Enseignement Supérieur et de la Recherche (MESR) a vocation à soutenir la politique de développement des infrastructures de la recherche. Avec l'élaboration d'une stratégie nationale, le MESR souhaite positionner la France en leader dans la construction d'une Europe des infrastructures [16, MESR] et soutient les initiatives nationales. Ainsi, à partir de 2005, le CNRS a mis en place le TGE-Adonis[11], une très grande infrastructure « visant à faciliter le tournant

---

[8] « L'objectif d'une grille de calcul est de concevoir une architecture informatique permettant de mettre à disposition des utilisateurs toutes les ressources dont ils ont besoin au moyen d'une interface simplifiée. » [15, CHARRON, p. 3]

[9] Ce projet a été mis en place en 2012, < http://www.chain-project.eu/ > [consulté le 10 novembre 2013].

[10] < http://www.ambafrance-cn.org/Seminaire-europeen-e-Infrastructure-for-e-Science.html > [consulté le 29 octobre 2013].

[11] Le Très grand équipement Adonis est une infrastructure de recherche nationale dont la mission principale est d'assurer l'accès et la préservation des données numériques produites par les sciences humaines et sociales. Il propose une grille de services à disposition des équipes de recherche : conservation des données, traitement, diffusion. Le TGE-Adonis et le Corpus-IR (une très grande infrastructure visant à produire des corpus numériques documentarisés collectifs reposant sur des

numérique de la recherche en Sciences humaines et sociales » [17, HUMA-NUM] qui propose aux chercheurs en Sciences humaines et sociales (SHS) un accompagnement et une technologie pour la gestion de leurs données. Cette infrastructure soutient des projets pilotes initiés par des laboratoires de recherche détenteurs de corpus de données pour lesquels les défis technologiques et documentaires posent des problématiques singulières.

En 2002, le projet Minerva a, quant à lui, voulu répondre au défi majeur que constitue la numérisation du patrimoine culturel et scientifique européen afin de constituer des corpus de données à mettre à disposition. En effet, il existe tout un ensemble de données de la recherche constituées avant l'ère du numérique qu'il convient de porter à la connaissance de la communauté scientifique et du grand public. La numérisation de ces données permet de proposer à la consultation en ligne des versions numérisées, tout en garantissant la préservation et la conservation de supports fragiles et peu manipulables. En France, le ministère de la Culture et de la Communication a lancé un grand projet national de numérisation, coordonné par la Mission de la Recherche et de la Technologie (MRT), qui a profité à différentes institutions (MMSH[12], BnF, etc.). La diffusion et l'exploitation des données scientifiques représentent un enjeu considérable pour l'IST qui, selon la définition du MESR :

> *« Regroupe l'ensemble des informations produites par la recherche et nécessaires à l'activité scientifique comme à l'industrie. » [16, MESR]*

Les missions des professionnels de l'IST consistent à valoriser et à faciliter la diffusion de ces informations. Elles sont donc étroitement associées aux TIC dans la mesure où leurs objectifs ont en commun de concevoir des structures et des outils d'analyse innovants qui répondent aux besoins de diffusion, de traitement et d'exploitation de l'information. La prise de conscience de la communauté internationale des bénéfices apportés par les outils de télécommunication a permis de fournir à la recherche dans le domaine des TIC les moyens de répondre aux enjeux de la diffusion des savoirs. C'est à travers le développement d'infrastructures, la création d'outils et d'applications qu'il sera possible de fournir un accès à l'information équitable et démocratique. L'Open Access a permis de créer l'impulsion nécessaire à l'élaboration de solutions techniques et technologiques au service du secteur de la recherche scientifique. Les institutions patrimoniales, culturelles et scientifiques ont mesuré les avantages à réunir et mutualiser leurs propres expériences et leurs compétences. De plus, le soutien de la classe politique internationale a permis de fournir les ressources et les moyens nécessaires à la création d'infrastructures favorisant les échanges et le partage des productions scientifiques.

---

formats ouverts) ont été fusionnés en 2013 pour devenir la TGIR Huma-Num du CNRS (Très grande infrastructure de recherche en humanités numériques).

[12] Maison méditerranéenne des sciences de l'Homme, < http://www.mmsh.univ-aix.fr/Pages/default.aspx > [consulté le 12 novembre 2013].

1.1.3 HAL : un exemple d'archive ouverte dédiée à la recherche

1.1.3.1 Le dépôt dans une archive ouverte, pour une visibilité et une diffusion immédiate

Les premières réalisations ont concerné essentiellement le développement d'archives ouvertes. En France, le projet HAL (Hyper Articles en Ligne)[13] permet aux chercheurs de déposer leurs articles, publiés ou non, qui deviennent alors accessibles gratuitement et de manière permanente. L'auto-archivage est aujourd'hui fortement encouragé par les organismes de recherche suite à la demande de l'Agence nationale de la recherche (ANR) en 2007 de verser dans le système national des archives ouvertes HAL toutes les publications issues des projets qu'elle finance. En effet, les archives ouvertes facilitent la diffusion et la mise à disposition des productions scientifiques tout en proposant un système d'archivage pérenne.

Le déposant porte l'entière responsabilité du contenu déposé, comme le précise le guide du déposant accessible en ligne :

> *« Le dépôt d'un document sur Hal est placé sous la responsabilité de la personne qui l'effectue ; il implique l'accord de l'ensemble des auteurs. Ces derniers conservent évidemment l'entière propriété intellectuelle de leur travail. Les documents déposés doivent avoir un contenu susceptible d'intéresser la communauté des chercheurs de leur domaine »*[14].

HAL a été développé par le CCSD[15] (Centre pour la communication scientifique directe) géré par le CNRS, et bénéficie d'une collaboration avec l'INRIA. HAL se présente comme un outil de dépôt et de diffusion : il ne permet ni le traitement, ni l'analyse des données. Le principe est de fournir le texte intégral et des métadonnées minimales permettant de décrire le contenu, les références de publication ainsi que le rattachement institutionnel du chercheur[16]. Ces métadonnées sont utilisées pour faciliter la recherche et les extractions de données qui complètent automatiquement les listes existantes de publication et d'institutions. Ces données sont moissonnées par le protocole OAI-PMH[17], ce qui favorise la visibilité des dépôts. Ce protocole d'échange des données assure plusieurs points d'accès à un document. Ainsi, un document dont les métadonnées sont renseignées et structurées sera accessible depuis plusieurs moteurs de recherche, bases de données et portails institutionnels interopérables. HAL est ainsi un point d'entrée vers l'archive ouverte de prépublications électroniques d'articles « ArXiv » où sont transférés automatiquement tous les dépôts de chercheurs dans les domaines des mathématiques, de la physique et de la biologie. Ce principe n'est garanti qu'à condition que le document ait une url pérenne, c'est-à-dire que son emplacement soit indépendant d'un

---

[13] < http://hal.archives-ouvertes.fr/ > [consulté le 5 novembre 2013].
[14] < http://www.ccsd.cnrs.fr/support/content/PDF/docHAL.pdf > [consulté le 13 novembre 2013].
[15] < http://www.ccsd.cnrs.fr/index.html > [consulté le 5 novembre 2013].
[16] Pour les détails, voir le guide d'utilisation en ligne, < http://www.ccsd.cnrs.fr/support.html > [consulté le 5 novembre 2013].
[17] Open Archives Initiative - Protocol for Metadata Harvesting ou « Protocole de Collecte de Métadonnées de l'Initiative Archives Ouvertes ». Ce protocole a été conçu dans le cadre de l'Open Archives Initiative. Pour les spécifications, consulter le site < http://www.openarchives.org/ > [consulté le 31 octobre 2013].

système ou d'une interface. Or HAL distribue des adresses http stables : cela a l'avantage de permettre aux déposants d'être cités et d'avoir une adresse de localisation permanente pour chacune de leur publication. HAL propose également différents services : l'import en masse d'articles au format XML[18], la consultation publique différée d'un document déposé, l'accès aux statistiques de consultation des dépôts, etc. Il est également possible de configurer des interfaces de dépôt et/ou de consultation pour répondre à des besoins spécifiques de certains domaines de recherche[19].

1.1.3.2 Un service d'archivage pérenne certifié

Pour les archives scientifiques mises en ligne, la norme d'archivage AFNOR AF Z 42-013 mise à jour en 2009 est en vigueur. Celle-ci définit les « Spécifications relatives à la conception et à l'exploitation de systèmes informatiques en vue d'assurer la conservation et l'intégrité des documents stockés dans ces systèmes ». Cette norme met l'accent sur le respect de l'intégrité du document : celui-ci doit pouvoir être lisible dans sa forme originale de manière permanente. De plus, l'archivage électronique des données numériques ou numérisées doit garantir la traçabilité des modifications et des enrichissements apportés au document.

Les services de l'archive ouverte HAL répondent à des enjeux de conservation et d'archivage à long terme. Tout dépôt au format .pdf est automatiquement envoyé au CINES (Centre Informatique National de l'Enseignement Supérieur) [18, CINES] qui fournit une solution pour la conservation à long terme du patrimoine numérique scientifique. Le CINES a conçu un système d'archivage en collaboration avec l'ABES[20] qui mutualise la plateforme pour tous projets d'archives reconnus d'intérêt national. Ce service est conforme à la norme ISO 14721 basée sur le modèle de référence OAIS[21], respecte le protocole d'échange standard (OAI-PMH), tient compte des évolutions technologiques liées aux formats et supports. Il offre également une puissance de calcul intensif aux chercheurs pour l'exploitation des données scientifiques (visualisation, modélisation, etc.). La Plateforme d'Archivage du CINES (PAC) a reçu l'accréditation DSA, Data Seal of Approval, attribuée aux centres de préservation numérique dont les procédures d'assurance qualité garantissent l'accessibilité et l'intelligibilité des informations qui leur sont confiées[22]. Ainsi, lorsqu'un chercheur dépose de nouvelles versions d'un document dans HAL, la traçabilité des dépôts successifs est garantie par une date certifiée par le système et le serveur conserve l'ensemble des versions. HAL représente un excellent exemple d'archive ouverte. D'une part, la plateforme de dépôt fournit un environnement technique et documentaire qui assure aux déposants et aux lecteurs une visibilité et une diffusion immédiate. HAL

---

[18] Extensible Markup Language. XML est un méta-langage dont les balises permettent de structurer le texte.

[19] Par exemple, la version HAL-SHS spécialisée dans les sciences humaines et sociales (SHS), <http://halshs.archives-ouvertes.fr/ > ou le serveur TEL (Thèses en Ligne) pour l'auto-archivage des thèses et des habilitations à la recherche, < http://tel.archives-ouvertes.fr/ > [consulté le 5 novembre 2013].

[20] Agence des Bibliothèques d'Enseignement Supérieur

[21] Reference Model for an Open Archival Information System. Il définit l'architecture et les fonctionnalités d'un système d'archivage.

[22] < http://www.cines.fr/spip.php?article804 > [consulté le 13 novembre 2013].

met à disposition une documentation et des supports détaillant chacune des fonctionnalités et des outils mis en ligne afin d'aider les démarches de dépôt. Lors du dépôt, le professionnel de l'IST et le chercheur peuvent être amenés à travailler ensemble pour en assurer la meilleure visibilité possible. Par ailleurs, HAL s'est adressé au CINES, un organisme externe agréé pour la prise en charge de la conservation et de l'archivage numérique à long terme. Cette mutualisation des compétences crée un système efficace qui répond aux besoins des chercheurs et est en conformité avec les normes et standards internationaux. Cette répartition garantit la viabilité des processus mis en place par l'équipe en charge des dépôts et par celle en charge de l'archivage.

1.1.3.3 Un projet évolutif pour être au plus près des besoins des chercheurs

La plateforme de dépôt HAL est actuellement en cours de réécriture : les évolutions les plus importantes concernent l'élargissement des types de documents déposables aux vidéos, la création d'un moteur de recherche et la possibilité de créer une page chercheur [19, CAPELLI]. Le moteur de recherche actuel ne permet pas de faire une recherche dans le contenu, c'est-à-dire sur le texte intégral, ce qui réduit les possibilités d'affiner sa recherche. Par ailleurs, suite au succès de la plateforme de dépôt d'images scientifiques MédiHAL[23] créée en 2010, une plateforme dédiée au dépôt des vidéos par des chercheurs et par des laboratoires ou des institutions de recherche est en développement. La nouvelle version de la plateforme HAL créera des liens entre les productions scientifiques, les documents annexés, les images et les vidéos déposées. HAL proposera également la création d'un référentiel « chercheur » sous forme de page personnelle sur laquelle l'utilisateur pourra rassembler les publications d'un auteur, intégrer son CV, sa bibliographie et des éléments statistiques. L'équipe en charge du projet HAL porte donc un regard réflexif sur les fonctionnalités et services qu'elle propose. Elle enquête et s'appuie sur les besoins des chercheurs afin de répondre au mieux à l'évolution de leurs pratiques. La mise en place d'indicateurs précis permet d'observer les pratiques du public ciblé et d'évaluer les usages qui sont faits de l'outil. Il s'agit, après analyse, de définir des priorités sur les éventuels développements ou améliorations à réaliser pour optimiser l'utilisation de la plateforme. Ce suivi encadré est primordial puisqu'il peut déterminer les conditions d'existence de la plateforme.

1.1.4 Conclusion

A partir des années 2000, la communauté internationale a posé les bases d'une « société de la connaissance » qui s'appuie sur Internet et les TIC. En parallèle, la communauté scientifique a exprimé sa volonté de mettre à profit le développement numérique pour ouvrir l'accès des productions scientifiques et mettre à disposition les publications et les données de la recherche à l'ensemble des chercheurs. Cette réflexion croisée a créé l'impulsion nécessaire à la mise en place des premières e-infrastructures dédiées aux données de la recherche afin d'y intégrer les réalisations nationales. Or les projets élaborés par les centres de recherche doivent répondre à un environnement et à des

[23] MédiHAL est une archive ouverte qui permet de déposer des images scientifiques et des documents iconographiques de science, < http://medihal.archives-ouvertes.fr/ > [consulté le 13 novembre 2013].

technologies en constante évolution. Comment intégrer ces évolutions ? Les équipes de recherche ont-elles les ressources suffisantes pour les mettre en place ? De quelle façon les chercheurs utilisent ces outils au sein de leurs activités de recherche ?

1.2 Internet et le Web : des supports techniques et technologiques de diffusion et de partage des données de la recherche

1.2.1 Le web de données : une nouvelle ère de l'Internet

1.2.1.1 Web sémantique et web de données

Depuis les origines du World Wide Web, l'environnement d'Internet a connu de profondes mutations auxquelles il a fallu s'adapter. Le réseau mondial que représente Internet permet la diffusion *« d'une masse en croissance exponentielle d'informations disponibles sur la Toile sous des formes multiples »* [20, PEUGEOT, p. 193]. Ces informations sont issues de secteurs divers et variés (scientifiques, industriels, commerciaux, gouvernementaux) ainsi que de contenus mis en ligne à travers les blogs, les réseaux sociaux, les sites de partage. Or, face à l'explosion du volume de données en ligne[24], des enjeux documentaires et technologiques sont à relever : garantir un accès pérenne à des données aux formats et natures divers et variés, faciliter leur recherche, leur consultation, leur échange, leur utilisation. En 2001, Tim Berners-Lee évoque pour la première fois le concept de « web sémantique » dont la BnF propose cette définition [21, BNF] :

> *« Un ensemble de technologies développé par le W3C (l'un des principaux organismes de normalisation du Web) visant à faciliter l'exploitation des données structurées, notamment en permettant leur interprétation par des machines.*
>
> *Le web de données (Linked Data en anglais) combine les technologies du web sémantique avec les principes fondamentaux du Web (protocole HTTP, identifiants URI), avec pour objectif la construction d'un réseau d'informations structurées, disponibles en ligne et facilement réutilisables dans de nombreux contextes. »*[25]

Avec le web de données, il ne s'agit plus seulement d'avoir accès à des pages HTML[26] reliées entre elles par des liens hypertextes comme sur le modèle du web de documents[27]. L'objectif est d'extraire

[24] La notion de « big data » désigne cet ensemble de données à croissance exponentielle qui requiert des infrastructures spécifiques pour leur gestion.

[25] < http://www.bnf.fr/fr/professionnels/web_semantique_donnees/s.web_semantique_intro.html > [consulté le 12 novembre 2013].

[26] L'HyperText Markup Langage, est un langage normalisé de marquage et de structuration des documents hypermédias sur le web [2, ADBS].

[27] Le web des documents permet de consulter des pages au format HTML reliées entre elles grâce à un système de renvois : l'hypertexte. L'internaute peut naviguer au sein d'un document, passer d'un document à un autre et accéder à des pages web externes. Dans ce système, les navigateurs et les moteurs de recherche parcourent des pages HTML générées et stockées dans le « web profond »,

des informations à partir des ressources en ligne : on parle alors de métadonnées qui sont des données sur les données primaires. Ces métadonnées décrivent les ressources et les relient entre elles à travers des liens typés, signifiants pour les machines puisqu'ils explicitent les relations entre chaque ressource. Les machines peuvent ainsi interpréter les relations entre les données et répondre de manière fine aux requêtes des internautes. Cette mise en relation des données construit un réseau documentaire global qui permet de naviguer entre les données enfin sorties de leurs silos. Cet espace repose sur des standards du Web développés par le W3C[28] qui garantissent l'interopérabilité des données par les machines. Les métadonnées deviennent donc des éléments essentiels pour décrire et structurer l'information contenue dans une ressource. Jean-Michel Salaün, chercheur en sciences de l'information, le résume parfaitement :

> *« Dans le web de données, l'unité de base n'est plus le document comme précédemment, mais la donnée. On ne relie plus des documents entre eux, mais des données entre elles. Le langage dominant n'est plus celui qui permet de rédiger un texte et de le relier à d'autres (HTML), mais celui (RDF) qui permet de décrire et relier entre elles les données […]. Le lien entre les données autorise la constitution de documents à la volée selon la navigation de l'internaute et pour éclairer son action » [23, SALAUN, p. 15].*

Le web de données contribue donc à une révolution numérique qui change le rapport à l'information. Grâce à l'exploitation des liens pertinents et significatifs entre les objets et du principe d'interopérabilité, les machines ont vocation à faciliter les recherches de l'internaute et à lui fournir un ensemble d'informations devenues accessibles et consultables.

Le web sémantique offre une technologie dont l'objectif est de garantir l'interprétation et l'utilisation des jeux de données ainsi constitués. A cet effet, la publication des données et de leurs métadonnées doit être réalisée selon un vocabulaire normalisé. Des schémas de description des métadonnées sont associés à chaque type de ressources :

- ISBD[29] pour la description bibliographique ;
- Dublin Core simple[30] un format générique pour la description de ressources numériques ;
- EAD[31] pour l'encodage des fonds d'archives et la description des manuscrits.

---

c'est-à-dire dans des bases de données non interrogeables qui forment des silos ne communiquant pas entre eux [22, BERMES]. Le contenu de ces documents étant peu structuré, les machines ne peuvent pas les exploiter : impossible pour l'usager de faire des requêtes sur une information précise contenue dans la page, ou d'interroger directement les bases de données.

[28] Le World Wide Web Consortium est l'organisme de standardisation du Web, <http://www.w3.org/Consortium/mission.html> [consulté le 12 novembre 2013].

[29] International Standard Bibliographic Description, < http://www.ifla.org/publications/international-standard-bibliographic-description > [consulté le 12 novembre 2013].

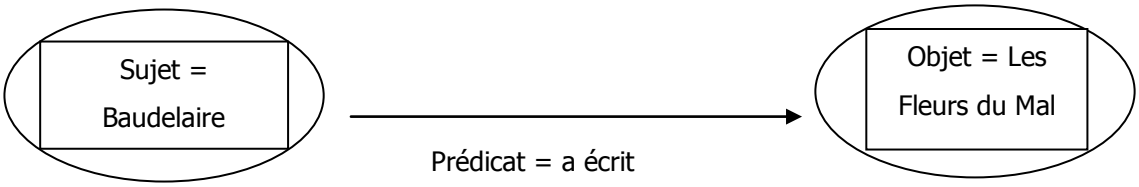
[30] Dublin Core Metadata Initiative, < http://dublincore.org > [consulté le 12 novembre 2013].

[31] Encoded Archival Description, <http://www.archivesdefrance.culture.gouv.fr/gerer/classement/normes-outils/ead/ > [consulté le 12 novembre 2013].

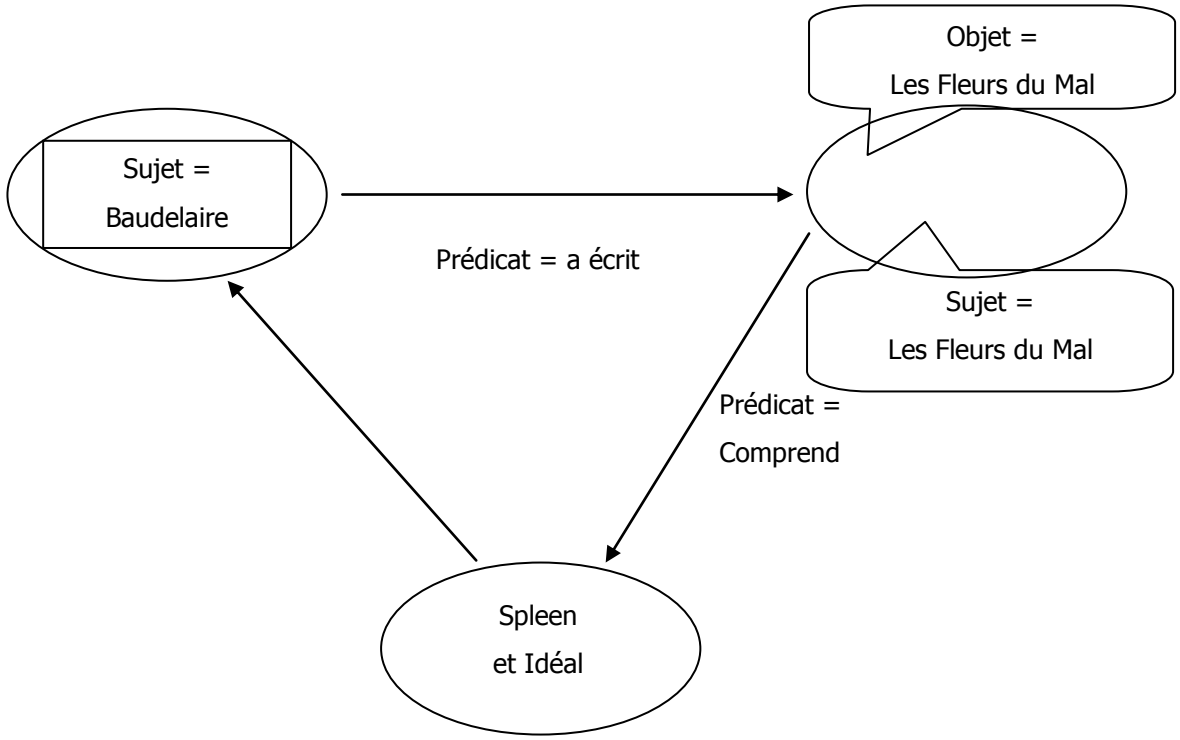
Le protocole OAI-PMH récupère ces métadonnées et permet leur échange. Les institutions déposent leurs données dans des entrepôts OAI qui les centralisent et les mettent à jour automatiquement. Des fournisseurs de services (ou moissonneurs) consultent et récoltent les données et leurs métadonnées et les mettent à disposition de l'internaute. Ce protocole implique que les détenteurs de données traitent et fournissent leurs fonds afin qu'ils soient exploités. Interopérabilité, ouverture et exploitation des données sont les fondements de cette approche qui recompose l'architecture du Web.

