L'archivage des données numériques pour la recherche par le Centre National pour la Numérisation de Sources Visuelles

Stéphane Pouyllau, Daniel Pouyllau, Marie-Dominique Mouton, Fabrice Melka

PROJET DE CENTRE DE COMPETENCES EN SCIENCES DE L'HOMME ET DE LA SOCIETE

Informatisation des ressources scientifiques pour la recherche : les fonds de scientifiques et la documentation cartographique et photographique en histoire des sciences et des techniques et sur les aires culturelles

1 Contexte

1.1 Les archives scientifiques

A la différence des sciences exactes, le patrimoine des sciences humaines et sociales reste encore largement inexploité. Seule la production scientifique elle-même a suscité des questionnements et des réalisations numériques (HAL) et peu d'initiatives ont été lancées concernant les archives et les documents primaires produits par les chercheurs eux-mêmes.

Il n'existe ainsi ni réglementation ni réel dispositif permettant la collecte et la conservation des matériaux documentaires issus de l'activité de recherche proprement dite. Ces fonds sont beaucoup plus fréquemment conservés dans les laboratoires, les bibliothèques, voire dans les familles, que dans des services d'archives.

1.2 Corpus/Domaines

1.2.1 Un savoir-faire : histoire des sciences et des techniques

Dans le cadre des SHS, les avancées les plus significatives en matière de création et d'innovation de corpus numérique ont été réalisées dans le domaine de l'Histoire des sciences et des techniques. Depuis 1992, plusieurs centres de recherche universitaires ont développé des stratégies d'informatisation des données qui font école aujourd'hui et qui s'exportent au niveau européen.

Sans avoir pour autant les moyens de certains de nos voisins européens, les centres de recherche et bibliothèques spécialisées dans l'histoire des sciences et des techniques ont investi depuis plusieurs années le domaine des TIC. Le CNAM (avec ABU et le CNUM), la BNF (avec Gallica), la Médiathèque d'histoire des sciences de la Cité des Sciences, proposent des ressources numériques en ligne et des accès à des dossiers documentaires en texte intégral.

1.2.2 Exemple d'un manque : Aires culturelles

Le patrimoine scientifique sur les aires culturelles nous semble avoir sa propre spécificité, d'une part grâce à la place importante qu'y occupent la variété des matériaux de terrain (notes, carnets, productions sonores, iconographiques ou audiovisuelles), d'autre part, hormis un réel efforts de conservation en ethnologie, en raison de l'absence de programme permettant de le sauvegarder et de le diffuser.

Surtout à partir des années 1950 s'est constitué un vaste champ scientifique producteur de sources originales dont la conservation et la valorisation constituent un enjeu important.

Il est indispensable d'accorder aujourd'hui une attention privilégiée à ces fonds (et archives personnelles), à leur conservation et à leur consultation. Ils sont conservés d'une part par des collecteurs (centres de documentation d'unité, bibliothèques de recherche), d'autre part en propre par les chercheurs, et pour lesquels, bien souvent, aucune des conditions minimales de conservation n'est remplie.

1.2.3 Nécessité de la numérisation

Ces données d'enquête ont tendance à disparaître avec le temps :

- dégradation physique des supports ;
- disparition des appareils de lecture de certains de ces supports ;
- séparation physique des originaux et de leurs analyses ;

- non publication des données, ni recensement ni catalogue des matériaux d'enquêtes ;
- dépendance de ces données vis à vis de leurs producteurs (formats, annotations, localisation…).

Tout autant de raisons pour se préoccuper dès maintenant de la transmission de ces données et connaissances qu'elles datent d'hier ou d'aujourd'hui.

1.3 Propositions pour une expérimentation

Avec ce projet, il s'agit pour nous de mettre en place et de valider un processus et des outils collaboratifs de travail informatiques et documentaires, des solutions techniques sûres, axées sur la pérennité et l'interopérabilité des données.

Ce système permet la création de réseaux, une décentralisation du travail et l'accession à des ressources réparties où chacun a la responsabilité de ses propres données. Ainsi le projet s'accroît à des rythmes différents suivant les nouvelles perspectives de financement et de collaboration.

La démarche se veut pragmatique : les partenaires de ce réseau se connaissent et partagent le même intérêt pour la conservation des archives scientifiques. Par ailleurs ils ont déjà eu pour certains l'occasion de travailler ensemble sur des projets similaires. Ils disposent de savoir-faire complémentaires et de ressources sur lesquelles travailler. Nous insistons sur l'importance de la participation des centres de documentation et des bibliothèques de recherche (centre de documentation REGARDS-ADES, Réseau ethnologie, UMS Caphes) et sur le rôle de relais des partenaires de ce projet auprès de réseaux scientifiques plus larges (GIS, RTP, Réseau Pléiade, etc.).