1.2.1.2 Le modèle RDF : un standard de description au cœur du web de données

Le modèle RDF (Resource Description Framework) est un standard de description des ressources web et de leurs données développé par le W3C qui s'exprime à l'aide de balises XML. Ce modèle a pour postulat de décrire tout objet à partir d'un triplet, sujet-prédicat-objet, qui explicite par une phrase simple les relations entre les données. On parle d'assertion.



Ce modèle permet de construire tout un ensemble de relations où l'objet d'une assertion devient lui-même le sujet d'une autre assertion et ainsi de suite.



Ainsi, chacune des informations autour d'une ressource est découpée, ce qui permet à chaque assertion de faire sens, indépendamment les unes des autres. Cela permet également de compléter l'information d'une ressource en ligne et de l'enrichir régulièrement : la granularité de l'information s'en trouve affinée. Ce modèle fait éclater les informations autour d'une donnée, ce qui facilite leur sélection et leur manipulation. Le web de données permet donc à l'internaute de naviguer et de rebondir entre des informations issues de sources diverses et variées.

A chaque ressource est attribué un identifiant pérenne, l'URI[32], qui localise la ressource, la désigne de manière univoque et exprime les relations qui ont été établies avec les autres données liées. Les URI sont attribuées et enregistrées par l'IANA[33] afin de garantir qu'elles soient interprétables par le web de données et qu'à une URI corresponde une seule ressource.

On construit alors un graphe qui offre pour chaque ressource un cadre de description, un identifiant unique et une localisation pérenne. Dans le domaine de la recherche scientifique, le cadre conceptuel de description des ressources en RDF fournit un modèle d'échanges et d'interconnexions des métadonnées garanti par les standards ouverts recommandés par le W3C. Pour les chercheurs, cela signifie pouvoir rechercher des informations ciblées extraites de leurs silos, tout en ayant la possibilité de sélectionner les données qui leur seront le plus utile. Quant aux URI, elles fournissent un accès pérenne : l'accès aux ressources n'est donc plus dépendant d'une interface ou d'un système et les citations et renvois aux sources d'informations en sont alors facilités.

[32] Uniform Resource Identifier
[33] Internet Assigned Numbers Authority

Les infrastructures pour les données de la recherche actuellement développées répondent-elles aux technologies du web sémantique ? Comment les chercheurs exploitent-ils les outils de diffusion de leurs données ? Quel rôle ont à jouer les professionnels de l'information dans ce nouvel environnement numérique ?

1.2.1.3 Les ontologies : une modélisation de la connaissance

Les ontologies sont des vocabulaires normalisés utilisés pour représenter de manière formalisée des connaissances et les décrire :

> *« Une ontologie définit un vocabulaire commun pour les chercheurs qui ont besoin de partager l'information dans un domaine. Elle contient des définitions lisibles en machine des concepts de base de ce domaine et de leurs relations. » [21, BNF]*

Les ontologies sont donc conçues pour être intelligibles par les machines afin d'en assurer leur interprétation, comme le précise Manuel Zacklad, directeur du laboratoire DICEN[34] au CNAM :

> *« Les ontologies formelles ne sont pas faites pour être directement exploitées par des usagers humains engagés dans une navigation hypertextuelle comme cela pourrait être le cas pour une classification documentaire ou un thésaurus. Au contraire, elles sont le plus souvent conçues pour être exploitées par des programmes informatiques (des agents de recherche automatique sur le web), l'utilisateur interagissant avec l'agent à l'aide d'un formulaire ou d'un autre type de langage de requête ». [24, ZACKLAD]*

Elles ont vocation à donner du sens aux données en décrivant des concepts et en précisant leurs relations dans un domaine de connaissance spécifique. Le standard de description des langages de représentation recommandé par le W3C, OWL[35] se construit sur le modèle de données RDF. Il permet de définir des classes et de *« contraindre plus précisément leurs description (en les décrivant comme union, intersection, complémentaire d'autres descriptions ou comme l'ensemble d'un certain nombre d'individus).* » [25, TRONCY]. Par exemple, L'ontologie FOAF[36] permet de décrire un profil utilisateur. Ce langage permet également de créer des domaines de relations ou des relations entre les concepts.

1.2.2 La plateforme Isidore, un portail d'accès aux données de la recherche en Sciences humaines et sociales (SHS)

1.2.2.1 Une plateforme de collecte et de traitement des données

En France, le projet ISIDORE[37] a été soutenu et développé par le TGE-Adonis[38] au CNRS. ISIDORE est une plateforme de collecte et d'indexation des données et métadonnées de la recherche

---

[34] DICEN – Dispositifs d'information et de communication à l'ère numérique.
[35] Web Ontology Language.
[36] Friend of a Friend, < http://xmlns.com/foaf/spec/ > [consulté le 12 novembre 2013].
[37] ISIDORE pour « Intégration de services, interconnexion de données de la recherche et de l'enseignement ». Ce projet a été mis en œuvre par la Très grande infrastructure de recherche Huma-

en SHS. Elle offre un accès unifié à différentes sources et ressources en accès libre produites en France. Toute institution détentrice de données scientifiques en SHS a ainsi la possibilité de contribuer à ce projet en versant ses propres données numériques dans la plateforme ISIDORE [26, CAPELLI]. Cette plateforme s'appuie sur les principes du web de données et met à disposition des données en accès libre. Depuis sa mise en place, ISIDORE a été intégré au projet européen DARIAH[39] dont la mission est de développer et soutenir la recherche et l'enseignement dans les disciplines des SHS en créant une infrastructure favorisant la mise en réseau d'outils, d'informations, de méthodologies et de personnes au service de la recherche en SHS.

La collecte des données et métadonnées dans ISIDORE est assurée par trois connecteurs informatiques asynchrones et un connecteur synchrone[40] (SRU/SRW) qui moissonnent l'information :

- le protocole OAI-PMH récupère les métadonnées. Celles-ci doivent être exprimées en Dublin Core simple[41] selon le modèle XML ;
- les flux de syndication tels que RSS, ATOM jouent un rôle d'entrepôt de données : un flux de syndication pour un item embarque un ensemble de métadonnées. Ces métadonnées sont décrites selon le schéma XML RSS 2.0 ou ATOM ;
- le protocole Sitemaps/XML, pour lequel les données et métadonnées doivent être fournies au format RDFa avec un enrichissement de la structuration en Dublin Core simple. Ce protocole permet au producteur de ressources de fournir à ISIDORE un fichier Sitemaps avec les pages web dont le contenu scientifique est pertinent à collecter : cela limite ainsi la collecte aux pages les plus pertinentes ;
- le connecteur synchrone SRU/SRW compatible avec le protocole Z-3950 permet d'étendre une recherche à des catalogues bibliographiques dont les données ne sont pas indexées par ISIDORE. Les métadonnées au format MARC21 et UNIMARC dans leur version XML sont compatibles avec cette application.

Ces connecteurs, en tant que barrière d'entrée à ISIDORE, sont volontairement faibles afin d'étendre à la grande majorité des laboratoires la possibilité d'intégrer ISIDORE. En effet, tous les centres de recherche n'ont pas les ressources nécessaires pour enrichir les métadonnées sur le standard RDF et la multiplicité des formats des contenus acceptés par ISIDORE leur permet d'être visibles et moissonnés sur cette plateforme. Les producteurs sont tenus de respecter des formats de description des données, tels que stipulés dans le guide de dépôt fourni sur le site[42]. Ce traitement à mettre en œuvre par le producteur est un préalable avant toute collecte par ISIDORE. Il a pour but de faciliter la

---

Num, et le Centre pour la Communication Scientifique Directe (CCSD). <http://www.rechercheisidore.fr/> [consulté le 13 novembre 2013].

[38] Le Très grand équipement Adonis est une infrastructure de recherche nationale dont la mission principale est d'assurer l'accès et la préservation des données numériques produites par les sciences humaines et sociales. Il propose une grille de services à disposition des équipes de recherche : conservation des données, traitement, diffusion. Le TGE-Adonis et IR-Corpus (une très grande infrastructure visant à produire des corpus numériques documentarisés collectifs reposant sur des formats ouverts) ont été fusionnés en 2013 pour devenir la TGIR Huma-Num du CNRS (Très grande infrastructure de recherche en humanités numériques).

[39] Digital Research Infrastructure for the Arts and Humanities, < http://dariah.eu/ > [consulté le 13 novembre 2013].

[40] Un connecteur synchrone effectue une recherche en temps réel dans des catalogues spécifiques qui ne sont pas indexés par la plateforme ISIDORE.

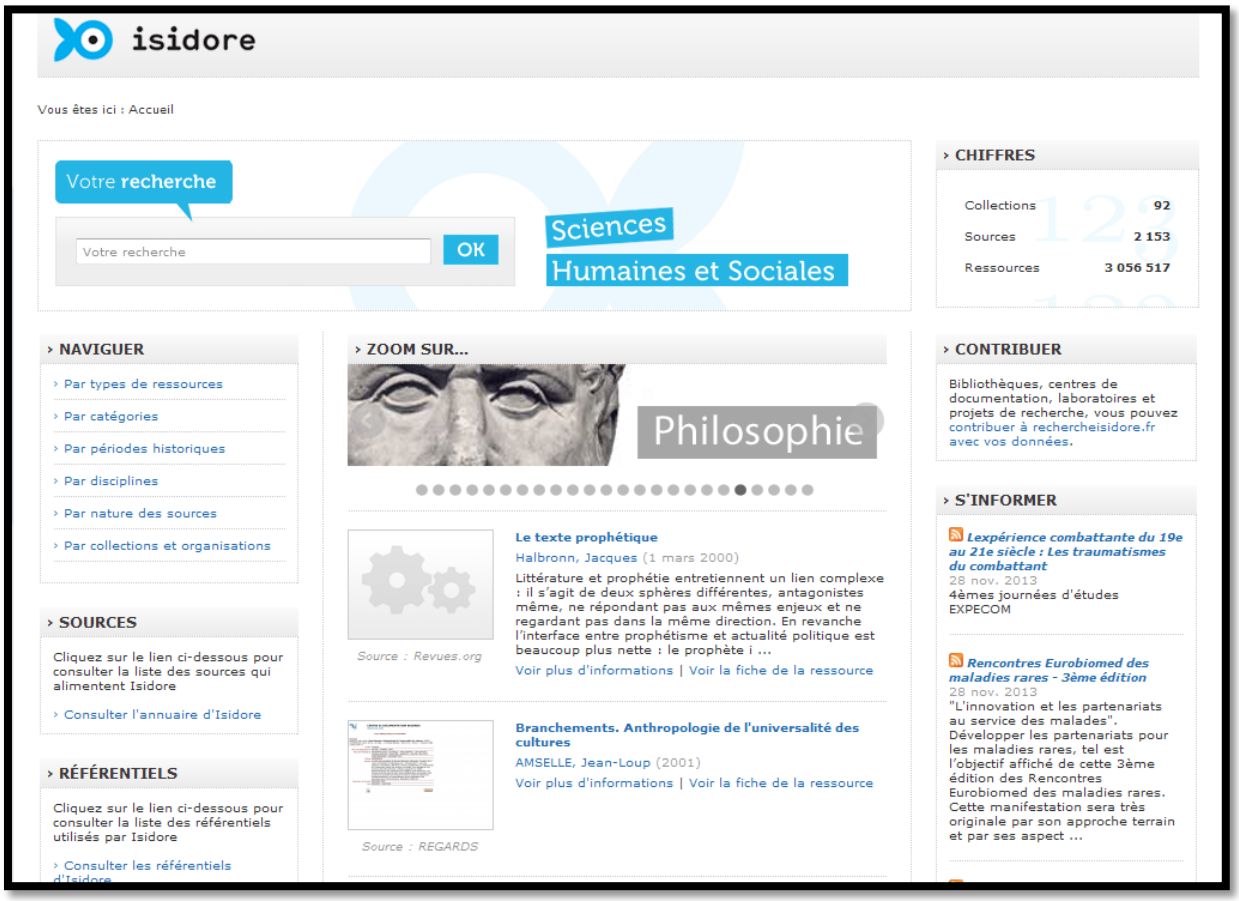
[41] < http://dublincore.org > [consulté le 10 novembre 2013].

[42] < http://www.huma-num.fr/sites/default/files/ressourcesdoc/guide_isidore_2012.pdf > [consulté le 10 novembre 2013].

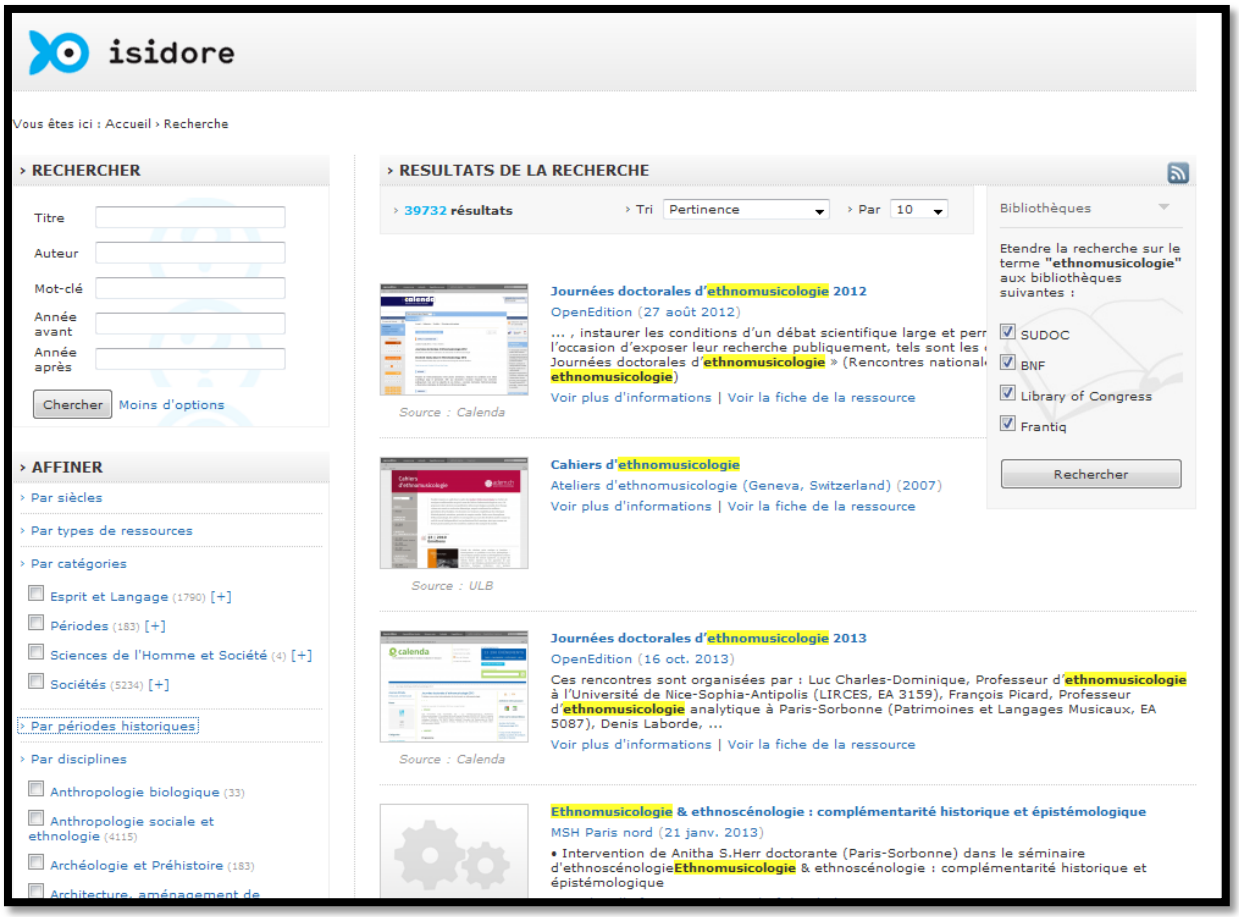
recherche des utilisateurs et d'optimiser la valorisation de ces ressources. Une fois la collecte effectuée, ISIDORE traite les données pour s'assurer de leur conformité vis-à-vis des spécifications d'ISIDORE, pour normaliser dans un format unique certains éléments (date, langues, etc.), pour aligner différentes valeurs aux référentiels utilisés par ISIDORE[43] (référentiel auteur d'ISIDORE, vocabulaire RAMEAU, les référentiels HAL-SHS et GeoNames, le thésaurus Pactols, etc.), et enfin pour convertir les données au format RDF. Un des points forts d'ISIDORE est que l'enrichissement automatisé des données indexées par le moteur peut être récupéré par les producteurs/contributeurs afin qu'ils améliorent les performances de leur propre système. Cette rétroaction fait partie des services à mettre en œuvre dans les infrastructures dédiées aux données de la recherche car ils apportent une valeur ajoutée à l'outil. Cela peut encourager des organismes, qui n'ont pas les moyens de traiter leurs données conformément au web de données, à contribuer au projet. D'autre part, ISIDORE fournit une expertise et des technologies qu'il n'est pas toujours facile à mettre en place dans de petites structures.

[43] Les référentiels sont structurés selon le vocabulaire SKOS qui permet d'utiliser les thésaurus, classifications et vocabulaires contrôlés dans le web de données.

1.2.2.2 Une interface conviviale et intuitive



L'interface utilisateur d'ISIDORE se veut assez ergonomique et intuitive pour ne pas mettre à disposition des utilisateurs une aide en ligne. La page d'accueil donne une vue simplifiée et structurée des différentes fonctionnalités proposées par ISIDORE. Ainsi, la barre de recherche simple permet d'exploiter les référentiels d'ISIDORE sous forme de suggestions de recherche au cours de la frappe (auto-complétion), les encarts ont des intitulés simples et décrivent en une phrase le contenu auquel ils renvoient sous forme de liens hypertextes. La colonne centrale propose un éclairage sur une discipline qui change chaque jour, ce qui dynamise la page d'accueil.



Après avoir lancé une requête, ISIDORE fournit une liste de résultats avec pour chaque ressource la présentation de son contenu sous forme de notice simple. On retrouve une recherche par facettes qui permet d'affiner les résultats, un lien hypertexte externe qui renvoie aux catalogues de bibliothèques, un raccourci sous forme de vignette qui permet de s'abonner au flux RSS de la requête pour en suivre l'actualité, un affichage des résultats par pertinence, auteurs, date, etc. La recherche avancée s'effectue sur des champs spécifiques avec l'opérateur booléen implicite « ET ». La gestion des opérateurs booléens n'est pas à l'initiative de l'utilisateur.

Lors de l'affichage d'une ressource sélectionnée, l'utilisateur a la possibilité d'enregistrer directement la référence bibliographique de la ressource dans Zotero (vignette dans la barre du navigateur), d'accéder à la ressource dans son contexte d'origine (par l'URI originale de la ressource ou par la capture d'écran), de bénéficier d'un contexte d'indexation enrichi (sous forme de nuage de mots-clés,

de liens hypertextes, d'une classification signifiante par discipline ou catégorie, d'enrichissements terminologiques issus de référentiels établis par des institutions reconnues, d'une visualisation de la couverture géographique, etc.), de partager la fiche sur les réseaux sociaux et de rebondir sur d'autres ressources.

Les différentes fonctionnalités qui viennent d'être décrites sont utilisées dans la plupart des sites à destination du grand public[44]. Ce parti pris éditorial a, semble-t-il, vocation à créer un contexte familier afin de faciliter la recherche et la consultation des ressources moissonnées par ISIDORE. Le chercheur est amené à consulter les ressources et à naviguer de manière autonome. L'exemple d'ISIDORE permet également d'envisager les avantages qu'apporte le web de données pour l'indexation des données de la recherche. En effet, certaines institutions reconnues ont développé, selon leur domaine d'application, des thésaurus et/ou vocabulaires contrôlés qu'ils exploitent dans leur propre système de catalogage. Le web de données offre le moyen de les mettre à disposition de l'ensemble de la communauté en SHS et de les croiser en créant de nouvelles relations entre ces vocabulaires. Ce tronc commun de références fiables et validées peut être réutilisé immédiatement par les producteurs de données et les chercheurs dans le cadre de leurs travaux, ce qui donne une cohérence et harmonise tout un domaine de recherche. Ainsi, non seulement le web de données favorise la mutualisation des connaissances, mais il permet de sortir les données de leurs silos et crée une dynamique de consultation (rebonds, extension des résultats, recherche fédérée, etc.).

1.2.3 Digital Humanities et IST

1.2.3.1 Nouvelles pratiques et auto-formation des chercheurs

Le développement du numérique dans le domaine de la recherche implique pour les chercheurs des pratiques et des méthodes bien éloignées de leur champ de recherche et de leur formation initiale. Il s'agit là d'un véritable défi à relever, en particulier pour les disciplines en Sciences humaines et sociales qui utilisent des données disparates et hétérogènes auxquelles le Web donne facilement accès. Le chercheur en SHS a dorénavant à sa disposition un ensemble de sources et de ressources, d'outils et d'applications qui peuvent lui être particulièrement utiles mais qui exigent des compétences particulières.

Les perspectives amorcées par le numérique ont de fait contribué à l'émergence d'une communauté dédiée aux SHS et à l'exploitation de ses ressources numériques : les Humanités Numériques ou Digital Humanities (DH). Ce mouvement qui soutient la diffusion des savoirs se situe au croisement des disciplines des SHS et des technologies numériques[45] [26b, INTD-CNAM]. En effet, le numérique a contribué à une évolution plus rapide qu'auparavant des méthodes de recherche, dont les DH ont pris conscience. En effet, pour nombre de chercheurs, il a fallu s'adapter et s'auto-former afin d'acquérir de nouvelles compétences pour mettre à profit les technologies nouvellement offertes dans le cadre de leurs activités de recherche. Certains chercheurs ont alors souhaité créer une dynamique

[44] Sites commerciaux, réseaux sociaux, etc.

[45] < http://dhi.intd.cnam.fr/ > [consulté le 15 décembre 2013].

transdisciplinaire qui favorise les échanges et les interactions entre les acteurs de la communauté des SHS [27, DACOS]. Le développement de listes de diffusion [28, DH] a contribué à donner naissance à un réseau dont la portée est aujourd'hui internationale. De nouveaux profils de chercheurs, alliant recherche académique et compétences numériques, ont vu le jour. Pourtant, ces compétences croisées ne sont pas toujours reconnues dans le domaine académique et peinent à être valorisées. En France, en particulier, c'est une vision encore très traditionnelle qui domine [29, TERRAS]. Aujourd'hui, il est devenu évident qu'il faut former les futurs chercheurs au contexte numérique. Actuellement, les cursus en SHS intègrent de plus en plus des cours sur les technologies numériques. Des séminaires et des formations spécialisés en DH se mettent en place pour accompagner et former le futur chercheur en SHS aux techniques de recherche et à l'utilisation des outils mis à sa disposition afin de le rendre le plus autonome possible. Il s'agit également de lui donner un regard réflexif sur la place des TIC dans son travail quotidien.

1.2.3.2 Le rôle des professionnels de l'information scientifique et technique

Le développement des TIC œuvre à l'ouverture des données et au développement d'un environnement résolument tourné vers le numérique. Pour autant, tous les chercheurs en SHS ont-ils pris la mesure des enjeux du numérique pour leur travail de recherche et sont-ils investis dans ces nouvelles pratiques ? De fait, tous les chercheurs ne s'inscrivent pas dans une démarche d'ouverture. Certains mettront en avant des questions de droit et de sécurité, ces aspects étant une préoccupation fondamentale qu'il est absolument nécessaire de prendre en compte. D'autres considèrent que participer bénévolement et de manière volontaire au partage et à l'enrichissement des données ne relève pas de leurs compétences, ni de leurs activités de recherche. En effet, si les pratiques de contributions et de partage développées par le web 2.0[46] ont été bien reçues dans le domaine privé avec le développement des réseaux sociaux et plateformes de partage en ligne, elles ne sont pas toujours aussi faciles à mettre en place dans un contexte de recherche scientifique. Par ailleurs, certains chercheurs rencontrent des difficultés dans un univers technique et technologique qu'ils considèrent comme hors de leur portée. On peut dès lors se demander si ces nouvelles approches méthodologiques ont du sens dans le cadre de leurs activités afin de mieux comprendre une certaine résistance au changement initié par les outils numériques et des réticences de la part de chercheurs. C'est dans ce cadre que les professionnels de l'IST ont un rôle essentiel à jouer. En effet, ils ont dû adapter leurs méthodes au contexte numérique, s'approprier les nouvelles technologies et répondre aux évolutions, plus ou moins subies par la communauté scientifique, des pratiques et des méthodes de recherche. Le périmètre d'action des IST s'est a fortiori étendu à l'environnement technologique

[46] Le web 2.0, appelé également web collaboratif, désigne un environnement technologique du web dans lequel les interfaces sont simplifiées et nécessitent moins de compétences techniques. Les internautes y ont découvert un univers à leur portée au sein duquel ils jouent un rôle central. Ils se sont approprié la Toile en en devenant les acteurs principaux. Ils partagent des ressources en ligne, indexent leurs documents avec un vocabulaire personnalisé, intègrent de nouveaux réseaux sociaux et/ou professionnels, travaillent en ligne sur des plateformes de travail collaboratif, créent des espaces d'échange et d'entraide sous forme de blogs, de forums, de commentaires, etc.

des chercheurs [30, DALBIN]. A l'heure du web de données, cela est d'autant plus vrai : l'exploitation des ressources en ligne et la gestion des données des chercheurs représentent un nouveau défi. La structuration de l'information, les modèles et formats de description des données et leur traitement sont autant de processus documentaires qui nécessitent une veille régulière. Or ces processus sont au cœur du savoir-faire des professionnels de l'information, ce qui en fait des interlocuteurs privilégiés pour le chercheur. Le professionnel de l'IST a les compétences requises pour accompagner le chercheur tout au long de son projet et peut l'aider à déterminer quels sont ses besoins afin de lui proposer ce qui lui sera le plus utile.

Par ailleurs, les infrastructures destinées aux chercheurs ont tout intérêt à être développées avec le soutien des professionnels de l'IST. En effet, leur mission étant la diffusion et la valorisation des travaux scientifiques, ils sont au plus près des pratiques des chercheurs. Ils sont confrontés au quotidien à l'évolution de leurs méthodes de recherche et aux problématiques qu'elles induisent. Ils appréhendent donc objectivement les attentes des chercheurs et possèdent une expertise et une expérience à mettre à profit lors de l'élaboration d'applications et d'outils de travail. Ils deviennent une force de propositions se positionnant du côté des chercheurs et de leurs besoins. Un des enjeux à relever est de faciliter l'autonomie des chercheurs dans leur prise en main des plateformes. C'est pourquoi il me semble que les professionnels de l'IST doivent être impliqués dès lors qu'une équipe de recherche projette de créer ou de s'intégrer à une plateforme de partage des données de la recherche. Ils servent d'intermédiaire entre les experts techniques et les équipes de recherche afin de rendre intelligible l'environnement technologique et complexe qui sous-tend toute infrastructure. Un autre aspect qui pourrait sembler anecdotique est celui de communiquer autour du projet. Le professionnel de l'IST doit avoir les moyens d'engager un plan de communication au moment de la mise en place d'un projet de valorisation des données au sein d'un laboratoire : présentation, démonstration, formation, aide à la mise en ligne (soutien individuel, suivi), actualités, alertes sur les dernières mises à jour ou sur les développements en cours, etc. Cela permet d'amorcer un changement en douceur et éventuellement, de répondre rapidement aux problèmes rencontrés par les chercheurs. La réactivité du professionnel de l'IST et le suivi régulier des utilisateurs et de leurs usages sont les conditions sine qua non pour créer un environnement propice à la mise en place de nouvelles pratiques au sein de l'équipe. Les professionnels de l'IST doivent également être les acteurs de la maintenance documentaire[47] : gestion des profils utilisateurs, mise à jour des vocabulaires et des listes contrôlées, suivi des développements, mise en place d'indicateurs statistiques, suivi et validation des contributions, etc. En effet, ces questions relèvent essentiellement du domaine de l'information et de la documentation. Les technologies actuelles ne permettent pas une automatisation de tous les processus documentaires et les machines, aussi puissantes soient-elles, ne sont pas encore capables d'être totalement autonomes. Ainsi, le rôle des professionnels de l'IST reste essentiel pour accompagner le chercheur à l'ère du numérique.

[47] La maintenance technique relevant des compétences des ingénieurs informatiques et/ou des développeurs.

1.2.4 Conclusion

Depuis qu'Internet est devenu le principal instrument de diffusion des connaissances et du savoir, les usages et pratiques de la communauté scientifique ont fortement été impactés. Nous assistons à un renouveau des méthodes de recherche en SHS qui s'inscrivent dans un monde du tout numérique. Le contexte politique est propice à la mise en place de projets de diffusion des données de la recherche. Cependant, on assiste à une évolution rapide de l'environnement technique et technologique (web de données, interopérabilité entre les systèmes, respect des normes et standards, etc.) auquel doivent faire face les équipes en charge des projets de diffusion des données. En France, les premiers projets pilotes de plateformes en SHS ont été mis en place entre 2007 et 2010. Quelques années plus tard, quel est le premier bilan que nous pouvons en retirer ? Ces plateformes ont-elles su répondre aux objectifs de l'Open Access ? Comment les projets concernant les données brutes des chercheurs ont-ils été reçus par les principaux intéressés ? Ces plateformes répondent-elles à l'ensemble des besoins des chercheurs en SHS ? Quels usages en font les chercheurs ? De quelle manière sont-ils impliqués dans le versement de leurs données ? A quelle(s) étape(s) du travail de recherche les systèmes mis en place sont-ils réellement utilisés ? Il paraît nécessaire d'établir des indicateurs précis qui rendent compte des apports et des développements nécessaires à l'optimisation de l'usage de ces plateformes de recherche par la communauté scientifique. L'ensemble de ces paramètres permettront de vérifier si les objectifs de départ ont bien été atteints.

D'autres questions d'ordre plus général peuvent se poser : comment éviter que les contraintes juridiques et linguistiques soient un frein à une diffusion internationale des données de la recherche ? Comment créer une synergie afin de développer des bonnes pratiques et des méthodes de travail communes à l'intention de la communauté scientifique en SHS ? Dans le cadre de l'ERA, certaines de ces questions ont été soulevées et des projets sont mis en place pour tenter d'y répondre. Les nouvelles pratiques des chercheurs en SHS deviennent en elles-mêmes un objet d'étude et d'analyse afin d'harmoniser et d'optimiser des méthodes de travail efficaces et pertinentes. Webdatanet[48] est un projet de plateforme de discussion dont le but est d'établir une méthodologie pour la collecte de données via le Web. Le projet Nedimah[49], Network for Digital Methods in the Arts and Humanities, propose aux chercheurs un espace où échanger et partager leurs retours d'expérience avec comme finalité de contribuer à la définition de bonnes pratiques autour de la gestion des données de recherche. Par ailleurs, le réseau Clarin, Common Language Resources and Technology Infrastructure, a pour mission de développer un portail centralisé avec un point d'accès unique[50] qui agrège l'ensemble des infrastructures pour la recherche et intègre des options de langue. Cette plateforme devrait favoriser le partage des données et la création d'un réseau international de chercheurs, indépendamment des obstacles linguistiques. Ces démarches soutiennent l'innovation, le partage d'expériences ainsi que l'élaboration et la promotion de bonnes pratiques. Il s'agit de veiller à

[48] < http://webdatanet.cbs.dk/ > [consulté le 29 octobre 2013].
[49] < http://www.nedimah.eu/ > [consulté le 29 octobre 2013].
[50] SSO : Single Sign-On.

harmoniser les initiatives nationales, européennes et internationales pour éviter de constituer de nouveaux silos de données. Cependant, comment ces infrastructures seront-elles maintenues dans le temps (financièrement, techniquement, humainement) ?

Deuxième partie

2 Les données de la recherche en ethnologie : des matériaux bruts à valoriser

2.1 Problématiques et enjeux pour la diffusion et l'exploitation des données ethnographiques

2.1.1 Des données collectées sur le terrain

2.1.1.1 Les données ethnographiques : une hétérogénéité de supports à exploiter

L'ethnologie est une discipline des SHS pour laquelle les données collectées sont, du point de vue des supports et des formats, particulièrement hétérogènes. En effet, un chercheur reviendra « du terrain » avec tout un ensemble de données qui pourront se présenter sous forme de documents papier et/ou électroniques (carnets de notes, cartes, dessins, reproductions de documents écrits, livres, etc.) et de documents audiovisuels tels que des images, des films, des enregistrements sonores. Ces documents représentent une source précieuse d'informations sur la société étudiée et sont nécessaires au travail d'analyse du chercheur. Or, la plupart du temps, ces données brutes, une fois qu'elles ont été exploitées, finissent oubliées et perdues au fond de cartons. Par la suite, ces données ne sont pas systématiquement déposées au laboratoire d'affiliation : elles sont considérées comme des archives personnelles et, à ce titre, privées. Dans le contexte numérique actuel, les projets mis en place proposent des systèmes dans lesquels le chercheur verse l'ensemble de ses données, de quelque nature qu'elles soient. Par ailleurs, les équipes en charge de ces projets développent et mettent à disposition des outils de traitement et d'analyse des données afin de donner une valeur ajoutée aux plateformes. L'objectif est de proposer dans un seul et même espace que les données soient déposées, conservées, voire archivées automatiquement et que l'on puisse documenter, traiter et analyser ces données sur une même interface. Ainsi, la plateforme devient un espace de travail au quotidien pour le chercheur, ce qui peut le convaincre d'y déposer ses archives. Pour que cela fonctionne, l'architecture qui sous-tend le système devra avoir été mûrement réfléchie afin d'offrir une interface conviviale et facile à prendre en main.

En ce qui concerne les données de terrain non nativement numérique (document papier ou analogique), il est nécessaire d'avoir recours à la numérisation avant de pouvoir les verser dans une plateforme. Ce procédé nécessite des moyens techniques et humains qui ne sont pas toujours à disposition en interne dans les laboratoires de recherche. A l'heure actuelle, le ministère de la Culture et de la Communication finance la numérisation du patrimoine culturel à partir d'appels à projet[51]. Le CNRS, quant à lui, soutient la numérisation uniquement dans le cadre de programmes de recherche. De nombreuses entreprises proposent également ce type de services[52] et les meilleures d'entre elles

[51] < http://www.culture.gouv.fr/culture/mrt/numerisation/index.html > [consulté le 22 novembre 2013].
[52] Cette question a été largement abordée dans la liste de diffusion « archives-son-audiovisuel » [31, ARCHIVES-SON-AUDIOVISUEL].

s'appuient sur les normes internationales pour fournir des données exploitables. La numérisation des données de la recherche fait partie du processus documentaire à mettre en œuvre pour alimenter les plateformes et il convient de soutenir financièrement les laboratoires de recherche qui en sont détenteurs. Le travail d'archivage et de réinterprétation des données non nativement numériques est complexe. En effet, il est particulièrement difficile de décrypter des annotations personnelles et de retranscrire de manière signifiante les informations collectées par un chercheur. Pour autant, cela ne doit pas être un frein à la numérisation car ces données représentent un fonds d'archives à mettre à disposition et à exploiter au même titre que les autres jeux de données. Des techniques de traitement automatique ou semi-automatique peuvent être, par la suite, élaborées à partir des corpus numérisés[53].

2.1.1.2 Un patrimoine culturel immatériel

Dans le domaine de l'ethnologie, les archives audiovisuelles collectées sur le terrain représentent un fonds patrimonial immatériel, tel que définit par l'UNESCO, qui relève d'une politique de sauvegarde et de valorisation promue par l'UNESCO [32, UNESCO] :

> *« On entend par "sauvegarde" les mesures visant à assurer la viabilité du patrimoine culturel immatériel, y compris l'identification, la documentation, la recherche, la préservation, la protection, la promotion, la mise en valeur, la transmission, essentiellement par l'éducation formelle et non formelle, ainsi que la revitalisation des différents aspects de ce patrimoine. »*

Ces archives réalisées in situ représentent une source d'informations précieuses et uniques sur la société étudiée. A l'heure actuelle, les plateformes de diffusion des données de la recherche répondent aux enjeux d'identification et de documentation de l'UNESCO. En effet, toutes les données déposées doivent être accompagnées de métadonnées qui fournissent un minimum d'informations sur le contexte de leur collecte. Ces métadonnées de description peuvent être : le titre (voire le titre original ou sa traduction), la date de collecte, des indications géographiques et culturelles, etc. Ces métadonnées permettent d'identifier le contenu mis en ligne. D'autres informations peuvent être ajoutées qui complètent les connaissances et/ou donnent un éclairage nouveau. La structuration de l'information permet à l'utilisateur de vérifier les données renseignées[54] et de les enrichir. Les plateformes dédiées aux données en SHS étant collaboratives, elles sont alimentées par des contributeurs. Cependant, il paraît primordial que ces informations proviennent de sources fiables. Or comment évaluer les contributions externes au laboratoire de recherche ? Une solution est la modération qui implique une vérification systématique de chacune des contributions. Il s'agit d'une tâche qui peut s'avérer chronophage et qui nécessite de désigner une ou des personnes référentes.

Les données ethnographiques constituent une source d'identité et de cohésion pour les populations qui en sont à l'origine. Elles sont également une représentation de la diversité culturelle,

[53] L'OCR ou reconnaissance optique de caractères est un outil qui convertit des documents scannés, PDF et images vers des formats modifiables et exploitables informatiquement.
[54] Mais également de rendre visible des champs non renseignés.

au cœur d'enjeux politiques et économiques nationaux de plus en plus prégnants. Or, une fois les données brutes de la recherche déposées dans une plateforme, celles-ci deviennent accessibles et peuvent être rapportées aux populations où elles ont été collectées et qui sont en droit de les réclamer. Le retour des données culturelles et patrimoniales à leur population est un enjeu important pour l'ethnologue. Cela fait partie de son devoir de les leur transmettre sous une forme brute et exploitable. Ces données brutes ethnographiques seront alors réutilisées à des fins différentes que celles de la recherche. Par ailleurs, il est tout à fait envisageable que ces populations contribuent elles-mêmes à compléter les informations, ce qui offre une nouvelle interprétation qui permet d'enrichir et de valoriser les données ethnographiques. La réutilisation des données ethnographiques et leur diffusion publique reposent sur des enjeux particuliers, liés à leur domaine d'application.

2.1.1.3 Les données ethnographiques, une diffusion publique sous condition : questions de droit et d'éthique

Une problématique complexe spécifique aux données ethnographiques est la gestion des droits d'accès. En effet, la diffusion de données, telles que les enregistrements sonores, les entretiens, les vidéos, les photographies, peut représenter un certain danger pour les personnes qui ont participé à l'enquête ethnographique, selon la teneur des informations fournies. Or c'est de la responsabilité du chercheur qui les diffuse de respecter l'anonymat des personnes et la confidentialité de certaines informations. Ainsi, il n'est pas toujours souhaitable de rendre publique les données ethnographiques sensibles. Une solution est d'anonymiser les données personnelles[55], comme les noms de personnes et de lieux, directement sur les archives. J'ai pu observer également certains chercheurs qui ne mentionnaient volontairement aucun nom dans les fiches de métadonnées. Se pose également la question de la propriété intellectuelle, du droit d'auteur et des droits d'exploitation de ces documents. Leur diffusion représente un enjeu complexe pour les chercheurs. Il est donc essentiel et impératif qu'ils définissent, au préalable, l'ensemble des utilisations qu'ils comptent faire de leurs matériaux de terrain. Cela leur permet de préparer des contrats qui stipulent les projets et objectifs de recherche ainsi que les formes de diffusion des données afin de prévoir et d'anticiper tout litige ultérieur. Cette démarche d'explicitation des projets de diffusion et d'exploitation est nécessaire aussi bien aux scientifiques qu'aux informateurs, sans oublier les centres de recherche et les institutions patrimoniales. A ce sujet, le carnet de recherche « Questions éthique et droit en SHS » met en ligne des articles qui décrivent les étapes pour constituer un environnement juridique et éthique formel et adapté aux enquêtes ethnographiques [33, GINOUVES]. Une fois les usages explicités, un contrat de cession de droit doit être rédigé et signé par le collecteur et l'informateur afin de préciser les modalités d'exploitation et de diffusion des informations collectées. Ces précautions facilitent la compréhension et la confiance entre l'informateur et le chercheur et sécurisent les exploitations des données qui seront faites a posteriori. Dans le cadre de recherches effectuées hors de France, les contrats doivent tenir compte de la législation du pays où est réalisé le terrain.

[55] L'anonymisation peut être totale ou partielle, permanente ou temporaire.

Ces contraintes sont à prendre en compte dans le cadre d'un projet de diffusion de données ethnographiques. Ce projet doit permettre une gestion des droits d'accès « sur mesure » afin que le chercheur puisse disposer des données et gérer leurs conditions d'accès selon différentes modalités (accès libre et téléchargement, accès restreint, etc.). De la même façon, il est utile de pouvoir automatiser la mise à disposition des données qui relèvent, au moment de leur dépôt, de dispositions légales particulières. Ainsi, une période d'embargo qui restreint l'accès à un document doit pouvoir être levée automatiquement grâce au paramétrage du système. Par ailleurs, la plupart des systèmes qui sous-tendent les droits d'accès fonctionnent par héritage. Ainsi, tous les items d'une collection en accès restreint hériteront d'un droit d'accès restreint. Pourtant, il peut arriver que dans certaines collections, le chercheur ait besoin de rendre disponible une partie des items. On devrait pouvoir accorder à ces items un statut d'exception dans lequel il serait possible de définir un droit d'accès différent de celui dont les items ont hérité.

Les données ethnographiques représentent donc un corpus singulier qui fait l'objet d'un consortium « Archives des ethnologues », soutenu par la TGIR-Huma-Num, qui a pour objectif de définir l'ensemble des processus de la chaîne de traitement des données collectées par les ethnologues pour garantir leur diffusion [34, ARCHIVES DES ETHNOLOGUES].

2.1.2 Mise en ligne des données ethnographiques : grille comparative de plateformes de recherche

2.1.2.1 Présentation de projets dédiés aux données de terrain

La collecte et l'analyse des données primaires[56] audiovisuelles font l'objet de projets de diffusion. Ces projets, menés par des institutions travaillant sur des disciplines diverses, diffèrent les uns des autres. Les plateformes[57] qui ont été sélectionnées ici offrent un panel représentatif des services mis à disposition des chercheurs pour l'exploitation des données de la recherche. Outre le fait que ces plateformes collectent des données audiovisuelles et ethnographiques[58], les données déposées dans l'ensemble de ces plateformes sont moissonnées par ISIDORE.

Les institutions, qui ont mis en place un projet de gestion des données audiovisuelles de la recherche et qui feront l'objet d'une étude, sont les suivantes :

- Lacito[59]et le projet du CRDO – Paris (Centre de ressources pour la description de l'oral - Paris) qui devient Cocoon pour « Collections de corpus oraux numériques ». Ce centre assure la gestion de corpus de ressources orales numérisées accompagnées d'annotations directes réalisées par des producteurs scientifiques.