Elle est aussi expérimentale : la diversité des domaines et des documents à considérer est source de fructueux questionnements et aboutira à la mise en place de méthodologies exportables et à la création d'outils génériques ; la dimension patrimoniale assurant la cohésion du traitement de ces ressources.

Deux domaines d'application :
- histoire des sciences et des techniques
- aires culturelles

Deux types de collection d'archives scientifiques sont disponibles :
- les fonds privés de scientifiques (documents écrits, iconographiques, sonores, audiovisuels)
- la documentation cartographique et photographique réunie dans les centres de documentation et les bibliothèques de recherche :
  - documents cartographiques et photographiques
  - cartes géologiques historiques d'Europe (projet européen HistMap financé par l'ESF)

Trois disciplines académiques :
- histoire moderne et contemporaine ;
- anthropologie sociale et Ethnologie ;
- géographie.

2 Objectifs : Des services informatiques et documentaires au service de la recherche permettant la création de ressources numériques : de la production à la diffusion, le rôle du centre de compétences.

Ce projet se propose de constituer un ensemble intégré, combinant veille patrimoniale, conservation et traitement d'archives et de collections, services liés à la recherche et diffusion des ressources numériques.

2.1 Sauvegarder et numériser : une patrimonialisation des fonds documentaires scientifiques

Nombreux, épars, uniques et peu connus, les fonds documentaires scientifiques méritent d'être localisés et préservés. Le patrimoine n'existe pas a priori, il se construit. Pour fonctionner

comme patrimoine il doit être connu et reconnu comme tel. Cette visibilité débouchant sur le principe de son accessibilité.

Il s'agit donc de produire un patrimoine scientifique à l'aide de pratiques reposant sur les notions d'héritage, de respect et de préservation, et qui implique l'idée de propriété collective et de mise à disposition auprès de la communauté scientifique.

Le centre de compétences se propose ainsi de collaborer à la mise en place d'un « chantier des collections » pour une numérisation et une informatisation des fonds de ses partenaires. Il pourra s'articuler avec la mission aux archives scientifiques du Réseau national des MSH.

2.2 Enrichir le projet de sauvegarde des fonds d'un catalogage informatisé

- regrouper, localiser, documenter et contextualité les fonds par la production de métadonnées ;
- professionnaliser nos pratiques documentaires en alignant nos méthodes de travail sur les standards internationaux et en rendant les descriptions universellement compréhensibles par l'emploi de normes.

2.3 Valoriser les ressources et en assurer la diffusion

Il s'agit d'introduire dans le champ des sciences humaines et sociales des matériaux et des savoirs constitués par les chercheurs depuis plusieurs dizaines d'années et n'ayant été exploités que par eux-mêmes, d'en assurer l'accès, la circulation et le partage afin de favoriser l'émergence de nouvelles problématiques scientifiques. Cela permettra également de s'interroger sur la constitution de ces savoirs ; questionnement sur l'outillage des recherches, les méthodes et nécessaire retour sur le travail accompli :

- augmenter la visibilité et favoriser la localisation des fonds ;
- standardiser les ressources créées ;
- produire différents types de métadonnées normalisées pour des usages variés et l'échange des données ;
- promouvoir des services et des outils informatiques liés à la préparation des données en amont par les chercheurs et en aval pour leur diffusion.

2.4 Le centre de compétences : rouage de la constitution d'une archive institutionnelle sur les ressources numériques.

Les trois rôles qu'assurera le centre de compétences pourront être :

- fournisseur de contenus pour l'archive institutionnelle et fournisseur de services pour les « auteurs » des contenus ;
- organisation de la syndication des métadonnées avec la création d'un entrepôt OAI de certaines ressources numériques : entrepôt thématique et/ou disciplinaire ;
- participation à des projets coopératifs avec d'autres institutions favorisant l'articulation et la reconstitution virtuelle d'epi-collections similaires ou complémentaires à travers la mise en œuvre de silos de documents numériques et de catalogues.

3 Modalités

3.1 Production et gestion des ressources

Il ne s'agit donc pas pour nous de « révolutionner » les pratiques dans le domaine mais de suivre des principes et des méthodes ayant déjà fait leurs preuves.

L'interopérabilité, l'accessibilité et la pérennité des données constituent les bases du système d'information à mettre en place. Il est ainsi nécessaire de s'appuyer sur des codes de bonne conduite, l'utilisation de formats de données ouverts, de standards, de protocoles et de logiciels, libres, déjà éprouvés.

3.1.1 Repérer, recenser, classer les fonds

- récolement des fonds déposés chez les différents partenaires (voir liste en annexe) décrivant ce qui est possédé et son classement (s'il existe) ;
- travail documentaire et préparation matérielle et scientifique des matériaux avec l'aide de chercheurs. Il s'agira de constituer des lots et de réunir toute la documentation annexe (papiers préexistants, mentions sur les supports, les conditionnements ou les étiquettes, transcriptions, traductions, carnet de recherche…), permettant de classer, de renseigner et d'enrichir ces archives ainsi que d'éclaircir les problèmes de droits de diffusion ;
- expertise de la qualité physique des documents ;
- travail d'enquête et de repérage sur les archives en possession des chercheurs et non encore déposées (ayant ou non quitté les institutions scientifiques).

3.1.2 Processus de numérisation

- préparation physique des fonds ;
- choix des formats de fichiers de référence numérique et de diffusion (textes, audio, images, vidéo)
- création d'enregistrements de référence (*masters*), véritables archives numériques destinées à la conservation et à la réalisation de copies de qualité et à celle, de taille moins importantes, destinées à la diffusion ;
- premier travail de nommage des fichiers et de gestion des données numériques. Les masters sont documentés à l'aide d'une première série de métadonnées : données administratives, informations techniques utiles à la conservation à long terme (type de scanner, format du fichier, date de la numérisation, etc.) ou données liées à la gestion des droits d'accès aux documents.

3.1.3 Catalogage et pérennisation de l'archive

Un travail d'analyse documentaire sera réalisé en vue de l'élaboration d'instruments de recherche et de la mise en place de bases de données descriptives des fonds. Pour cela la production de métadonnées nous permettra de renseigner toutes les ressources disponibles de nos collections. Elles précisent le lieu de conservation, la cote, l'intitulé du fonds, le nom de la personne à l'origine du fonds, une présentation du contenu, les modalités de consultation et de reproduction. Pour les documents elles indiquent les participants (auteur, informateur, interprète, etc.), les lieux et dates des enquêtes, les durées, les formats, etc. Ce sont aussi ces métadonnées qui explicitent le lien entre les ressources (notice, fichier numérisé et annexes d'une archive).

La pérennisation de ces analyses est assurée par le choix d'un format de codage de ces métadonnées. Exprimées d'une manière normalisée, elles pourront être lues, recherchées et échangées.

Notre projet est fondé à la fois sur la norme internationale de description des archives ISAD(G)[1] et sur la norme EAD[2], dérivées de XML[3], pour l'administration et la localisation des enregistrements archivistiques. Standard mondial, indépendant des plates-formes techniques, ce vocabulaire contrôlé, employé pour le codage de métadonnées complexes et hautement structurées, est une bonne solution pour la description des fonds d'archives et des collections spécialisées. Il restitue et décrit des corpus documentaires complexes, caractérisés par l'hétérogénéité des supports qui les composent, comme peuvent l'être les archives de chercheurs, et intègre des liens vers les documents numérisés.

Ces normes et ces méthodes de travail permettent d'une part d'établir un cadre de classement adéquat au principe du respect des fonds des chercheurs et d'autre part

[1] International Standard for Archival Description (General). 2e édition du Comité international des archives : http://www.ica.org/biblio/isad_g_2e.pdf

[2] DTD EAD : Document Type Definition Encoded Archival Description, maintenu par la Bibliothèque du Congrès : http://www.loc.gov/ead/
Bulletin francophone de la Direction des Archives de France sur l'EAD : http://www.archivesdefrance.culture.gouv.fr/fr/publications/DAFbuldtd.htm

[3] *eXtensible Markup Language* - langage de balisage extensible

l'enrichissement du catalogue au fur et à mesure du traitement de ceux-ci. L'aspect évolutif de la production nous donne à voir rapidement les grandes sections et sous-sections jusqu'au document lui-même.