[56] Ces données peuvent être nativement numériques ou avoir été numérisées.
[57] Seule exception : la base de données de la phonothèque du MMSH qui n'est pas, à proprement parlé, une plateforme de diffusion.
[58] Les données ethnographiques peuvent représenter une partie ou l'ensemble des données prises en charge.
[59] Le Lacito, Langues et civilisations à tradition orale (UMR 7107), est un laboratoire de recherche pluridisciplinaire qui se consacre à l'étude des langues de tradition orale.

- LPL[60] et le projet CRDO – Aix-en-provence (Centre de ressources pour la description de l'oral - Aix-en-Provence) qui devient SLDR, Speech and Language Data Repository. Ce centre numérique offre une plateforme de dépôt des données orales et linguistiques aux laboratoires et aux chercheurs indépendants ainsi qu'un système d'archivage pérenne.
- CREDO (Centre de recherche et de documentation sur l'Océanie). Ce centre a développé ODSAS, une plateforme d'archivage et d'annotations des données de terrain des chercheurs dans le domaine de l'ethnologie.
- Phonothèque de la MMSH, Maison méditerranéenne des sciences de l'Homme. La phonothèque gère la base de données Ganoub qui réunit des enregistrements du patrimoine sonore sur l'aire méditerranéenne.

Le projet Cocoon a vocation à rendre compte des technologies actuelles et des processus à mettre en œuvre dans le cadre d'un projet de diffusion des données de la recherche, par une approche pédagogique. En effet, un onglet « documentation » présente, de manière synthétique, la grille de services fournis par les plateformes dédiées aux ressources audiovisuelles et les processus qui les sous-tendent. Par ailleurs, les fonctionnalités de recherche sont innovantes et ergonomiques : on navigue sur une frise chronologique et/ou sur une carte géographique. La bulle d'informations qui s'ouvre, lorsque l'on pointe sur une date ou un lieu, donne un aperçu des ressources disponibles. Quant à la base de données Ganoub de la MMSH, elle fournit des outils de recherche que je qualifierais de traditionnels. En effet, la base de données repose sur des fonctions de recherche relativement complexes en regard de ce qui se fait aujourd'hui : opérateurs de comparaison, multiplicité des champs interrogeables, thésaurus, etc. Des outils facilitant la recherche, telle que l'auto-complétion, les possibilités d'affiner ou d'étendre les résultats de la recherche, n'ont pas été développés. Ce type de base de données s'adresse, plus particulièrement, aux personnes dont on suppose qu'elles maîtrisent et connaissent déjà la base Ganoub, c'est-à-dire aux documentalistes ou professionnels de l'IST dont le cœur de métier est l'exploitation des fonctions de recherche des différentes bases de données. Par contre, un carnet de recherche en ligne régulièrement mis à jour dynamise la valorisation des collections déposées à la MMSH [35, GINOUVES]. L'intérêt d'analyser ces deux projets est de mettre en valeur les atouts dont ils disposent afin de les proposer à l'ensemble de la communauté scientifique en charge de données audiovisuelles. La plateforme SLDR offre une interface riche et fournie qui peut déstabiliser l'utilisateur lors des premières visites. Cependant, elle repose sur un système complexe très bien organisé qui propose des développements extrêmement intéressants dans le cadre d'un projet de diffusion des données : téléchargement d'outils de traitement, mise en ligne de référentiels, gestion fine des droits de diffusion automatisée, etc. Un point fort de SLDR est qu'il met en ligne l'ensemble de la documentation qui donne à voir l'aboutissement des réflexions et des pratiques qui ont été mises en œuvre lors de la conception du projet. Le principal inconvénient est qu'il est difficile de retrouver une information sur le contexte de

[60] Le LPL, Laboratoire parole et langage (UMR 7309), est un laboratoire pluridisciplinaire dont les axes de recherche portent sur l'étude du langage et de la parole.

mise en place du projet[61], alors même que ces informations sont précieuses. La plateforme ODSAS a l'avantage de proposer un système de multicouches qui permet de sauvegarder l'ensemble des versions d'un même objet qui a été exploité. Ce système qui semble opérant oblige à restreindre le nombre d'utilisateurs puisqu'il nécessite un espace de stockage important. Le nombre de fichiers qui figure dans les statistiques est, de fait, biaisé, c'est-à-dire qu'il prend en compte chaque manipulation et annotation, enregistrées dans une même session, réalisées sur un même objet. Par contre, il propose des outils de manipulation de l'image qui font défaut aux autres projets. De plus, il offre la possibilité d'organiser soi-même les documents et de créer des relations entre eux selon les besoins du chercheur.

Les plateformes Cocoon et ODSAS ont en projet de s'inscrire dans le web de données en exprimant leurs métadonnées en RDF. Le projet Cocoon va encore plus loin puisqu'il travaille à l'identification de référentiels existants sur lesquels s'appuyer pour harmoniser ses données (FOAF pour les personnes, ORG pour les organisations, GeoNames pour les indications géographiques, Lexvo pour les langues).

2.1.2.2 Grille comparative des projets en gestion des données audiovisuelles

Le tableau 1 a pour objectif d'identifier précisément les services mis à disposition des chercheurs dans chaque projet. Le tableau 2 présente les modalités qui sous-tendent le versement de données dans les plateformes.

[61] Cette documentation n'est pas concernée par les fonctions de recherche qui se concentrent uniquement sur les ressources déposées.

Tableau 1: Identification des projets de diffusion des données ethnographiques de la recherche en SHS

Identification du projet	Nom de l'institution	Lacito / CRDO - Paris	LPL / CRDO – Aix-en-Provence	Centre de recherche et de documentation sur l'Océanie	Maison méditerranéenne des sciences de l'Homme
	Nom du projet	Cocoon – Collections de corpus oraux numériques	SLDR – Speech and Language Data Repository (SLDR/ORTOLANG)	ODSAS – Online Digital Sources and Annotation System	Base de données Ganoub – Phonothèque de la MMSH
	url	http://cocoon.tge-adonis.fr/exist/crdo/index.htm	http://crdo.up.univ-aix.fr/	http://www.odsas.net/	http://phonotheque.mmsh.univ-aix.fr/
	Objectifs	Gestion de corpus de ressources orales numérisées et portail de diffusion	Centre numérique de diffusion de données primaires	Plateforme de ressources numériques	Base de données
	Nature des données	Audio, vidéo, texte	Audio, vidéo, image, texte	Audio, vidéo, image, texte	Audio, vidéo
	Modalités de consultation	Streaming + téléchargement	Streaming + téléchargement	Streaming (téléchargement à venir)	Streaming (player sur poste local)
	Documents en ligne	5 657 enregistrements (4/12/2013)	345 161 documents (4/12/2013)	101 575 fichiers (4/12/2013)	6 878 documents (4/12/2013)
	Accompagnement au projet	Oui : éditorial et technique	Oui : technique, éditorial (normalisation, enrichissement)	Non	Oui : numérisation, catalogage, conservation
	Réseaux	OLAC	ORTOLANG		BnF, FAMDT, ARCADE
	Multilingue	Non	Oui	Oui	Oui
	Dépôt	Oui : identifiant OAI	Oui : identifiant PID	Oui	Oui
	Diffusion	Oui (Cocoon)	Oui (CC-IN2P3)	Oui	Oui
	Archivage	Oui (CINES) : identifiant ARK	Oui (CINES) : identifiant ARK	Oui	Non

Tableau 2 : Traitement des données avant leur versement

Nom du projet		Cocoon – Collections de corpus oraux numériques	SLDR – Speech and Language Data Repository (SLDR/ORTOLANG)	ODSAS – Online Digital Sources and Annotation System	Base de données Ganoub – Phonothèque de la MMSH
Outils en ligne		Documentation pour la gestion des ressources orales (présentation synthétique, liens externes, etc.)	Référentiels lexicaux et outils de traitement de la langue	Outils d'analyse images et vidéos (manipulation, séquençage)	Thésaurus thématique, entités nommées (auteur, collectivités)
Traitement des dépôts	Ouverture d'un compte utilisateur	Demande par mail	Demande sur formulaire	Demande par mail	
	Modération des dépôts	Oui	Oui	Non	
	Traitement par le déposant	Oui	Non (traitement en interne)	Oui	Non (traitement en interne)
	Métadonnées	Dublin Core simple	Sans restriction	Dublin Core	
	Format conseillé	Audio : WAV, FLAC Vidéo : MPEG-4, MKV Texte : XML, TXT Unicode, PDF	Sans restriction	Audio : MP3 Vidéo : MPEG-4 Image : JPEG Texte : PDF, DOC	Non renseigné
Gestion des droits d'accès	Déposant en ligne	Oui	Oui	Oui	Non
	Statut des accès	Licence Creatives Commons	Libre (Creative Commons) / Restreint : réservé (groupe d'utilisateurs ou accès individualisé)	Libre / Restreint (groupe d'utilisateurs ou accès individualisé)	Libre (en ligne ou à la phonothèque) / Restreint (consultation sur demande)

2.1.2.3 Des services et des acteurs dédiés aux projets et aux réalisations de diffusion des données de la recherche en SHS

La plupart des projets analysés ici prennent en charge la variété des supports et des formats des archives audiovisuelles collectées sur le terrain. Pour faciliter les contributions, tous les formats sont acceptés lors du dépôt ou sont traités en interne par les institutions (Ganoub – MMSH). Cependant, certains formats sont conseillés pour la diffusion, généralement des formats non compressés[62], et pour le téléchargement, pour lequel les formats compressés sont préconisés (rapidité du téléchargement).

Pour chaque projet, une première modération est mise en œuvre avant tout dépôt. Il est nécessaire de justifier de son statut scientifique et de ses projets de recherche pour obtenir un espace personnel de dépôt et de consultation. Par la suite, les dépôts sont, soit systématiquement vérifiés avant leur mise en ligne, soit ils relèvent de l'entière responsabilité du déposant (projet ODSAS). La modération de chaque dépôt implique un travail de vérification, éventuellement de re-traitement afin d'assurer leur conformité avec les spécifications liées au système, et de validation qui nécessite des ressources humaines importantes et reporte la disponibilité des données déposées.

La modération a plusieurs objectifs :

- garantir la fiabilité et l'intérêt scientifique des ressources déposées
- faciliter la visibilité et l'utilisation des données (vérification du traitement des données)

D'autres services sont également fournis : la diffusion, l'exploitation et l'archivage des données. Les équipes en charge des plateformes proposent également un accompagnement aux équipes de recherche détentrices de données audiovisuelles qui souhaitent les partager en ligne et les valoriser. En effet, la conversion des formats, le nommage des fichiers, la description des métadonnées ne sont toujours pas des tâches simples à réaliser. Ces tâches demandent du temps et des compétences techniques documentaires. Il est donc apparu important de soutenir les projets en apportant une aide et une expertise dès leur conception, pour suivre et faciliter la mise en œuvre des processus de traitement des données. L'intérêt est d'assurer l'alimentation et l'usage réguliers des plateformes. L'accompagnement technique et/ou éditorial tient donc compte des difficultés que peuvent rencontrer des équipes ou des institutions non spécialisées dans le domaine de l'information et de la documentation.

Le service d'archivage n'est pas obligatoire et est mis en place automatiquement sans que le déposant n'ait à intervenir. La plateforme de dépôt fonctionne alors comme une passerelle qui envoie les données au centre d'archivage. Ce dernier coordonne lui-même l'ensemble des traitements des données avant leur archivage (conversion des formats,

---

[62] Les formats non compressés contiennent le signal numérique brut qui assure son intégrité et permet des calculs plus fins lors des traitements.

vérification de la compatibilité avec les règles internes au système, etc.). L'archivage garantit la pérennité et la conservation des données selon les normes et standards internationaux ainsi que l'intégrité de l'information et la traçabilité des enrichissements. Intégrité signifie que l'information doit être lisible et disponible, dans le temps, sous sa forme originale, sans subir aucune altération. La traçabilité détermine d'avoir un accès à l'historique de chacune des modifications et de chaque enrichissement apportés au document original. Il s'agit donc de conserver de manière permanente l'ensemble des versions associées à un original.

Actuellement, la TGIR Huma-Num[63] pilote la grille de services qui tient compte de l'ensemble des processus à mettre en place lors de la conception de projets de diffusion des données de la recherche en SHS :

- Le service de traitement
  - accès à une grille de calcul et à des supercalculateurs issus des fermes de calcul du CC-IN2P3[64] pour le traitement des données
  - développement d'outils de traitement (logiciels de conversion, analyse et rendu interactif des données et des documents)
- Le service de diffusion :
  - hébergement de systèmes d'informations (sites, serveurs, plateformes communes de gestion et de diffusion) pour l'exploitation des données (hébergement des serveurs au CC-IN2P3)
- Le service d'archivage à long terme
  - Convention avec un centre agréé, le CINES.

Chacun des projets étudiés a ses propres points forts qu'il serait opportun d'exploiter. S'intéresser de près aux développements technologiques de la plateforme SLDR fournit des bases solides pour anticiper les besoins techniques et prévoir des fonctionnalités pertinentes, en amont du projet. ODSAS, qui a été mis en place par un chercheur, répond au plus près aux besoins d'une discipline particulière et propose des outils d'analyse pour les images et les vidéos, ce qui complète l'éventail des outils qui sont, essentiellement, orientés vers les documents sonores. Le web sémantique dispose de technologies qui semblent tout à fait appropriées pour optimiser le fonctionnement de ces plateformes. Ainsi, s'appuyer sur les référentiels utilisés dans la base de données Ganoub de la MMSH pour développer un vocabulaire commun sur les ressources orales serait utile pour harmoniser l'indexation des données. De la même façon, s'inspirer de l'interface de Cocoon, qui utilise des référentiels pour faciliter la recherche de données grâce à une représentation graphique, aiderait à mettre en place des systèmes de navigation et de recherche ergonomiques. Par ailleurs, encourager la description des ressources en RDF serait l'occasion de créer des relations entre ces jeux de données, ce qui faciliterait leur accès et garantirait une plus grande visibilité.

---

[63] La TGIR – Huma-Num soutient les projets qui ont analysés dans le cadre de ce mémoire.
[64] Centre de calcul de l'institut national de physique nucléaire et de physique des particules, <http://cc.in2p3.fr/ >, [consulté le 25 novembre 2013].

2.2 Le projet TELEMETA, une plateforme collaborative de partage d'archives sonores

2.2.1 Contexte de développement de TELEMETA

2.2.1.1 Le CREM, un centre de recherche à l'initiative d'un projet pilote soutenu par le CNRS

Le CREM est le Centre de recherche en ethnomusicologie affilié au LESC (Laboratoire d'ethnologie et de sociologie comparative – UMR 7186) du CNRS. Ce laboratoire dispose d'un fonds d'archives sonores important issu des missions de terrain de chercheurs dans le monde entier. En 1932, André Schaeffner[65], qui revient de son premier terrain en Afrique, ouvre une phonothèque au MET. La première collection est constituée des disques enregistrés lors de l'exposition coloniale de 1931. Après la rénovation du musée, la phonothèque deviendra la phonothèque du Musée de l'Homme en 1937. A la fin des années 60, un laboratoire d'analyse du son et une équipe de recherche du CNRS sont créés au sein du Musée de l'Homme. Des dépôts issus des terrains de chercheurs, des dépôts privés, des achats et des échanges avec des institutions diverses viennent alimenter et compléter le fonds de la phonothèque. Ces archives sonores se présentent sous des supports variés qui suivent l'histoire des techniques de l'enregistrement sonore : cylindres gravés, disques à gravure direct, disques 78 tours, bandes magnétiques, disques microsillons, cassettes audio, cassettes DAT, CD audio. Une partie du fonds sonore du CREM a été publiée par les éditions Musée de l'Homme-CNRS. Ces archives, constituées de 3 500 heures d'enregistrements de terrain inédits et d'environ 3 700 heures de documents publiés, représentent un fonds historique et patrimonial inestimable qui a bénéficié d'un programme de sauvegarde et de valorisation du CNRS. De fait, les supports analogiques sont soumis à l'obsolescence des machines et sont voués à une détérioration certaine qui limite toute exploitation. C'est dans le cadre du projet de méta-portail AnthropoNet [37, ANTHROPONET], qu'a été proposé, en 2008 au TGE-Adonis, le développement de TELEMETA, un outil web multimédia open source de diffusion des archives sonores à destination de la communauté scientifique et du grand public. Ce projet-pilote a vu le jour suite à la constatation qu'il n'existait sur le marché aucune application « open source » qui permette l'accès, la gestion, la conservation et la diffusion des données sonores de la recherche. Depuis 2009, Joséphine Simonnot, ingénieur au CREM, est le chef de projet TELEMETA dont le développement est assuré par la société Parisson. Différents experts ont été consultés lors de la conception de cette plateforme : Pascal Cordereix (BnF) et Véronique Ginouvès (MMSH[66]) en tant que spécialistes du

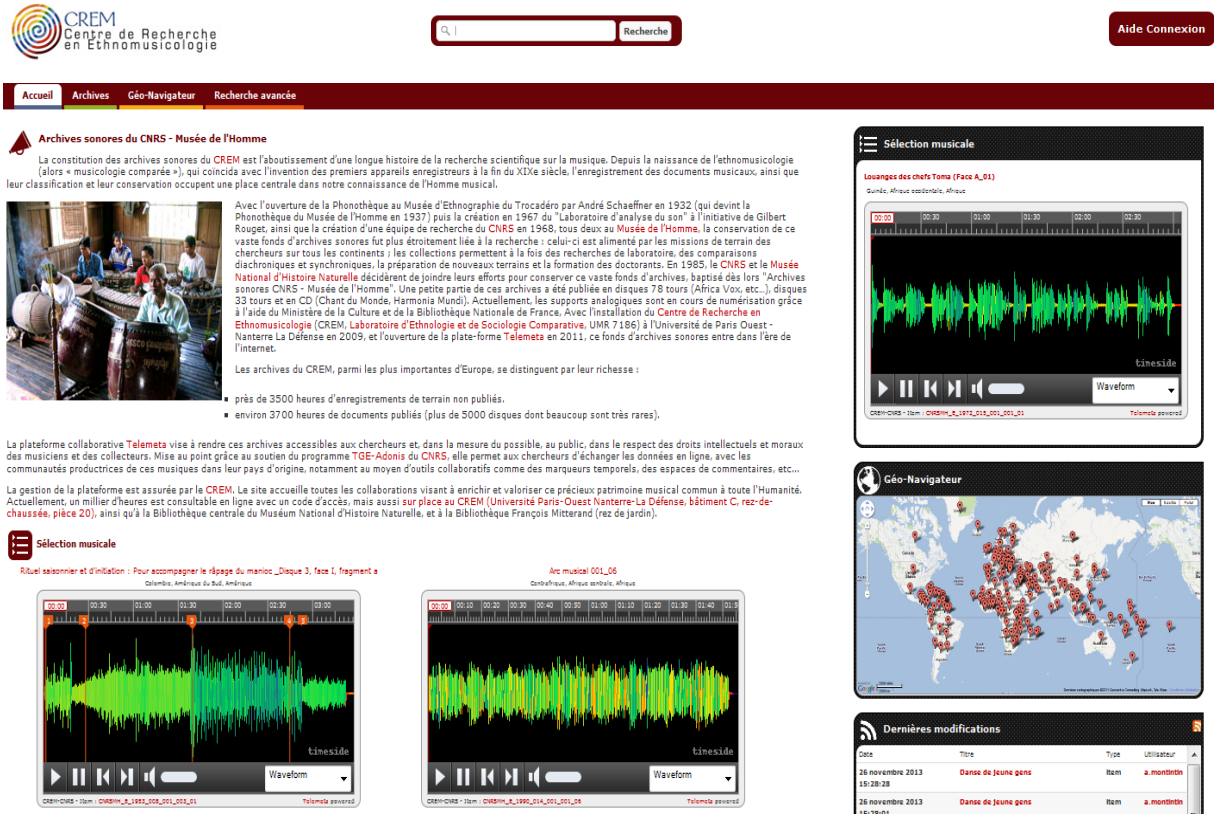
[65] André Schaeffner est un scientifique, spécialiste de l'organologie, qui créa le département d'ethnomusicologie du Musée d'Ethnographie du Trocadéro (MET), devenu Musée de l'Homme qu'il dirigea de 1929 à 1965 [36, ROUEFF].
[66] Maison méditerranéenne des sciences de l'Homme

catalogage et de l'archivage des documents numériques sonores, Jean-Marc Fontaine (LAM[67]), spécialiste de la conservation des supports d'enregistrement sonores, et Hugues Genevois (LAM), expert en perception et en représentation du sonore. Le suivi est également assuré par Aude Julien-Da Cruz Lima, responsable de la gestion et de la valorisation des archives du CREM depuis 2010.

Dès sa conception, TELEMETA s'est appuyé sur le format de description des métadonnées Dublin Core (ISO 15836) afin de permettre le moissonnage des métadonnées et d'assurer ainsi la visibilité des données mises en ligne. Le cahier des charges fonctionnel de 2009, principalement orienté vers les besoins techniques liés à l'application web, a été coordonné par le chef de projet, Joséphine Simonnot, et le développeur, Guillaume Pellerin[68]. Avant TELEMETA, les archives du CREM étaient accessibles en interne depuis le SGBD[69] 4D. La migration de la base de données a été réalisée à partir d'une étude faite par la société Samalyse qui a identifié les données prioritaires à convertir (cote, intitulés des champs, repérage des doublons, accès) et qui a également préconisé une nouvelle organisation de la structure de la base (relations entre les tables, énumérations, équivalences Dublin Core) afin d'améliorer la recherche d'informations et d'assurer des fonctions de recherche puissantes et intuitives au projet TELEMETA. Par ailleurs, pour alimenter la plateforme, l'équipe technique du CREM a mis en place des projets de numérisation des supports analogiques. Le ministère de la Culture et de la Communication a apporté son aide dans le cadre du Plan national de numérisation entre 2003 et 2008. A partir de 2008, un accord cadre entre le ministère de la Culture et le CNRS a permis de poursuivre cette campagne de numérisation des archives du CREM. Depuis 2009, le CREM collabore également avec le service audiovisuel de la BnF pour la conservation des supports inédits et du fonds de 78 tours édités sur le site de Tolbiac. A l'heure actuelle, la moitié du fonds inédit a été numérisée et mise en ligne (sur un total de 800 collections inédites), ce qui représente environ 2 téras octets.

[67] Lutheries-Acoustique-Musique de l'Institut Jean le Rond d'Alembert, LAM/IJLRA. Depuis mai 2012, le LAM est doté de sa propre plateforme TELEMETA.
[68] Guillaume Pellerin a lui-même été formé au LAM/IJLRA.
[69] Système de gestion de base de données

2.2.1.2 Un outil collaboratif de partage et d'analyse des données



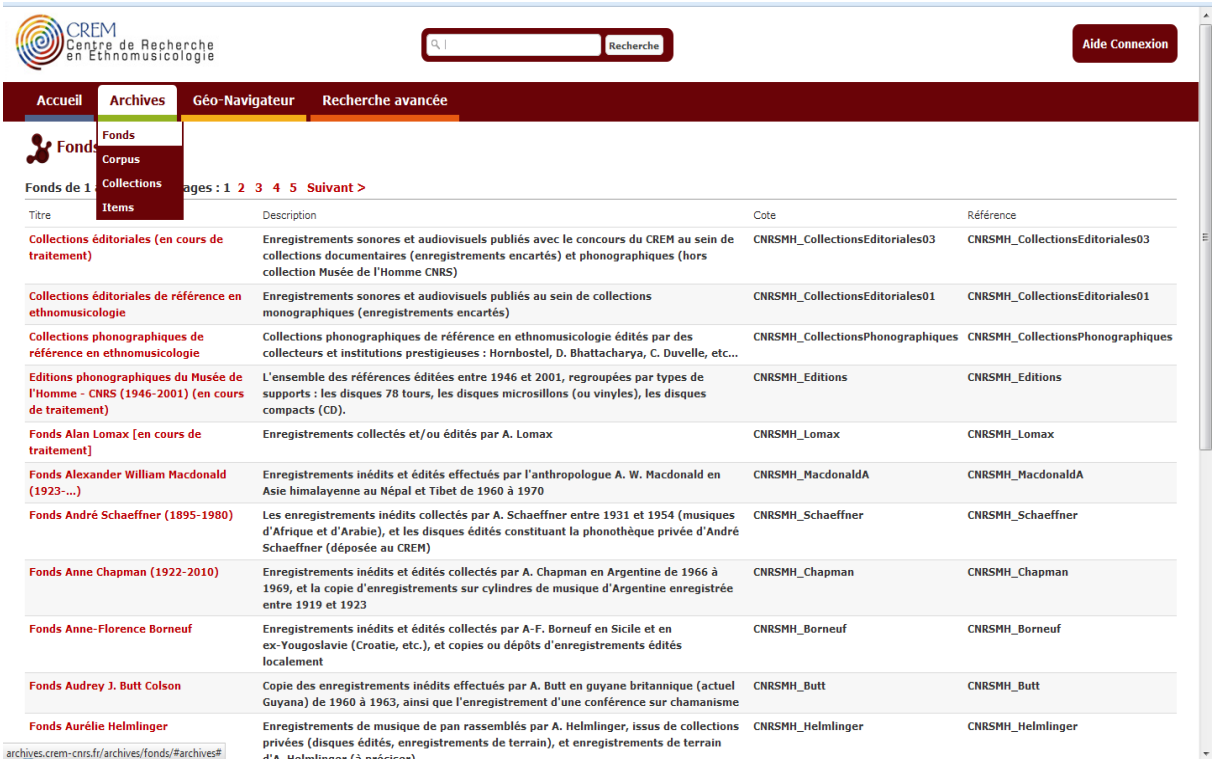
TELEMETA est un projet dont le dispositif fonctionne sur le principe d'une participation active des chercheurs. C'est pourquoi l'on parle de plateforme collaborative de travail. Les contributions reposent sur le dépôt de documents sonores collectés dans le cadre d'activités de recherche et sur une description du contexte d'enregistrement à partir de fiches à documenter. Les fiches de métadonnées se présentent sous forme de formulaires dont les champs, organisés en pavé, apportent des renseignements de natures diverses. Certains champs non renseignés n'ont pas été masqués volontairement à l'affichage afin de signaler à l'utilisateur que l'information est manquante.

Trois possibilités s'offrent à l'utilisateur pour ses besoins de recherche :

- la barre de recherche simple permet de faire une requête sur un ensemble de métadonnées de la base ;
- un formulaire de recherche avancée propose de combiner différents critères avec, pour les champs liés à une énumération (listes de mots), une auto-complétion automatique ;
- un géo-navigateur donne accès aux items en pointant sur le lieu de leur collecte. Ce géo-navigateur utilise le thésaurus GeoEthno, conçu pour l'indexation géographique dans le domaine de l'ethnologie élaboré par la bibliothèque Eric-de-Dampierre située à la MAE[70].

[70] Maison de l'archéologie et de l'ethnologie

L'architecture du site s'appuie sur 4 niveaux hiérarchiques[71] qui organisent le contenu en ligne et permettent de créer des relations « parent-children » (par exemple, il est possible de relier une collection, le « children », à plusieurs corpus). Cela permet alors d'accéder à un ensemble d'enregistrements à partir de plusieurs points d'entrées et évite le cloisonnement des collections.

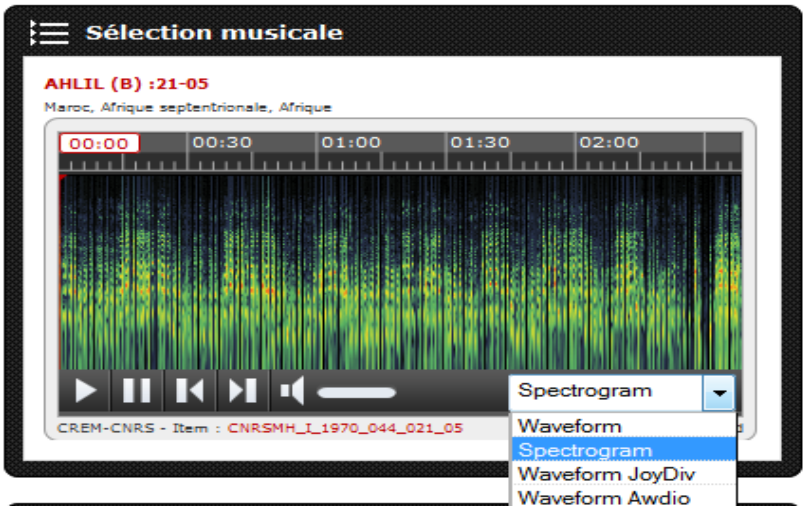


Lorsqu'un document a été défini comme libre de droit (accès « full »), la fiche de métadonnées est accessible et l'enregistrement sonore est proposé à l'écoute sur un player audio qui apparaît sur la fiche. Il existe plusieurs modalités pour l'accès restreint à un item : seules les métadonnées sont accessibles (accès « metadata »), soit l'accès est interdit à l'ensemble des données d'un item (accès « none »). Pour les statuts « metadata » et « none », une boîte de dialogue précise les conditions d'accès et la démarche à suivre pour consulter les items. Ces accès sont définis par le chercheur lui-même.

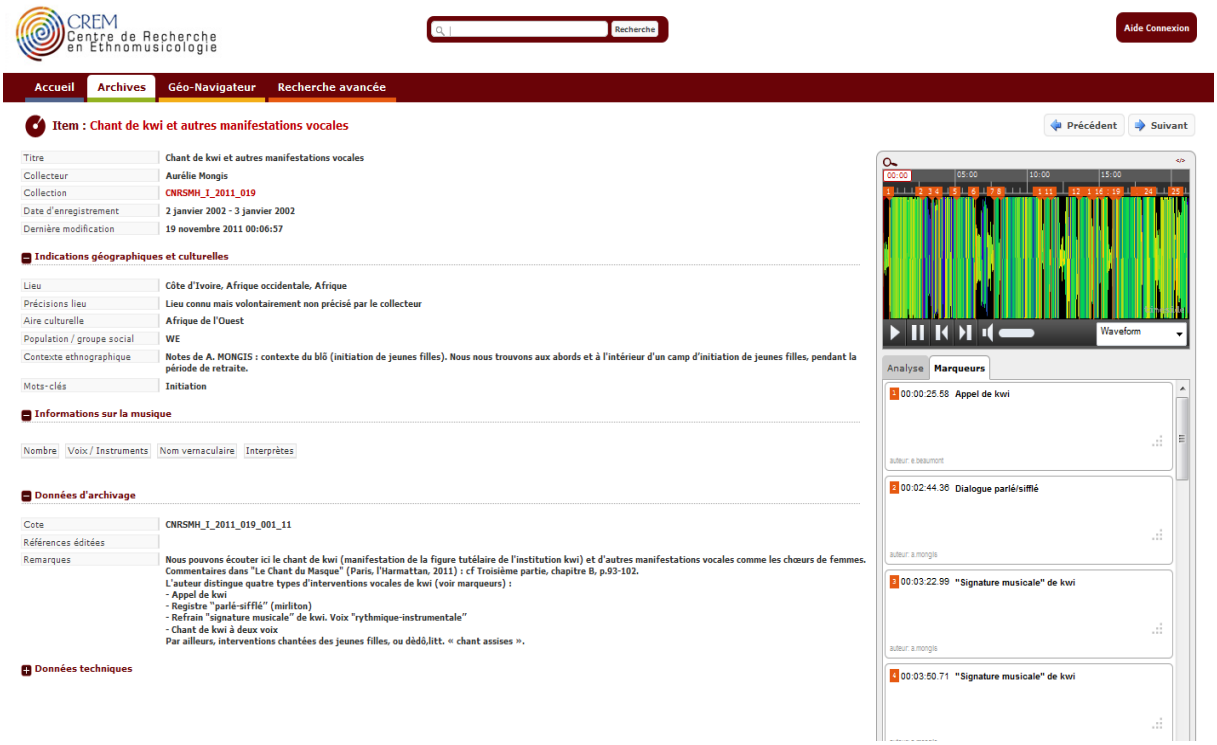
Sur la page d'accueil, trois players audio sont mis à la disposition de l'utilisateur. Ces players permettent l'écoute en ligne des archives libres de droit sélectionnées au hasard dans le fonds (au format non compressé WAV). A partir du player audio, l'accès à la fiche de

[71] Les différents objets qui décomposent le modèle de métadonnées sont : le fonds (1er niveau), le corpus (2e niveau), la collection (3e niveau) et l'item (4e niveau) qui correspond à une plage d'enregistrement.

métadonnées de l'item en écoute s'effectue sur le titre et la cote (liens hypertextes). Le player audio propose une visualisation de la forme d'onde du signal et de son spectrogramme en trois dimensions : temps, fréquence, intensité.

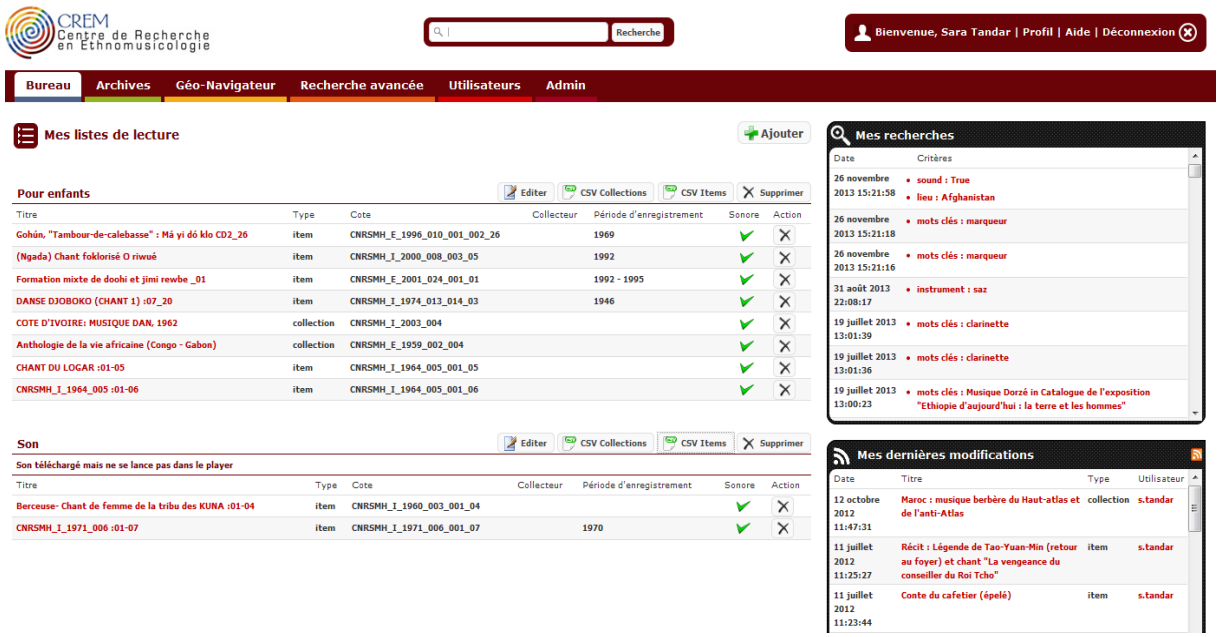


Un outil de séquençage a également été développé qui permet la pose de marqueurs temporels[72] directement sur l'onde pour annoter un continuum sonore. A ces marqueurs sont associés des espaces de commentaires qui décrivent les séquences.



[72] Cette fonctionnalité n'est accessible qu'aux utilisateurs disposant d'un compte personnel. Ces marqueurs pourront être visibles pour un simple visiteur selon les règles d'accès définies.

Cette palette d'outils innovants mis à disposition de l'utilisateur facilite le travail d'analyse des documents sonores. Cependant, il est nécessaire de disposer d'un compte personnel pour avoir accès aux outils de traitement ainsi qu'aux fonctionnalités d'édition. Le mode édition donne la possibilité de mettre en ligne des documents sonores, de remplir et/ou de compléter les fiches documentaires associées, d'y joindre des médias (images, vidéos) et des liens pointant vers des ressources externes (par exemple, un lien vers la page personnelle du chercheur). D'autres fonctionnalités viennent compléter l'espace de travail du chercheur. Ainsi, il est possible de créer des listes de lecture et de faire un export au format CSV des métadonnées des fiches collections et items.



Trois profils auxquels sont associés des niveaux d'utilisation différents ont été prévus :

- l'administrateur a accès à l'ensemble des ressources de la session TELEMETA et définit le statut de l'utilisateur qui demande l'ouverture d'un compte ;
- le documentaliste émet ou modifie le contenu d'un média et de ses métadonnées et enrichit les énumérations (listes) ;
- le chercheur émet ou modifie le contenu d'un média et de ses métadonnées. TELEMETA fournit donc un espace de travail qui donne accès à des enregistrements documentés qui peuvent être enrichis, analysés et exploités.

Cette plateforme répond à des normes d'interopérabilité des données afin d'assurer la visibilité des chercheurs et de leurs travaux, notamment par le moissonnage des données dans ISIDORE. En 2013, 25 596 fichiers[73] sont en ligne et 46 013 fiches items ont été

[73] Il s'agit essentiellement de fichiers sons. Depuis 2012, des fichiers vidéo sont également mis en ligne.

créées, ce qui équivaut à plus de 50% des items pour lesquels des fichiers son ont été joints. Depuis 2011, 170 comptes utilisateurs ont été ouverts.

2.2.2 Perspectives de développement et d'améliorations

2.2.2.1 Un projet évolutif : avantages et contraintes

Le projet TELEMETA a été conçu, dès l'origine, comme une plateforme évolutive, et non comme un produit final ou finalisé. Cela signifie qu'il est possible d'adapter le projet et de proposer des développements et des améliorations en continu afin de répondre aux besoins des utilisateurs. L'objectif principal est de disposer rapidement d'une version exploitable de la plateforme qui permette de mettre en ligne et de documenter les archives, quand bien même l'ensemble des fonctionnalités et des résultats n'est pas fixé, ni arrêté. Cette démarche favorise l'interaction entre le développeur, le concepteur et les utilisateurs et fournit les moyens de réévaluer les contraintes et les nécessités d'exploitation de l'outil.

Dans le cadre de mon activité au CREM en tant que chargée de ressources documentaires, j'ai été amenée à contribuer régulièrement à l'alimentation de la base de données. Cette expérience m'a permis de constater des problèmes qui se posaient régulièrement et qui nécessitaient d'être mis à l'ordre du jour d'un programme de développements à venir. L'étude et l'analyse de ces points de vigilance ont été réalisées en collaboration avec Aude Julien-Da Cruz Lima en 2012. Il s'agissait de lister et de décrire les problèmes rencontrés, de faire des propositions d'améliorations et d'établir un ordre des priorités dans les développements.

Les principaux points abordés ont été :

- La gestion des droits d'accès
- La traçabilité des contributions
- Le formulaire de saisie en mode édition
- La gestion des modifications
- La recherche plein texte

Les droits d'accès nécessitent d'être affinés suivant les demandes formulées par certains chercheurs. En effet, à l'origine, trois statuts définissant les droits d'accès aux ressources ont été paramétrés. Ils reposent sur un principe d'héritage[74] : « full », « metadata » et « none ». Un nouvel accès « partial » a été prévu afin que certains items libres de droit puissent devenir accessibles, indépendamment de l'accès à la collection à laquelle ils appartiennent. Par ailleurs, actuellement, le profil chercheur donne accès, sans restriction, à l'ensemble des archives mises en ligne. Ce principe peut faire tort au projet car certains chercheurs refusent de rendre accessibles leurs données, et le principe de fournir automatiquement les données à l'ensemble des profils chercheurs peut freiner leurs

[74] Ainsi, les items d'une collection en accès « metadata » héritent automatiquement de ce statut.

contributions. Par ailleurs, il s'avère que certains chercheurs souhaiteraient rendre accessible une partie de leurs données en accès restreint, à des personnes identifiées. Le chercheur devrait donc pouvoir sélectionner les données qu'il veut rendre disponibles de manière ponctuelle, ainsi que les personnes à qui il accorde cette autorisation, sans modifier le statut d'origine des items. Cette procédure serait à l'initiative du chercheur concerné. Cela implique de créer un nouveau profil d'utilisateur qui serait paramétré par le chercheur : sélection des items, des personnes, des actions autorisées[75] et du temps de disponibilité des items. Ce compte aurait alors un statut partiel et temporaire. Pour les items, dont la restriction d'accès est justifiée par des dispositions légales particulières, prévoient, aujourd'hui, des métadonnées qui précisent les attributs des droits d'accès. Ces métadonnées déterminent la date de début de la restriction[76] et calculent directement la date de fin de la restriction d'accès. Ce paramétrage, qui s'effectue dès le dépôt, automatise la mise à disposition publique du document.

Un autre point de vigilance qui me paraît essentiel est qu'il n'existe pas, à l'heure actuelle, de fonction de publication qui permette de rendre public ou privé, le séquençage réalisé avec les marqueurs. Or, il me semble utile que l'espace de travail du chercheur devienne également un environnement personnel, c'est-à-dire que lorsque le chercheur considère que l'analyse en cours n'est pas finalisée et qu'il ne désire pas la mettre à disposition, il puisse conserver sa session de travail en l'état sur un espace personnel privé.

La plateforme TELEMETA ne dispose pas encore d'un système qui assure la traçabilité des contributions, bien qu'un bloc situé sur la page d'accueil affiche les dernières modifications. Ce bloc fournit des informations uniquement sur la date de modification, le titre de la fiche modifiée et le nom de l'utilisateur qui a effectué la modification. Pour autant, il ne s'agit pas d'un historique en tant que tel. En effet, il est impossible de savoir ce qui a été précisément modifié, ou de consulter la version immédiatement antérieure à la dernière modification. La mise en place d'un historique permettrait de garantir un suivi des contributions qui éviterait que la dernière version enregistrée n'écrase la version précédente. Bien au contraire, chaque version serait sauvegardée et conservée dans le temps. Ce système anticiperait sur les développements nécessaires à la mise en place du projet d'archivage pérenne des données déposées dans TELEMETA par le CINES. Par ailleurs, pour les archives les plus anciennes qui disposent de fiches documentaires papier, cela garantirait que les métadonnées d'origine, qui représentent un contexte historique spécifique, ne soient

---

[75] Le chercheur peut avoir besoin de définir le périmètre d'action que peut effectuer l'utilisateur : simple consultation à tout ou partie des données, consultation et annotation, etc.

[76] Le SLDR utilise ce principe. Il met à disposition, sur son site, une documentation qui précise ce système de gestion et d'automatisation des droits d'accès, < http://crdo.up.univ-aix.fr/SLDRdata/doc/show/lyon-oct2012/GestionDroitsAcces.pdf > [consulté le 25 novembre 2013].

pas modifiées ou corrigées pour répondre aux évolutions terminologiques et idéologiques actuelles et futures. Bien entendu, le projet TELEMETA n'a pas pour ambition d'archiver les données déposées, mais il paraît essentiel, dans un premier temps, de sauvegarder et de conserver au moins deux versions du même objet : la première version qui a été déposée, qui sera considérée comme l'originale, et la dernière version qui a été mise à jour et enregistrée. Cette proposition concerne autant le fichier son ou vidéo que la fiche de métadonnées.

Les autres points relèvent de fonctionnalités spécifiques au back-office qui concernent, plus particulièrement, l'ergonomie du formulaire de saisie (mode édition), l'exploitation des listes et des notices d'autorité des entités nommées associées à certains champs du formulaire de saisie, ainsi que l'amélioration du moteur de recherche.

Le formulaire de saisie présente plusieurs problèmes d'ergonomie qui entravent la lisibilité des données saisies et ne facilitent pas la navigation. Concernant la lisibilité des données saisies, il s'agit de modifier les dimensions des cadres de saisie de certains champs. Il serait également envisageable de proposer une icone de visualisation qui permette de vérifier le rendu des données saisies à l'affichage de la fiche. D'autre part, la gestion des modifications n'a pas été prévue. Les opérations de recherche et de manipulation des données devraient être intégrées au projet et s'appuyer sur le modèle des SGBD[77] : modification d'un ou plusieurs champs d'une fiche sélectionnée ou sur un lot de fiches, modification ou suppression d'un terme ou d'un ensemble de termes dans un champ, etc. De la même façon, la gestion des énumérations devrait pouvoir reposer sur ce modèle. Cela est d'autant plus nécessaire que les listes actuelles ont été extraites à partir des données issues de la base de données 4D et qu'elles nécessitent d'être nettoyées et harmonisées. Les énumérations devraient également pouvoir intégrer des référentiels existants, en particulier les entités nommées, telles que les noms de personnes, les institutions, etc., référentiels qui ont pu être réalisés par d'autres centres de recherche.