Afin d'assurer qualité et cohérence aux instruments de recherche électroniques ainsi créés, il est nécessaire d'élaborer un guide de recommandations pour les personnes encodant. Les règles DTD EAD sont génériques et pas toujours suffisantes dans un contexte archivistique précis. Il est indispensable avant de commencer l'édition des documents XML de définir des niveaux de contraintes et des recommandations supplémentaires :

- un schéma XML propre au projet (syntaxe détaillée des contenus, précisions dans l'usage des balises) ;
- des outils de vérification et de validation du document et du corpus (grammaire permettant de vérifier sa structure ou les liens hypertextes.) ;
- Il s'agit de mettre en place différents types de contrôle, de la contrainte formelle, en passant par l'incitation très forte et les règles de bonne pratique.

3.2 Exploitation et diffusion de ces ressources

Le centre de compétences pourra développer des services intégrant l'utilisation systématique des métadonnées. Ceci permettra entre autre d'améliorer leur qualité (recommandations, règles d'encodage) et de les réutiliser dans d'autres contextes (différents niveaux de granularité des métadonnées).

Trois types d'exploitation des ressources numériques seront développés en bénéficiant des avantages de la standardisation des métadonnées :

- pour les chercheurs/communauté (portail de ressources spécialisé, etc.) ;
- pour les documentalistes (constitution de ressources documentaires et pédagogiques) ;
- pour le public plus large (culture scientifique et technique).

3.2.1 Diffusion des instruments de recherche

Réalisation des outils de consultation permettant aux utilisateurs d'accéder à la description précise des fonds et des documents et de les localiser : mise en ligne d'instruments de recherche (catalogues, inventaires, index, moteurs de recherche, etc.)

3.2.2 Valorisation des ressources numériques :

- publication en ligne de recueils critiques de sources ;
- constitution d'e-atlas spécialisés (possibilité de partenariat avec d'autres centres travaillant sur les SIG) ;
- publications d'outils documentaires (répertoires thématiques de ressources liées, itinéraires au sein des ressources patrimoniales, dossiers thématiques spécialisés, bibliographies rétrospectives).

3.2.3 Assistance à la constitution de lexiques

Aide à la constitution d'un vocabulaire d'indexation contrôlé (thésaurus et ontologie) sur nos domaines d'études (Histoire des sciences et des techniques et Aires culturelles) en partenariat avec d'autres centres de compétences (linguistique par exemple).

3.2.4 Syndication de nos ressources patrimoniales

L'utilisation systématique du XML et des normes telles que l'EAD, nous permettra de constituer ou de participer à des entrepôts de ressources numérisées (utilisant le protocole OAI-PMH). La mise en place de flux RSS assurera, de façon complémentaire, la syndication des ressources patrimoniales traitées.

3.3 Groupe de travail et organisation d'ateliers technologiques

Le centre de compétences propose la mise en place de plusieurs types d'accompagnement :

- aide en amont aux projets chercheurs/groupes de chercheurs/communautés de chercheurs en matière d'informatisation des données de et pour la recherche (dépôt, normes/formats/standards/base de données/écriture web/diffusion) ;
- sensibilisation aux questions juridiques et déontologiques ;
- production de guides de pratiques.

4 Organisation

Maître d'œuvre : Centre Alexandre Koyré, Centre de Recherche en Histoire des Sciences et des Techniques (UMR n°8560 – CNRS/EHESS/Muséum National d'Histoire Naturelle/Cité des Sciences et de l'industrie)

Partenaires :
- UMR 8054 Laboratoire MALD (CNRS/Université de Paris 1 Panthéon Sorbonne)
- UMR 5185 ADES, Centre de Documentation REGARDS (CNRS/Université de Bordeaux 3/Université de Bordeaux 2)
- UMS 2267 CAPHES (Centre d'Archives de Philosophie, d'Histoire et d'Edition des Sciences - CNRS/ENS)
- UMR 7535, Laboratoire d'Ethnologie et de sociologies comparative, Bibliothèque Eric-de-Dampierre.

Partenaires réseaux :
- GIS RAFID
  - 24 équipes d'enseignement et de recherche (2 UMR) ; 19 bibliothèques et centres de documentation.
- GIS Réseau Amérique latine
  - 14 universités (5 UMR) ; le CNRS ; l'IRD ; l'EHESS ; Réseau européen REDIAL ; 14 bibliothèques et centres de documentation.
- Réseau Ethnologie
  - 4 bibliothèques et centres de documentation.
- Réseau Pléiade
  - Gallica ; CNUM/CNAM ; Médiathèque d'histoire des sciences de la Cité des Sciences et de l'Industrie.
- Réseau thématique pluridisciplinaire (RTP) Etudes africaines

Partenaires internationaux :
- REDIAL
  - 11 pays européens, 38 organismes membres

Annexe : Présentation des partenaires