Le moteur de recherche a également fait l'objet de propositions pour améliorer la pertinence des résultats. Après avoir réalisé quelques tests, j'ai pu constater que le moteur de recherche prenait en compte l'ensemble des termes de la requête, ce qui incluait également les mots vides, tels que « à », « les », etc. Cela occasionnait un bruit conséquent des résultats de recherche. Il a donc été envisagé d'intégrer une liste de mots vides, en téléchargement libre sur Internet. De même, il a été proposé d'élargir la recherche à l'ensemble des champs et des tables liées de la base de données. L'inconvénient est le risque de ralentissement de la recherche. C'est pourquoi la création et la mise à jour

[77] Système de gestion de base de données qui exprime les opérations de recherche et de manipulation des données sous forme de requêtes, par exemple SQL qui est une technologie utilisée dans TELEMETA.

régulière d'index associés aux champs interrogeables sont indispensables. Ces deux développements sont indissociables pour garantir une recherche plein texte performante. De plus, une extraction des données automatisée pourrait permettre de compléter automatiquement les énumérations.

2.2.2.2 Le projet TELEMETA LAM

Depuis mai 2012, la plateforme TELEMETA gérée par le LAM[78] est en ligne. J'ai participé, en 2012, à l'élaboration du plan de classement actuellement en ligne. Ce plan de classement a été réalisé à partir d'un fichier d'inventaire, réalisé en 2006, du fonds d'archives du LAM. Ce fonds est constitué d'environ 3 000 supports d'enregistrements divers[79] qui représentent l'histoire des recherches en acoustique musicale effectuées au LAM. Malheureusement, ces supports sont fragiles à manipuler et nécessitent d'être numérisés pour assurer leur sauvegarde et les mettre à disposition des chercheurs et du public. La numérisation, réalisée en interne, est en cours et aujourd'hui, TELEMETA a mis en ligne environ 230 archives[80] et leur documentation associée. Dans le cadre du projet DIADEMS, soutenu par l'ANR[81], un réseau de compétences s'est constitué afin d'améliorer l'accès au fonds d'archives sonores et son indexation pour faciliter le repérage et l'accès au contenu. La mission du projet DIADEMS est de développer des outils innovants pour automatiser, totalement ou en partie, l'indexation des enregistrements dont la longueur, le contenu et la qualité sont variables d'un support à l'autre. La collaboration entre des équipes de recherche axées sur les technologies[82] et des équipes spécialisées dans le domaine de l'ethnomusicologie et de la musicologie[83] assure une complémentarité des compétences au service des besoins des chercheurs. La représentation ou visualisation des résultats d'analyse fait également l'objet d'un projet de développement. Ces outils nécessitent d'entamer une réflexion sur l'architecture sous-tendant le système TELEMETA. En effet, les résultats d'analyse devront pouvoir être réutilisés par les chercheurs, ce qui implique qu'ils aient accès aux différentes versions d'un même objet. De fait, un objet original qui sert de support à une analyse devra être récupérable sous sa forme originelle, ainsi que sous les diverses formes qu'il prendra après les analyses successives qui seront réalisées. Comment accéder aux différentes versions de l'objet ? Comment restreindre l'accès à une version en

---

[78] Le laboratoire du LAM contribue à la conception de la plateforme depuis 2008.
[79] Dans la grande majorité, il s'agit de bandes magnétiques, mais on trouve également des cassettes audio, des cassettes DAT, des disquettes et des CD audio.
[80] Au LAM, un item correspond à un support, à la différence du CREM où un item correspond à une plage.
[81] Agence nationale de recherche
[82] Les laboratoires de linguistique et d'acoustique IRIT, LIMSI et LAM fournissent un environnement et un savoir-faire technologique.
[83] Les laboratoires du CREM-LESC et du Musée National d'Histoire Naturelle fournissent un corpus de ressources documentaires qui sera utilisé pour tester les outils.

cours d'analyse sans qu'elle n'hérite automatiquement du droit d'accès défini pour l'objet original ? Ces questions, en cours de réflexion, sont fondamentales pour assurer une exploitation optimale des analyses réalisées à partir des outils développés.

Par ailleurs, les améliorations à apporter au formulaire d'édition et aux fonctions de recherche, détaillées pour le projet TELEMETA géré par le CREM, sont également à mettre en œuvre dans le TELEMETA du LAM. En effet, il me paraît nécessaire de se concentrer, en priorité, sur les points évoqués dans le chapitre précédent, avant même de s'engager totalement dans le développement d'outils de traitement. De fait, si la plateforme ne répond pas à des spécifications basiques de recherche et d'édition qui sont à la source du travail de documentation et de mise en ligne des données, il existe un risque bien réel que les chercheurs n'utilisent pas la plateforme. Cependant, ces développements nécessitent un budget qui n'est pas toujours disponible, alors que des crédits sont accordés pour développer des outils d'analyse. C'est là, à mon avis, un problème récurrent pour ce type de projets : obtenir des financements pour la numérisation de documents audiovisuels ou pour l'amélioration de fonctionnalités basiques est plus complexe que de monter de nouveaux projets innovants qui facilitent l'exploitation des données. Or, une chose est sûre, si les données ne sont pas versées sur les plateformes, les outils ne pourront pas être utilisés.

2.2.3 Conclusion

A travers l'étude et l'analyse de différents projets dont le point commun est la diffusion des données audiovisuelles, il apparaît clairement que, pour un même type de jeux de données, les systèmes qui sous-tendent les plateformes diffèrent d'un projet à un autre mais présentent chacun des points forts qu'il serait pertinent de mettre à profit. La principale difficulté est de concevoir une plateforme qui réunit une interface lisible et ergonomique, un mode d'édition facile à prendre en main, un moteur de recherche puissant et efficace, des outils de traitement avancés et innovants et une gestion fine des droits d'accès. De fait, l'objectif est de créer un environnement au sein duquel le chercheur soit le plus autonome possible. Mais cette autonomie n'est pas un but simple à atteindre. A l'heure actuelle, le web sémantique fournit des technologies qui devraient contribuer à rendre plus performantes les fonctions de recherche et à l'éclatement des silos de données. Pour autant, le web sémantique ne peut pas tout et, il reste nécessaire, avant toute réalisation, de tenir compte de points de vigilance qui, s'ils ne sont pas anticipés en amont du projet, pourraient devenir des contraintes difficiles à régler sur le court et moyen terme. Bien entendu, il ne s'agit pas de monter un projet trop ambitieux qui nécessiterait alors des moyens et des ressources conséquentes, mais de comprendre les besoins de l'utilisateur final, mais aussi et surtout, de prendre en considération la réalité du contexte de travail du chercheur. En effet, les avantages de l'Open Access sont évidents et encouragés par la communauté scientifique, mais les problématiques liées à la mise à disposition des données sont peu mises en avant.

De fait, choisir la plateforme qui réponde le mieux à ses besoins, y déposer ses données et éventuellement les traiter en amont, les documenter, cela n'est pas toujours considéré comme des tâches qui relèvent des activités de recherche. Ces processus sont chronophages et peuvent se révéler relativement complexes. L'accompagnement est nécessaire mais s'agit-il uniquement d'accompagner ou de faire à la place de ? L'exemple du TELEMETA du CREM donne à réfléchir. Ce site qui a vocation à partager les données et à fournir les moyens de les exploiter, devient, avant tout, une simple plateforme de consultation. Seuls deux chercheurs contribuent régulièrement et utilisent les outils de travail mis à disposition. Comment se prémunir de ces inconvénients ? Quels sont les points de vigilance lors de la mise en place d'un projet de diffusion des données ? Comment impliquer les chercheurs et les encourager à mettre en ligne leurs ressources ? Ces points seront abordés dans la dernière partie.

Troisième partie

3 Recommandations pour la mise en place d'un projet de diffusion et de gestion des données collectées par les chercheurs en SHS

3.1 Etude et conception du projet de diffusion des données

Je me propose d'illustrer mes propos en prenant l'exemple du projet de plateforme TELEMETA, mis en place dans le laboratoire du LAM et auquel je participe de manière active. Je suis actuellement en charge de l'organisation des archives sonores du LAM et de leur mise en ligne. Je m'appuierai également sur les observations que j'ai effectuées dans le cadre du projet TELEMETA géré par le CREM, pour aborder des points particuliers.

3.1.1 Définition du périmètre du projet

Pour toute institution productrice de données sonores et audiovisuelles, il est nécessaire de définir le périmètre du projet afin de déterminer les problèmes et contraintes rencontrés dans l'environnement de travail et les avantages qui résulteraient à mettre en place un projet de diffusion et de valorisation du fonds. Le laboratoire du LAM dispose d'environ 3 000 supports physiques d'enregistrements sonores, disséminés dans ses locaux. Il s'agit essentiellement de bandes magnétiques[84] fragiles à manipuler. Par ailleurs, ces supports ne sont pas conservés dans des conditions de conditionnement et de stockage favorables à leur préservation et sont donc menacés de dégradation. Au sein du LAM, un expert assure le maintien en condition opérationnelle des équipements de lecture. Mais la fragilité des supports ainsi que l'usure des équipements de lecture impliquent que soient effectuées au plus vite des copies numériques de consultation. Dans le cadre du LAM, la mise en place d'une plateforme de diffusion du fonds d'archives était donc devenue nécessaire : il fallait pouvoir mettre à la disposition de la communauté scientifique un fonds d'archives sonores précieux et unique sur l'étude des instruments de musique et l'étude de la perception.

Le laboratoire du LAM n'ayant jamais eu à sa disposition de documentaliste, le fonds d'archives a été pris en charge par un chercheur étroitement associé aux activités du laboratoire depuis sa création, Michèle Castellengo. Selon les études et les analyses qui ont encadré les enregistrements, certains supports possèdent un code de couleur et de forme directement apposé sur les boîtiers. Ce code reste obscur pour l'ensemble des chercheurs du LAM, à l'exception de Mme Castellengo qui en est l'auteur. L'organisation actuelle du fonds

[84] Il existe également des cassettes audio, cassettes DAT, disquettes et CD audio.

est donc étroitement liée à ce chercheur, qui est également le principal interlocuteur à disposer d'informations sur les contextes d'enregistrement des archives. Ce chercheur représente donc la mémoire du fonds d'archives sonores du LAM et il convient de profiter de sa présence et de sa disponibilité pour documenter les archives.

L'évaluation des ressources internes est également à envisager à cette étape du projet. Le chef de projet TELEMETA collabore au projet TELEMETA du LAM. Une équipe de chercheurs du LAM participe à sa mise en œuvre, coordonnée par le directeur du laboratoire qui gère également le budget. Quant aux profils des chercheurs associés au projet, ils ont été définis selon des besoins spécifiques liés à la documentation des archives et au développement d'outils d'analyse. Au LAM, la numérisation du fonds sonore [38, THERON] est réalisée sur place dans le cadre d'un projet plus global de traitement et de préservation des supports analogiques. Par ailleurs, ce laboratoire a l'avantage de disposer, dans ses locaux, d'un ingénieur en charge de la maintenance informatique. Celui-ci a été impliqué dans le projet en tant que gestionnaire du système. Il est également prévu qu'il gère les profils utilisateurs et leurs droits d'accès. En ce qui concerne la documentation et l'organisation des archives, les moyens étant limités, il a été décidé de recruter temporairement une chargée de ressources documentaires.

3.1.2 Etat des lieux du fonds documentaire

Après les constatations qui ont permis d'évaluer l'urgence à mettre en place ce projet au LAM, un état des lieux s'impose pour identifier le fonds documentaire. Existe-t-il un catalogue ? Un inventaire a-t-il été récemment réalisé ? Comment les supports sont-ils référencés ? Existe-t-il un système de cote ? Est-il nécessaire de normaliser le système d'identification ? Quelles sont les informations à disposition sur les contextes d'enregistrement ? Sont-elles suffisantes pour décrire le contenu ? Pendant quelle durée faut-il conserver ces documents ?

Au sein du LAM, un fichier d'inventaire au format .xls avait été réalisé par un doctorant en 2006. Ce fichier propose un classement thématique des archives qui a été remanié afin d'élaborer, en concertation avec le maître d'ouvrage, un plan de classement définitif qui organise le contenu mis en ligne sur TELEMETA à partir des objets d'étude en cours au sein du laboratoire. Les étapes suivantes ont été le récolement systématique des archives et la création et la normalisation d'une règle de nommage des supports[85] à faire apparaître sous forme de cote sur le support lui-même ainsi que sur son boîtier. Par la suite,

[85] La cotation se présente sous la forme AA 000, c'est-à-dire deux lettres qui précisent le type de support dont il s'agit et trois chiffres qui correspondent au numéro de ligne où le support apparaît dans le fichier .xls d'inventaire. Pour une question de temps et de ressources, il a été choisi de partir du fichier d'inventaire qui a été, par la suite, réorganisé pour en faciliter la lecture.

en prévision de la numérisation des supports analogiques, des lots ont été constitués sur la base des critères de priorités définis avec l'équipe de recherche[86]. Ces lots se décomposent comme suit :

- archives présentant des problèmes techniques spécifiques et nécessitant un traitement particulier avant leur numérisation ;
- archives pouvant être numérisées sans traitement ;
- archives à ne pas numériser (copie de sauvegarde quand il existe un original, copie d'enregistrements édités, etc.).

Ces lots ont ensuite été stockés selon leur thématique. L'ensemble des informations liées à leur localisation et à leur statut (numérisé, en cours de numérisation, etc.) fait l'objet d'un document mis en ligne sur un espace de travail collaboratif.

Les informations[87] qui contextualisent chaque enregistrement se présentent sur des supports divers. Elles peuvent être inscrites directement sur le boîtier du support, soit sur le support lui-même ou encore sur une feuille volante jointe au support. Certains renseignements ont également été reportés sur des carnets de note manuscrits. Par ailleurs, des documents annexes, tels que des photographies, des livrets, des sonagrammes, etc., existent. L'ensemble de ces documents représente une documentation qu'il s'agira d'associer au fichier son et il est donc nécessaire de prévoir également leur traitement et leur numérisation. A l'heure actuelle, le projet s'appuie uniquement sur le fonds existant.

Cet état des lieux doit permettre :

- d'identifier la typologie des documents sonores : identification des supports d'enregistrement (analogique et/ou numérique) et leur volumétrie ;
- de lister les indications minimales sur les fichiers numériques (format, durée, niveau de compression, taille, etc.) ;
- d'énumérer les supports et les formats de la documentation associée aux enregistrements ;
- de regrouper les informations à disposition pour un même enregistrement ;
- de créer des lots d'archives à mettre en ligne et de définir les priorités ;
- d'élaborer un plan de classement ;
- de réorganiser le fonds en vue du projet de numérisation et de sa conservation.

3.1.3 Les contraintes de diffusion des données sonores

Les questions d'accès aux documents sonores mis en ligne font partie des principales contraintes à avoir à l'esprit lorsque l'on met en place un projet de valorisation. En effet, les modalités de diffusion et d'utilisation doivent avoir été dûment stipulées par le collecteur dans le cadre d'un contrat signé par l'ensemble des participants. Ce contrat doit

[86] Les priorités ont été définies selon trois axes : bonne qualité sonore, support fragile menacé de dégradation, enregistrements disposant d'une documentation scientifique donnant une valeur ajouté au fichier en ligne.

[87] A minima, un support a un titre qui renseigne sur l'objet d'étude de l'enregistrement. Le plus souvent, sont renseignés les dates d'enregistrement ainsi que les noms des intervenants (musiciens, interprètes, équipe de recherche, etc.).

être exigé lors du dépôt des données puisqu'il va faciliter leur valorisation. Pour les archives qui n'ont pas fait l'objet de contrat ou pour lesquelles aucune information sur les questions juridiques n'a été explicitée, il s'agira d'établir en interne des règles précises de recherche des ayants droits. Cette démarche est longue et coûteuse mais est obligatoire si l'on veut pouvoir mettre à disposition ces archives. Depuis 2011, un groupe de travail se penche sur la rédaction d'un guide de bonnes pratiques pour la diffusion électronique des données en Sciences humaines et sociales[88].

Une fois les questions de droit et d'éthique réglées, il est important de définir de manière précise et détaillée des groupes d'utilisateurs et les profils associés à chaque groupe. Ces profils vont déterminer le périmètre des actions que peut effectuer chacun de ces groupes. Pour un profil, des droits étendus ou restrictifs seront accordés concernant aussi bien l'accessibilité au contenu (front-office) qu'aux droits d'accès à la plateforme (back-office). Il s'agit d'établir qui peut intervenir sur quoi et quelles sont les actions possibles. En parallèle, les droits d'accès aux fichiers en ligne sont à étudier. Ainsi, un document libre de droit doit avoir un statut qui inclut un système de paramétrages qui accorde l'autorisation à une diffusion publique sans restriction. Les droits d'accès devraient être au maximum automatisés afin d'éviter des interventions humaines ponctuelles. En effet, si aucun système ne permet d'alerter qu'une donnée est entrée dans le domaine public, le risque principal est que le producteur de données oublie de la rendre publique. Il s'agit donc d'anticiper sur les développements techniques à mettre en œuvre lors de la conception du projet. Ainsi, pour les données devenues publiques à partir d'une période déterminée, il faudrait pouvoir rendre automatique leur diffusion, comme sur le modèle de la plateforme SLDR. Des métadonnées supplémentaires dans lesquelles sont renseignées les dates et la période d'inaccessibilité seront alors également à prévoir.

3.2 Le traitement documentaire

3.2.1 Les documents sonores : normalisation des fichiers son

Les fichiers numériques mis en ligne doivent être, au préalable, traités de manière à disposer de fichiers homogènes. Des règles de nommage à l'initiative du producteur de données sont à établir avant la réalisation du projet et doivent être précisées aux contributeurs. Ces règles vont permettre d'identifier les données déposées et de les retrouver facilement. Dans le cadre du projet TELEMETA, le nom du fichier son correspond à la cote de la fiche documentaire à laquelle il est associé. Il est conseillé de proposer des outils qui permettent d'automatiser le renommage des fichiers sur un ensemble de données, ce qui facilite le travail. Par ailleurs, définir les formats des fichiers est une étape cruciale. En

[88] < http://phonotheque.hypotheses.org/5218 > [consulté le 2 décembre 2013].

effet, se pose la question de leur obsolescence. C'est pourquoi il importe de privilégier la conversion des fichiers dans des formats stables, interopérables, libres et ouverts. Les formats de diffusion sont à définir selon les usages prévus dans le cadre du projet. Ainsi, il peut être nécessaire d'envisager plusieurs formats :

- un format de conservation ;
- un format de consultation ;
- un format pour le téléchargement ;
- un format d'archivage [39, TGE-Adonis].

Le format de consultation est, généralement, un format sans compression qui restitue l'intégrité du signal à l'écoute. Le format de téléchargement est, par contre, compressé afin de permettre un téléchargement rapide. Une fois ces choix effectués, il est important de rédiger un guide à l'attention des contributeurs qui indique les règles de nommage et les formats des fichiers conseillés pour leur mise en ligne.

3.2.2 Structuration des métadonnées

Le formulaire de métadonnées contient un ensemble d'informations aussi bien sur le contenu que sur les données techniques du fichier. Il est donc nécessaire d'organiser ce formulaire de façon à anticiper sur les usages qui en seront faits :

- les informations sur le contenu et leur degré d'importance (obligatoire ou facultatif) ;
- les informations techniques et leur degré d'importance ;
- les champs interrogeables ;
- les champs qui seront moissonnés.

Certains champs pourront être liés à des listes qui serviront à harmoniser leur contenu. Dans le cadre du TELEMETA du LAM, ces listes ont permis de préciser des informations techniques. Par exemple, une liste des supports a été réalisée. Ainsi, lorsque l'on renseigne le type de support, un certain nombre de champs associés à ce support en particulier et donnant des indications spécifiques sont à renseigner. Des référentiels servant à l'indexation des documents peuvent être également mis en relation avec certains champs.

Les formats de description des métadonnées doivent être libres et interopérables et reposer sur des standards internationaux. A minima, le schéma Dublin Core simple garantit l'échange des données. Le Dublin Core est composé de quinze descripteurs génériques qui permettent de décrire, de façon large, le contenu, les droits de propriété intellectuelle et l'instanciation. Pour affiner la description du document, le Dublin Core qualifié offre des éléments supplémentaires, des qualifieurs, qui précisent l'acception. D'autres formats de description, selon le type de données, sont détaillés dans le Guide des bonnes pratiques numériques fourni par le TGE-Adonis [40, BURNARD].

3.2.3 Indexation des fichiers en ligne pour une recherche efficace

Prévoir un champ d'indexation libre repose sur un principe qui est largement plébiscité par les chercheurs. En effet, cela leur permet d'organiser les données en ligne et de les décrire avec des termes appropriés qui relèvent d'une connaissance approfondie du terrain et du domaine de recherche. Il pourrait être envisagé de réutiliser ces termes d'indexation libre sous forme de nuages de tags, sur le modèle du web 2.0. Cette forme d'indexation n'empêche pas d'avoir un champ d'indexation normalisée relié à des référentiels propres à la discipline. Actuellement, le CREM élabore, en collaboration avec le musée du Quai Branly et la BnF, un référentiel commun pour l'ethnomusicologie. La première étape du projet porte sur la constitution d'un thésaurus sur les instruments de musique extra-européens[89]. Ce référentiel harmonisera l'indexation des informations sur la musique et aboutira à la mise en place d'une recherche fédérée entre ces trois institutions, détentrices de données ethnomusicologiques. De plus, les référentiels reliés aux tables de la base de données permettent de faciliter la recherche en proposant une aide à la saisie grâce à l'auto-complétion, par exemple.

Les fonctionnalités de recherche doivent être mûrement réfléchies. En effet, si un utilisateur se retrouve confronté à du bruit ou du silence lorsqu'il fait une recherche, il y a un fort risque qu'il décide de ne plus consulter les données, ou de ne pas contribuer à l'alimentation de la plateforme. L'indexation est une étape primordiale pour améliorer les fonctionnalités de recherche. Cependant, la réflexion doit porter également sur les champs qui seront interrogeables en cas de recherche simple, à partir d'un formulaire de recherche avancée ou encore en back-office afin de réaliser des modifications sur un lot de fiches ou de fichiers. Actuellement, c'est une des faiblesses de la plateforme TELEMETA : elle ne dispose pas d'un outil puissant de recherche qui faciliterait la recherche en front-office et le travail de gestion de la base de données en back-office. Un autre point à aborder est la configuration des résultats affichés. De fait, lorsqu'on lance une requête dans TELEMETA, les termes de la requête n'apparaissent pas à l'affichage de la liste des résultats. Une option intéressante serait d'envisager d'afficher la liste des résultats avec, pour chaque résultat, un extrait d'un des champs qui contient le terme de la requête qui serait lui-même surligné. Cela garantirait à l'utilisateur que les résultats de la recherche correspondent bien à sa requête initiale. Actuellement, les listes de résultats peuvent être affinées aux documents pour lesquels un fichier son est en ligne, bien qu'il n'existe aucun moyen de savoir si ces documents sont consultables ou non, sauf en cliquant dessus. Il pourrait être opportun de

[89] A terme, il s'agira de proposer deux autres thésaurus sur la voix et la danse afin de constituer un référentiel comprenant l'ensemble des axes de recherche qui ont cours en ethnomusicologie.

proposer un système de couleur sur l'encoche actuelle qui désigne les documents qui ont un fichier son en ligne. Ainsi, l'encoche verte indiquerait que le document est consultable et libre de droit, une encoche orange (par exemple) indiquerait que l'accès aux métadonnées est libre mais l'accès au fichier son est restreint. Une encoche rouge stipulerait que l'accès à l'ensemble des données est restreint. Par ailleurs, il serait judicieux de permettre à l'utilisateur de trier la liste des résultats[90] selon un ordre de pertinence, des dates d'enregistrement, etc.

3.3 Elaboration du site

3.3.1 Structuration de l'information du site

Un point à ne pas négliger est de travailler sur l'architecture de l'information du contenu des pages web et du site proprement dit, afin d'organiser l'ensemble des contenus et de réaliser le parcours des internautes dans le site. Structurer l'information permet de définir le système de navigation sur le site, en adéquation avec les besoins du public cible. Dans un premier temps, il faut lister toutes les informations que l'on souhaite mettre en ligne et les classer selon leurs typologies. Puis il faut organiser les informations en catégories et, enfin, les hiérarchiser par ordre d'importance sous forme de rubriques et sous-rubriques jusqu'à atteindre l'information elle-même. Cette hiérarchisation permet d'évaluer le nombre de niveaux qu'il faudra prévoir dans l'arborescence du site. A cette étape, les rubriques, qui seront utilisées dans la barre de navigation, feront l'objet d'une réflexion sur leurs libellés qui doivent être simples, logiques et limpides pour l'internaute. A l'issue de ce travail, on disposera d'une arborescence qui structurera le contenu et reliera les pages entre elles. Par ailleurs, définir la ligne éditoriale est un élément nécessaire à prévoir et à intégrer dans cette arborescence, avant de réaliser le site. Ensuite, élaborer le système de navigation devra permettre d'établir comment l'internaute se déplacera d'une page à une autre et comment il se repérera dans le site. Cette première étape réalisée, il faudra s'intéresser à la maquette conceptuelle des pages du site. Il s'agira alors d'organiser le contenu des pages de manière à homogénéiser la lecture et d'obtenir une cohérence visuelle. Pour concevoir la mise en page, on élaborera, en premier lieu, le zoning qui donnera une première idée de la configuration de la page sous forme de blocs. Puis, on créera le gabarit d'une page type qui en fournira le squelette graphique. Une charte graphique devra également être précisée.

[90] Actuellement, l'ordre d'affichage repose sur la cote du fichier, ce qui n'est pas nécessairement un tri pertinent pour l'utilisateur.

3.3.2 Configuration d'accès aux différentes versions d'un même objet

Diffuser les données brutes de la recherche implique que les données soient exploitées par le chercheur, c'est-à-dire réutilisées dans le cadre d'analyses dont les résultats serviront à des études théoriques. Or, à partir d'une donnée brute, plusieurs analyses successives peuvent être effectuées par des chercheurs différents. Comment accéder à la donnée originelle lorsqu'elle a été exploitée plusieurs fois ? De la même façon, une analyse à partir d'une version X doit pouvoir être réalisée. Comment accéder à une version spécifique de l'objet qui a subi un ou plusieurs traitements et que le chercheur souhaiterait réutiliser ? La configuration du système d'accès aux différentes versions d'un même objet doit donc faire l'objet d'une réflexion en amont. Il s'agit d'élaborer un modèle de récupération des données sur la base d'une arborescence qui stipule les chemins d'accès aux versions d'un même objet. La racine serait l'objet dans sa forme originale et les ramifications donneraient accès aux différentes versions issues des divers traitements. Cette configuration permettrait de prévoir différents usages pour un même objet et d'organiser l'accès aux différentes formes que cet objet prendra au cours des analyses, ce qui faciliterait leur accessibilité pour les utilisateurs.

3.3.3 Mise en place d'indicateurs

La mise en place d'indicateurs quantitatifs et qualitatifs permet d'obtenir des statistiques de consultation du site ainsi que de rendre compte de la réalité des usages de la plateforme par les utilisateurs. Ces indicateurs doivent permettre de suivre le parcours d'un utilisateur au sein de la plateforme. Qui consulte le site : un visiteur occasionnel ou un utilisateur disposant d'un compte ? Quelles pages sont visitées ? Combien de temps l'internaute reste-t-il sur chaque page ? Quel est le temps global de connexion au site ? Quel est le nombre de connexions quotidiennes ? Ces statistiques fournissent des indications d'ordre quantitatif sur la consultation et la navigation dans le site. Cependant, d'autres indicateurs de type qualitatif seraient pertinents à mettre en place. A cet effet, décortiquer les processus d'activités réalisées au sein de la plateforme paraît être la première étape. Par exemple, sur la plateforme TELEMETA, plusieurs tâches peuvent être décomposées :

- En mode édition : édition d'un formulaire de métadonnées, chargement du fichier son, chargement de médias associés ;
- En mode recherche : type de requête (sur des instruments de musique, sur des dates, sur des aires géographiques, etc.), pertinence des résultats ;
- Analyse du fichier son : pose de marqueur, annotations ;
- En mode consultation : lecture du fichier son, ouverture des médias associés.

L'identification des différentes tâches fournit une base pour établir des indicateurs précis qui permettent d'évaluer les usages des utilisateurs de la plateforme par rapport à ce qui avait

été envisagé à sa conception. Ces indicateurs doivent être croisés pour obtenir une analyse fine et sont à tester pour vérifier leur pertinence. Ils servent de support pour une réflexion sur l'adaptation du produit et son évolutivité. Par ailleurs, des indicateurs de mesure de satisfaction sont également à envisager. L'interprétation des résultats permet d'évaluer le projet au regard de ses objectifs de départ et des moyens investis pour les atteindre.

3.4 Stratégie de communication

3.4.1 Accompagnement au changement

Avant même la réalisation effective du projet, c'est-à-dire dès sa conception, impliquer les futurs utilisateurs peut être un excellent moyen de les familiariser avec l'idée d'un nouvel environnement de travail. Ainsi, il s'agit de s'adresser directement à eux pour définir leurs attentes et leurs besoins, leur exposer les étapes d'élaboration du projet et les informer régulièrement de son état d'avancement. De même, faire des démonstrations des fonctionnalités à partir du prototype ou de la version bêta de la plateforme apporte une dimension concrète à un projet qui relève encore du domaine conceptuel. Par ailleurs, confronté à des utilisateurs plus ou moins autonomes et dont les comportements diffèrent, il est important d'analyser les différents publics types à qui est destiné l'outil. A chaque public, des services adaptés doivent être mis en place. Ainsi, pour les utilisateurs les plus autonomes dans un environnement numérique, il peut suffire de mettre en ligne une aide sous forme de tutoriel ou de guide qui serve de base de travail et fournisse l'ensemble des règles à respecter pour la mise en ligne de leurs données. Pour d'autres, proposer des formations collectives et/ou individuelles est un service important qui peut créer une dynamique d'utilisation de l'outil. Parfois, il s'agira de se rendre disponible pour être aux côtés de l'utilisateur afin de participer à la mise en ligne des fichiers et de leur documentation.

3.4.2 La valeur ajoutée du produit réalisé

Il est crucial de mettre en avant les fonctionnalités les plus performantes et les plus pertinentes lors des démonstrations auprès du public ciblé. Ainsi, le projet TELEMETA ne propose pas uniquement de fournir une plateforme où déposer des documents sonores, mais également de mettre à disposition des outils de traitement et d'analyse du contenu qui donnent des résultats réutilisables et exportables dans le cadre de projets de publications. Cette approche donne une vision plus globale des avantages à mettre ses données en ligne : le chercheur n'est plus dans une démarche chronophage d'archivage de ses données. Il s'investit dans le dépôt et le partage de ses archives afin de pouvoir exploiter les données en ligne dans un espace de travail unique qui intègre l'ensemble des fonctionnalités dont il a besoin de manière ergonomique et simplifiée. Le projet prend alors une valeur ajoutée.

3.4.3 La valorisation des dépôts

Certaines pratiques liées au développement du web collaboratif peuvent être mises à profit dans un cadre scientifique. En effet, valoriser les collections en ligne peut être réalisé sous forme d'articles, dans un blog régulièrement mis à jour, qui donnent à voir l'évolution des contributions. Un blog offre également une visibilité accrue auprès d'un public qui n'est pas toujours facilement atteignable, et permet d'actualiser les informations à diffuser de manière simplifiée. Cette méthode apporte un regard dynamique et moins formel sur le projet. Par ailleurs, cela évite d'avoir à diffuser des informations par mail au risque de surcharger la boîte mail des utilisateurs. Bien au contraire, à partir du moment où la création d'un blog a fait l'objet d'une communication auprès des principaux intéressés, c'est à leur initiative qu'ils viendront consulter les articles du blog afin de les parcourir à leur propre rythme et selon leurs besoins immédiats. Cela permet également d'ouvrir un espace de parole dans lequel les lecteurs peuvent témoigner, partager leurs expériences, proposer des améliorations, etc. Bien entendu, ce type d'outil exige de contribuer régulièrement à son actualisation et à son alimentation. Il faut alors envisager cette contrainte comme un moyen détourné de mettre en place sa propre veille sur des sujets importants. Cette initiative demande de l'investissement et, surtout, elle impose d'avoir un service ou des personnes rattachées en permanence au projet qui seraient les principaux contributeurs.

Conclusion

Depuis les années 2 000, une mutation profonde et symptomatique du paysage technologique a contribué à faire évoluer les pratiques et les usages de la communauté scientifique en France et dans le monde. Internet est devenu l'instrument incontournable sur lequel repose les objectifs de partage des données de la recherche et de leur exploitation. L'environnement numérique est devenu le point névralgique à partir duquel de nouvelles formes de services, d'applications et d'outils sont élaborées afin de répondre à une croissance exponentielle de données numériques à mettre à disposition des chercheurs. Les données brutes collectées sur fonds public sont au cœur des préoccupations politiques actuelles puisqu'elles ouvrent un champ d'investigation et d'exploration infini dont les retombées peuvent être aussi bien d'ordre économique, humain qu'industriel. Par ailleurs, leur consultation est nécessaire à la compréhension des productions scientifiques finalisées sous forme de résultats d'analyse ou de publications. Avec la création d'un Espace européen de la recherche, une réflexion commune sur les infrastructures à développer a été élaborée. Une dizaine d'années plus tard, il s'agit d'évaluer les résultats de la stratégie mise en œuvre à l'échelle européenne et d'établir l'état d'avancement des projets concernant la diffusion des données de la recherche. Les Sciences humaines et sociales ont bénéficié d'une dynamique inspirée de la constatation que ce domaine de recherche n'avait pas encore mis à profit le numérique pour construire un environnement technologique alliant des capacités de calcul puissantes et des services garantissant l'exploitation des données mises en ligne. Par ailleurs, les promesses du web de données offrent de nouvelles perspectives pour optimiser la publication des données et assurer leur lisibilité et leur exploitation par les machines, à travers la création de relations signifiantes entre les données. Les donnés des SHS ont la particularité d'avoir une valeur scientifique et patrimoniale qui leur confère un statut spécifique rattaché à la valorisation de la recherche et du patrimoine culturel. Ce contexte a fourni les moyens à des centres de recherche en SHS, producteurs de données brutes, de mettre en place des projets spécifiques selon les objets d'étude et les disciplines internes à chacun. La comparaison entre différentes plateformes de diffusion des données sonores et audiovisuelles des chercheurs met en perspective des pratiques communes et des systèmes performants qui ont chacun leurs points forts. Ainsi, pour pallier aux difficultés liées à la complexité d'un environnement technologique évolutif, les structures proposent aussi bien d'accompagner les producteurs de données dans leur projet de mise à disposition des données, qu'une grille de services afin de faciliter leur travail de traitement des données avant versement. Cependant, l'un des défis majeurs à relever est de réussir à impliquer le chercheur et de lui fournir un espace de travail offrant une gamme complète d'outils de traitement et d'analyse des données. L'objectif sous-jacent est que le chercheur soit le plus autonome possible et qu'il dépose, documente, exploite les données lui-même. Les inclinations naturelles à l'autonomie ont été, à mon sens, trop prises à la légère. Peut-être

parce qu'elles semblaient évidentes : qui ne voudrait pas disposer d'un espace personnel de travail réalisé sur mesure ? De fait, la communauté des Digital Humanities est activement impliquée dans les pratiques numériques, mais il s'avère que cette communauté ne représente pas la majorité des chercheurs en SHS et leurs pratiques. De nombreux chercheurs, qui ne contestent pas l'intérêt à mettre à disposition les données de la recherche, peuvent se sentir perdus dans un environnement qui ne leur est pas familier, et de ce fait, qui ne leur semble pas facile à prendre en main. Les concepteurs se doivent donc de développer des interfaces simplifiées et ergonomiques. Par ailleurs, il est devenu indiscutable de former les futurs chercheurs pour les familiariser avec les outils qui leur sont destinés, et pour leur permettre d'appréhender les nouvelles méthodes de recherche issues des technologies numériques. Mais qu'en est-il des chercheurs actuels ? Ne devraient-ils pas bénéficier de formations adaptées ? La mise en place d'un accompagnement au changement et la disponibilité des concepteurs sont des conditions importantes pour assurer le suivi du projet et encourager les chercheurs à partager leurs données. D'autre part, la mise en place d'indicateurs de gestion et de consultation donne à voir les usages réels de l'outil et la satisfaction de ses utilisateurs. L'élaboration d'une stratégie de communication offre les moyens de valoriser les corpus en ligne auprès des contributeurs et des chercheurs et de diffuser des informations ciblées avec d'éventuels commentaires et/ou retours d'expérience. Ces éléments permettent d'établir des mesures précises et concrètes du degré d'exploitation de la plateforme de diffusion afin de mettre en œuvre des développements et des évolutions pour s'assurer d'atteindre les objectifs initiaux. En conséquence, le rôle des professionnels de l'information et de la documentation reste important et incontournable. Qu'il s'agisse du traitement des données, de la conception du projet sur le versant documentaire, de l'accompagnement au changement auprès du public cible, du suivi des indicateurs de gestion, le professionnel de l'information et de la documentation doit devenir un membre de l'équipe projet dès sa conception. Ses compétences se déclinent d'un bout à l'autre de la chaîne documentaire à mettre en place : élaboration d'une politique de dépôt et de définition des droits d'accès à partir des conditions légales de diffusion, structuration de l'information (niveau métadonnées), veille documentaire (niveau formats et standards du Web), traitement des données, architecture globale de l'information (niveau site), etc. L'implication d'un acteur spécialiste des questions documentaires permet de monter un projet réaliste et orienté vers les besoins des chercheurs. Acteur qui pourra, en outre, assurer le suivi du projet après sa réalisation et la maintenance documentaire. Cela contribuera à garantir que le projet, même s'il évolue dans le temps, repose sur une base documentaire solide qui facilite le travail du chercheur pour qu'il n'ait plus alors qu'à se préoccuper d'exploiter les données disponibles en ligne.

Le contexte numérique de la recherche scientifique questionne en profondeur le décalage actuel qui existe entre une vision politique qui pousse au partage des données et une implication des chercheurs qui s'avère encore fragile. Les scénarios d'usage des plateformes mises en place ainsi que la ré-utilisation des données répondent-ils aux besoins réels des chercheurs ? Une étude fine de ces questions permettrait de mieux appréhender les freins rencontrés au sein des équipes de chercheurs et d'apporter des solutions ciblées.

Bibliographie

La bibliographie analytique suivante, arrêtée au 5 décembre 2013, est classée selon une numérotation qui correspond à l'ordre d'apparition de la référence dans le mémoire. Elle suit donc le plan du mémoire. Un index alphabétique d'auteurs est également proposé.

Introduction

[1] OCDE. Principes et lignes directrices pour l'accès aux données de la recherche financées sur fonds public [en ligne]. Paris : OCDE, 2007 [consulté le 23 octobre 2013]. [p. 9]
<http://www.oecd.org/fr/science/sci-tech/38500823.pdf>
L'OCDE proposent des orientations aux responsables politiques et scientifiques pour élaborer des politiques et des bonnes pratiques autour des données numériques de la recherche financés par des fonds public et faciliter leur accès, leur exploitation et leur gestion.

[2] ADBS. Espace « Accéder à la doc professionnelle », « Vocabulaire de la doc » de l'Association des professionnels de l'information et de la documentation [consulté le 15 novembre 2013]. [p. 10] [p. 22]
<http://www.adbs.fr/vocabulaire-de-la-documentation-41820.htm?RH=OUTILS_VOC&RF=OUTILS_VOC >
L'Association des professionnels de l'information et de la documentation met en ligne un vocabulaire technique dans le domaine de l'information-documentation publié en 2004.

[3] ROUGE-DUCOS, Isabelle. Bulletin des Archives de France sur la conservation à long terme des documents électroniques [en ligne]. Bulletin des Archives de France, n° 10, 2003 [consulté le 25 octobre 2013]. [p. 10]
<http://www.archivesdefrance.culture.gouv.fr/static/1674>
Il s'agit d'un compte-rendu du colloque intitulé « La valorisation et la pérennisation des données scientifiques et techniques » qui s'est déroulé du 5 au 7 novembre 2002 à Toulouse.

Première partie

[4] BERNERS-LEE, Tim. Site du W3C (World Wide Web Consortium) [consulté le 25 octobre 2013]. [p. 13]
< http://www.w3.org/People/Berners-Lee/Longer.html >
Le W3C est une organisation internationale qui définit les standards du Web. Le lien mentionné renvoie à la bibliographie de Tim Berners-Lee, le fondateur du World Wide Web.

[5] COLLECTIF. Budapest Open Access Initiative [en ligne]. Budapest : 2002 [consulté le 5 novembre 2013]. [p. 13]

<http://www.budapestopenaccessinitiative.org/read >
Cet appel à l'initiative de chercheurs revendique le droit à la diffusion libre et gratuite des résultats scientifiques et initie le mouvement de l'Open Access ou Libre Accès.

[6] COLLECTIF. Déclaration de Berlin sur le Libre Accès à la Connaissance à la Science exacte, Sciences de la vie, Sciences humaines et sociales [en ligne]. In Site du Max Planck Open Access. Munich : 2003 [consulté le 4 novembre 2013]. [p. 14] <http://openaccess.mpg.de/286432/Berlin-Declaration>
Les citations en français proviennent de la version française issue des « pdf versions Berlin Declaration » : < http://openaccess.mpg.de/68042/BerlinDeclaration_wsis_fr.pdf >
Cette déclaration met en évidence le rôle qu'Internet peut et doit jouer pour un accès universel aux connaissances scientifiques. Le mouvement de l'Open Access prend de l'ampleur puisqu'y adhérent des organismes scientifiques et des institutions culturelles et patrimoniales qui prennent conscience des enjeux du numérique.

[7] COLLECTIF. Déclaration de principes - Construire la société de l'information: un défi mondial pour le nouveau millénaire [en ligne]. In Site du Sommet mondial sur la société de l'information Genève 2003 – Tunis 2005, document WSIS-03/GENEVA/DOC/4-F, publié le 12 mai 2004 [consulté le 23 octobre 2013]. [p. 14]
< http://www.itu.int/wsis/docs/geneva/official/dop-fr.html >
Lors de ce sommet tenu à Genève, les participants ont rédigé une déclaration qui promeut la place des TIC dans le développement d'une société de l'information favorisant le partage des connaissances et du savoir. Les TIC sont considérés comme le moyen d'atteindre les objectifs de développement humain des Nations Unies.

[8] COLLECTIF. Déclaration de principes - Construire la société de l'information: un défi mondial pour le nouveau millénaire [en ligne]. In Site du Sommet mondial sur la société de l'information Genève 2003 – Tunis 2005, document WSIS-05/TUNIS/DOC/007-F, publié le 18 novembre 2005 [consulté le 5 novembre 2013]. [p. 15]
< http://www.itu.int/wsis/docs2/tunis/off/7-fr.html >
Dans cette déclaration qui fait suite à la déclaration de Genève, les signataires réaffirment le rôle des TIC comme vecteur de construction d'une société de la connaissance et s'engagent à élaborer une stratégie commune de gouvernance d'Internet.

[9] SWAN, Alma. Principes directeurs pour le développement et la promotion du Libre Accès [en ligne]. Paris : UNESCO, 2013 [consulté le 23 octobre 2013]. [p. 15]
< http://unesdoc.unesco.org/images/0022/002220/222085f.pdf > ISBN 9230010522

Ce guide à destination des décideurs et des organismes et institutions de financement internationaux a pour objectif de permettre l'évaluation et l'adoption d'une politique de libre accès à l'information scientifique et d'en comprendre les enjeux stratégiques, économiques et politiques.

[10] CNRS, UNIVERSITE D'ORLEANS, UNIVERSITE DE TOUR. Site de la Cellule Mutualisée « Europe Recherche » [consulté le 30 octobre 2013]. [p. 16]
< http://cellule-europe-recherche-centre.fr/programmes-europeens/thematique-prioritaire-3-technologies-de-linformation-et-de-la-communication/ >
Suite à la mise en place d'un Espace européen de la recherche, cette plateforme propose une veille sur les programmes européens et fournit des informations en français sur les appels à projets.

[11] SIREN, Jarkko. Strategies for Scientific and Technological Information in Europe. In JEZEQUEL, Armelle, coord. Actes des FRéDoc 2011 : L'IST au prisme de l'Europe [en ligne]. Bordeaux : halshs-00868912, version 1. Octobre 2013 [consulté le 5 novembre 2013]. [p. 16]
< http://halshs.archives-ouvertes.fr/halshs-00868912 >
Cet article retrace les décisions et directives énoncées par la Commission Européenne pour un accès aux informations scientifiques à l'échelle européenne.

[12] COMMISSION DES COMMUNAUTES EUROPEENNES. Communication de la commission au Parlement Européen, au Conseil et au Comité Economique et Social Européen sur l'information scientifique et technique à l'ère numérique : accès, diffusion et préservation [en ligne]. Bruxelles : 2007 [consulté le 13 novembre 2013]. [p. 16]
< http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2007:0056:FIN:FR:PDF >
Ce document alerte l'Union Européenne sur la nécessité de prendre des mesures politiques pour accélérer la mise en place d'infrastructures pour diffuser l'information scientifique et pour préserver les contenus numériques.

[13] HAMEAU, Thérèse. Panorama des projets européens en faveur du libre accès à l'IST. In JEZEQUEL, Armelle, coord. Actes des FRéDoc 2011 : L'IST au prisme de l'Europe [en ligne]. Bordeaux : halshs-00868912, version 1. Octobre 2013 [consulté le 5 novembre 2013]. [p. 16]
< http://halshs.archives-ouvertes.fr/halshs-00868912 >

Dans cet article, l'auteur présente différents projets européens portant sur le Libre Accès ainsi qu'une analyse des problématiques juridiques et financières qui se posent pour les chercheurs.

[14] DESRUELLES, François-Xavier. Le programme e-Science au Royaume-Uni [en ligne]. In Site Veille technologique internationale, publié le 31 mars 2007 [consulté le 25 octobre 2013]. [p. 17]
< http://www.bulletins-electroniques.com/rapports/smm07_021.htm >
Ce rapport est une présentation synthétique en français du projet e-Science développé par le Royaume-Uni en 2000 : acteurs, financement, projets, etc.

[15] CHARRON, Thierry. Le projet européen de grille de calcul (DataGrid) : Concept de grilles de calcul. Objectifs du projet et état d'avancement [en ligne]. Paris, Lyon : Conservatoire National des Arts et Métiers - Centre d'enseignement de Lyon, 2003 [consulté le 28 octobre 2013]. [p. 17]
< http://cjt.eljako.org/cnam/pr_detail.php?id=82 >
L'auteur introduit sa présentation du projet européen DataGrid mené par le CERN par une définition et un historique du concept de grille de calcul.

[16] MESR. Site du ministère de l'Enseignement Supérieur et de la Recherche [consulté le 12 novembre 2013]. [p. 17] [p. 18]
< http://www.enseignementsup-recherche.gouv.fr/cid72588/la-strategie-nationale-des-infrastructures-de-recherche.html >
< http://www.enseignementsup-recherche.gouv.fr/cid20438/les-missions-de-l-information-scientifique-et-technique.html >
Dans le domaine de l'information scientifique et technique, une documentation importante présente les projets mis en place dans le cadre de la stratégie nationale de recherche et de l'élaboration d'infrastructures de recherche. Le site offre également l'ensemble des informations concernant les projets européens et internationaux en matière d'innovation pour la recherche.

[17] HUMA-NUM. Site Huma-Num, la TGIR des humanités numériques [consulté le 28 octobre 2013]. [p. 18]
< http://www.huma-num.fr/ >
Site à destination des chercheurs qui propose un accompagnement et des outils de gestion des données dans le domaine des SHS.

[18] CINES. Site du Centre Informatique National de l'Enseignement Supérieur [consulté le 13 novembre 2013]. [p. 20]
< http://www.cines.fr/ >
Le CINES est un établissement public national dont les missions sont d'offrir une puissance de calcul intensif pour l'exploitation des données et une plateforme d'archivage pérenne des données reconnues d'intérêt national.

[19] CAPELLI, Laurent. Evolution de l'archive ouverte HAL-SHS. In THATCamp Paris 2012. Non actes de la non-conférence des humanités numériques. Paris : 2012 [consulté le 13 novembre 2013]. [p. 21]
< http://books.openedition.org/editionsmsh/297> ISBN : 9782735115273
Dans cet article, l'auteur fait un état des lieux des développements en cours pour l'amélioration et l'optimisation du projet d'archive ouverte HAL-SHS.

[20] PEUGEOT, Valérie. Le web de données laisse-t-il une place au bien commun ? In Libres savoirs. Les biens communs de la connaissance. Caen : C & F Edition, 2001, p. 193-210. ISBN 2915825060 [p. 22]
Cet article s'intéresse aux enjeux posés par la diffusion et le partage des données publiques et privées sur Internet.

[21] BNF. Espace « professionnels » du site de la Bibliothèque nationale de France [consulté le 12 novembre 2013]. [p. 22] [p. 26]
< http://www.bnf.fr/fr/professionnels.html >
Cet espace destiné aux professionnels de l'information et de la documentation propose tout un ensemble d'informations pratiques sur les sujets relatifs au catalogage, à l'archivage des données, aux concepts du web sémantique, etc.

[22] BERMES, Emmanuelle. Web de données 1 : Qu'est-ce que le web de données ? [en ligne]. In Site de l'ADBS, publié le 9 mars 2010, mis à jour le 30 mai 2013 [consulté le 12 novembre 2013]. [p. 23]
<http://www.adbs.fr/web-de-donnees-1-qu-est-ce-que-le-web-de-donnees-130123.htm >
L'ADBS organise des « 5 à 7 » tous les deux mois pour ses adhérents autour d'un spécialiste. Ici, il s'agit de l'intervention d'Emmanuelle Bermès qui présente l'histoire du développement du Web et de ses évolutions majeures.

[23] SALAUN, Jean-Michel. Du document à la donnée et retour – La fourmilière ou les Lumières. In Le document numérique à l'heure du web de données – Séminaire INRIA, Carnac – 1er -5 octobre 2012. Paris : ADBS Editions, 2012. ISBN 2843651427 [p. 23]
L'auteur aborde la notion de document et son évolution depuis l'avènement du numérique et au regard des nouvelles technologies initiées par le web de données.

[24] ZACKLAD, Manuel. Evaluation des Systèmes d'Organisation des Connaissances. In Les Cahiers du numérique, vol. 6, n° 3, p. 133-166. Cachan : Lavoisier, 2010. ISSN 1622-1494 [p. 26]
L'auteur propose une grille d'analyse des systèmes d'organisation des connaissances adaptés à la gestion de l'information documentaire numérique.

[25] TRONCY, Raphaël. Owl, un « chouette » langage pour représenter des ontologies. In Documentaliste, Sciences de l'information vol. 48, n° 4. Dossier Web sémantique, web de données… Quelle nouvelle donne, p. 4. Paris : ADBS, 2011. ISSN 0012-4508 [p. 26]
Dans cet article très bref, l'auteur expose les principes du langage de représentation d'ontologies, Owl en l'illustrant d'exemples.

[26] CAPELLI, Laurent, KILOUSHI, Nadia, MINEL, Jean-Luc [et al.]. Comment contribuer, avec ses données numériques, à ISIDORE ? [en ligne]. Version 2. Janvier 2012 [consulté le 11 novembre 2013]. [p. 27]
< http://www.huma-num.fr/sites/default/files/ressourcesdoc/guide_isidore_2012.pdf >
Ce guide précise l'ensemble des bonnes pratiques et les conditions techniques à mettre en œuvre afin de faciliter la collecte des données et leur traitement par ISIDORE.

[26b] INTD-CNAM. Digital Humanities International – Veille sur les humanités numériques et champs associés [en ligne]. [consulté le 15 décembre 2013]. [p. 32]
< http://dhi.intd.cnam.fr/ >
Ce blog, arrêté courant 2012, présente les Digital Humanities et propose une actualité sur le domaine à l'international.

[27] DACOS, Marin. Manifeste des Digital Humanities [en ligne]. Publié le 26 mars 2011 [consulté le 11 novembre 2013]. [p. 32]
< http://tcp.hypotheses.org/318 >
Ce manifeste définit les Digital Humanities comme une communauté de pratiques et une discipline à part entière dans le domaine des Sciences humaines et sociales qui soutient

l'Open Access et veut contribuer au développement des cyberinfrastructures destinées aux chercheurs.

[28] DH. Liste de diffusion des Digital Humanities. [p. 33]
< archives-son-audiovisuel@listes.revues.org >
Cette liste de diffusion permet aux chercheurs et acteurs des DH d'échanger des informations et de soutenir leur mouvement.

[29] TERRAS, Mélissa. Un regard jeté sous le chapiteau : les humanités numériques et la crise de l'inclusion. In THATCamp Paris 2012. Non actes de la non-conférence des humanités numériques. Paris : 2012 [consulté le 13 novembre 2013]. [p. 33]
Dans cet article, l'auteur retrace le parcours réalisé par un chercheur dans le contexte numérique actuel de la collecte des données à leur exploitation.

[30] DALBIN, Sylvie. Métiers et organisation professionnelle. Perspectives européennes. In JEZEQUEL, Armelle, coord. Actes des FRéDoc 2011 : L'IST au prisme de l'Europe [en ligne]. Bordeaux : halshs-00868912, version 1. Octobre 2013 [consulté le 5 novembre 2013]. [p. 33]
< http://halshs.archives-ouvertes.fr/halshs-00868912 >
Dans cet article, l'auteur aborde du point de vue des professionnels de l'IST les perspectives et les moyens qui leur sont attribués dans une dynamique européenne au profit de la recherche.

[31] ARCHIVES-SON-AUDIOVISUEL. Liste de diffusion. [p. 38]
< archives-son-audiovisuel@listes.revues.org >
Cette liste de diffusion ouverte offre de nombreuses occasions d'échanges entre des personnes dont les profils sont très divers mais dont les préoccupations s'articulent autour des archives audiovisuelles.

Deuxième partie

[32] UNESCO. Texte de la Convention pour la sauvegarde du patrimoine culturel immatériel [en ligne]. Paris : 2003 [consulté le 22 novembre 2013]. [p. 39]
< http://www.unesco.org/culture/ich/fr/convention/ >
Ce texte, adopté par l'ONU le 17 octobre 2003, marque la prise de conscience internationale des enjeux liés à la conservation et à la sauvegarde du patrimoine culturel immatériel.

[33] GINOUVES, Véronique, GARRET, Pascal [et al.]. Carnet de recherche « Questions éthique et droit en SHS » [consulté le 25 novembre 2013]. [p. 40]
< http://ethiquedroit.hypotheses.org/ >
Ce carnet de recherche réunit des articles rédigés par des experts sur les questions juridiques et éthiques liées à la diffusion de corpus numériques iconographiques, sonores et audiovisuels des chercheurs par les scientifiques et les gestionnaires de projets de valorisation de ces ressources.

[34] ARCHIVES DES ETHNOLOGUES. Site du consortium [consulté le 25 novembre 2013]. [p. 41]
< http://ethnologues.corpus-ir.fr/ >
Ce site informe sur le contexte de mise en place et sur l'état d'avancement du consortium « Archives des ethnologues ».

[35] GINOUVES, Véronique. Carnet de recherche « Les carnets de la phonothèque » [consulté le 25 novembre 2013]. [p. 42]
< http://phonotheque.hypotheses.org/ >
Ce carnet de recherche valorise le fonds d'archives sonores déposés à la phonothèque de la Maison méditerranéenne des sciences de l'Homme. Il fournit également une veille sur les actualités liées aux sources orales.

[36] ROUEFF, Olivier. L'ethnologie musicale selon André Schaeffner, entre musée et performance [en ligne]. In Revue d'Histoire des Sciences Humaines, vol. 1, n° 14. Villeneuve d'Asq : Presse universitaires du Septentrion, 2006, p. 71-100 [consulté le 23 novembre 2013]. [p. 48]
< www.cairn.info/revue-histoire-des-sciences-humaines-2006-1-page-71.htm >
Cet article retrace le parcours d'André Schaeffner, de ses débuts en tant que critique littéraire et musical à sa nomination au poste de directeur du département d'ethnomusicologie du Musée de l'Homme, en mettant en évidence les axes de recherche et d'études qui ont contribué à sa réflexion.

[37] PROJET ANTHROPONET. In site de l'IRI. Publié le 16 décembre 2009 [consulté le 25 novembre 2013]. [p. 48]
< http://www.iri.centrepompidou.fr/projets/anthroponet/ >
Cet article présente le projet de méta-portail Anthroponet, ses objectifs et les acteurs impliqués. Ce projet a pour cadre la gestion, la diffusion, l'indexation des ressources audiovisuelles à destination du chercheur en Sciences humaines et sociales.

[38] THERON, Dominique. Ecrire un cahier des charges de numérisation de collections sonores, audiovisuelles et filmiques [en ligne]. Paris : Ministère de la Culture et de la Communication, Comité de pilotage numérisation, Bibliothèque nationale de France, 2009 [consulté le 25 novembre 2013]. [p. 62]
<http://www.culture.gouv.fr/culture/mrt/numerisation/fr/technique/documents/cahier_charges_numerisation.pdf >
Ce document est à destination de toute personne disposant de données sonores ou audiovisuelles sur des supports analogiques et qui souhaiterait externaliser leur numérisation. Ce guide fournit l'ensemble des bases nécessaires à la rédaction du cahier des charges.

Troisième partie

[39] TGE-ADONIS. Guide méthodologique pour le choix de formats numériques pérennes dans un contexte de données orales et visuelles [en ligne]. Version 0.5, publié le 26 mai 2011. Numéro ADONIS/SIAF/CINES-GM-0.5 [consulté le 2 décembre 2013]. [p. 65]
< http://www.archivesdefrance.culture.gouv.fr/static/4923 >
Dans le cadre d'un projet d'archivage de données sonores et audiovisuelles, ce guide propose une méthodologie et une étude des formats qui répondent à l'éligibilité de la conservation à long terme des données orales et audiovisuelles.

[40] BURNARD, Lou, DARDENNE, Nadine, DAVID, Sophie. Guide des bonnes pratiques numériques [en ligne]. Version 2, publié le 15 septembre 2011 [consulté le 2 décembre 2013]. [p. 65]
<http://www.huma-num.fr/sites/default/files/ressourcesdoc/guide_des_bonnes_pratiques_v2.pdf >
Ce guide se présente comme un guide méthodologique qui établit les étapes à suivre pour mettre en place un projet de diffusion des données numériques dans une unité de recherche ou à l'initiative des chercheurs.

Index des auteurs

SALAUN, Jean-Michel [23]
SIREN, Jarkko [11]
SWAN, Alma [9]
TERRAS, Mélissa [29]
TGE-ADONIS [39]
THERON, Dominique [38]
TRONCY, Raphaël [25]
UNESCO [32]
ZACKLAD, Manuel [24]

Annexes

Annexe 1 – Liste des abréviations

ABES : Agence bibliographique de l'enseignement supérieur

ANR : Agence nationale de la recherche

CINES : Centre informatique national de l'enseignement supérieur

CRDO : Centre de ressources pour la description de l'oral

CREDO : Centre de recherche et de documentation sur l'Océanie (UMR 6574)

CREM : Centre de recherche en ethnomusicologie – CREM/LESC (UMR 7186)

DH : Digital Humanities

DSA : Data Seal of Approval

ERA : Espace européen de la recherche

IANA : Internet Assigned Numbers Authority

IST : Information scientifique et technique

LAM : Laboratoire Lutheries – Acoustique – Musique/IJLRA

MESR : Ministère de l'Enseignement Supérieur et de la Recherche

MMSH : Maison méditerranéenne des sciences de l'Homme

MRT : Mission de la recherche et de la technologie

OAI : Open Archives Initiative

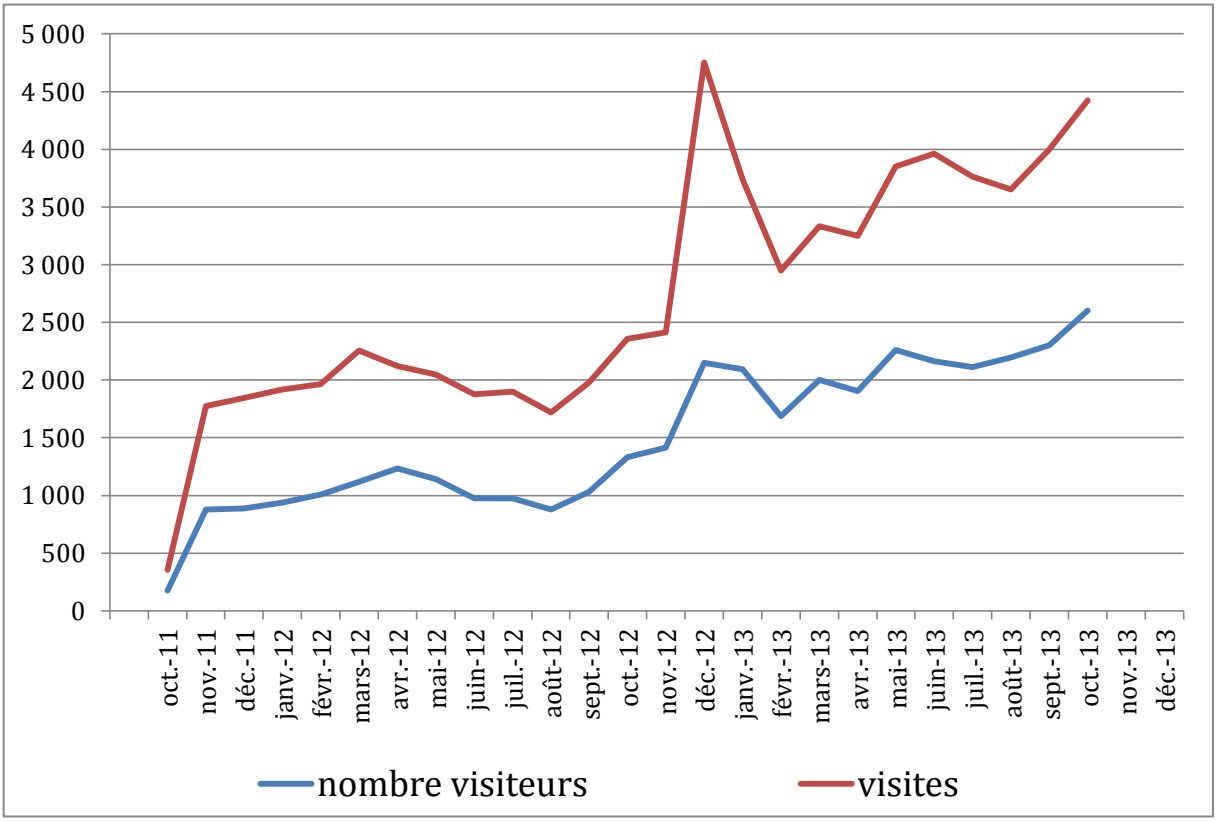
OAI – PMH : Open Archives Initiative – Protocol for Medata Harvesting

OAIS : Open Archival Information System

TIC : Technologies de l'information et de la communication

Annexe 2 – Statistiques de fréquentation de la plateforme TELEMETA du CREM

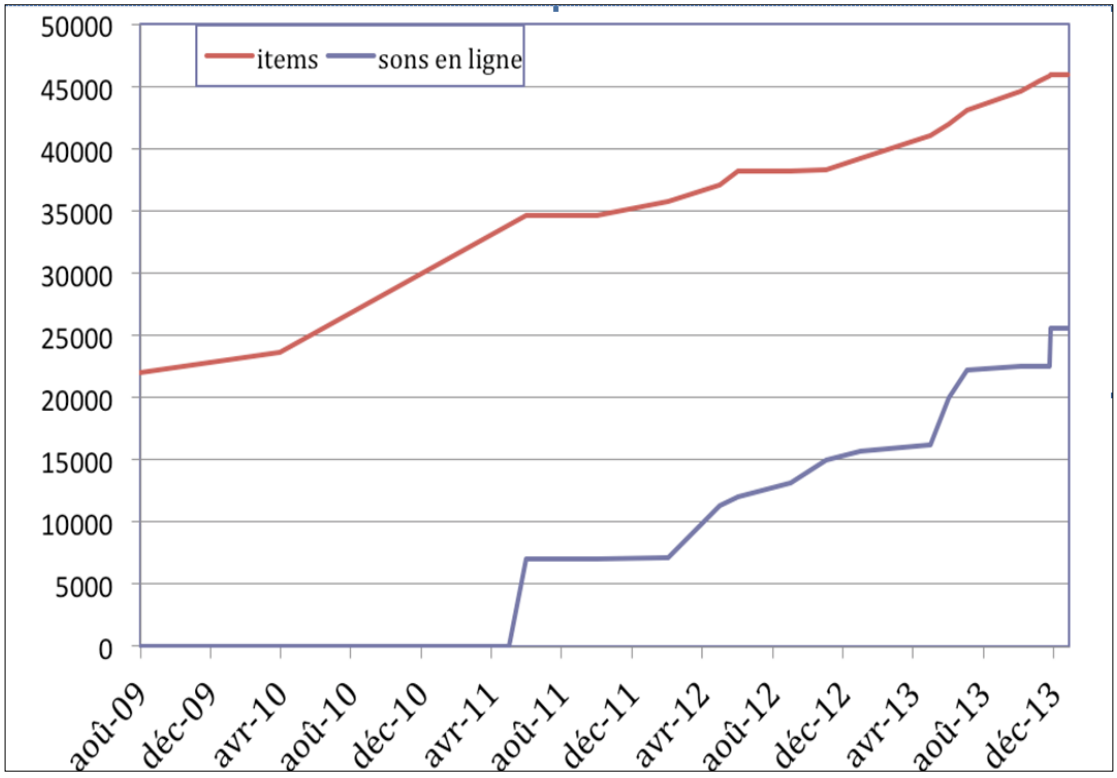
Evolution de la fréquentation des archives du CNRS-Musée de l'Homme diffusées sur TELEMETA :



Source : Statistiques TELEMETA – CREM

Annexe 3 – Evolution des dépôts dans la plateforme TELEMETA du CREM

Evolution du nombre d'items (fiches documentaires) et de fichiers son mis en ligne :



Source : Statistiques TELEMETA – CREM