Construire le web de données pour les sciences humaines et sociales

Stéphane Pouyllau, Shadia Kilouchi

Construire le web de données pour les sciences humaines et sociales

Fonds documentaires scientifiques de la recherche en SHS

Note informationnelle du Centre national pour la numérisation de sources visuelles

Stéphane POUYLLAU.
Ingénieur de recherche au CNRS,
créateur du CN2SV, TGE ADONIS.

Version 2.4 (12 septembre 2010)

Résumé :

Les fonds documentaires et plus largement les données sources pour la recherche dans les Sciences humaines et sociales (SHS) ont commencé à prendre le tournant du numérique, de plus en plus de données, servant à faire de la recherche en SHS, sont nativement numériques.

Il s'agit de mettre en œuvre une importante politique de conservation et de diffusion numérique des fonds, acquis la plupart du temps sur fonds publics depuis plus de 40 ans. Cela implique de la numérisation, de la redocumentarisation, de développer des accès multiples tout en assurant l'interopérabilité des données et en plaçant ces fonds dans le web de données. Depuis 2005, les choses ont évoluées dans le bon sens, mais l'informatisation des fonds documentaires reste faible, il s'agit maintenant d'atteindre une masse critique, de gérer un passage à l'échelle supérieure afin de positionner les données des SHS dans l'extension du web que sera dans quelques années le web de données.

Le web à 20 ans, il est devenu l'espace la diffusion des publications scientifiques (revues, sites, ouvrages, colloques, etc.) qui ont été les premières à l'utiliser comme vecteur de diffusion massive. Mais, pour le moment, il reste relativement vide de « données brutes » : il est temps d'inter-connecter les publications et fonds documentaires, mais aussi les fonds entre eux, afin de construire des espaces de données plus larges, mondiaux, dans le cadre de l'open access quand cela est possible afin d'offre aux scientifiques de demain des corpus de données numériques, documentés, accessibles, interopérables et pérennes.

Cette proposition expose le potentiel, les réalisations en cours et trace des perceptives opérationnelles et fonctionnelles pour la mise en œuvre d'un projet de conservation et de diffusion numérique des fonds documentaires et des données « brutes » de la recherche en SHS. Nous y avons ajouté des propositions concrètes en matière de consolidation du Centre national pour la numérisation de sources visuelle qui travaille, depuis 2005, avec plusieurs laboratoires et le TGE ADONIS.

Ce document est rédigée par Stéphane Pouyllau*, Ingénieur de recherche au Centre National de la Recherche Scientifique.

Elle est issue d'une note de réflexion sur les *digital humanities* développée en 2005 et remis à jour régulièrement en 2007, 2008, 2009 dans le cadre du CN2SV et en relation avec le TGE ADONIS, elle profite des réflexions et discussions de travail menées depuis 2005 avec Christine Blondel (CNRS, Centre A.-Koyré), Fabrice Melka (CNRS, Cemaf), Alain Michel (MCF, Université d'Evry), Shadia Kilouchi (contractuelle CNRS, CN2SV), Delphine Usal (CNRS, CN2SV), Gautier Poupeau (Ecole Nationale des Chartes, puis Atos Origin), Marin Dacos (CNRS, CLEO), Pierre Mounier (EHESS, CLEO), Richard Walter (CNRS, IRHT et TGE ADONIS), Robert Vergnieux (CNRS), Marie-Hélène Wronecki (Contractuelle CNRS, CN2SV), Yannick Maignien (CNRS, TGE ADONIS), Jean-Luc Minel (TGE ADONIS, MoDyCo - Université de Paris 10), ainsi qu'avec Alain Oguse (Interactives Solutions).

* Stéphane Pouyllau est ingénieur de recherche au CNRS, il est actuellement responsable du pôle *digital humanities* (SHS numériques) du Très Grand Équipement ADONIS. Il a co-fondé, en 2005, le Centre national pour la numérisation de sources visuelles (CN2SV) du CNRS dont il en assure la direction technologique, le management organisationnel et en coordonne les réalisations.

1) Introduction

Le web est l'un des vecteurs principaux de la diffusion des données de recherche en sciences humaines et sociales. Il permet de diffuser et d'éditer presque tout les matériaux utilisés par le chercheurs et l'enseignant : de la bibliothèque à la publication électronique en passant par le séminaire, le colloque, la revues et le livre. L'utilisation du web comme outil d'édition, de publication et de diffusion a permis de démultiplier les accès aux documents et à l'information. Depuis 20 ans, l'effort a plus porté sur la mise à disposition de documents numériques (ouvrages, articles, corpus) que sur la structuration de l'information contenue dans ces documents : il est vrai que l'essor des moteurs de recherche traditionnels depuis les années 90 (d'altavista à google) ont permis d'atteindre, petit à petit, ces milliards de documents qui compose le web aujourd'hui. En revanche, la publication électronique des contenus des bases de données, qui ont toujours leurs propres structurations, pose encore des questions et des difficultés qui font que le web, s'il est plein de documents et relativement vide de données et d'informations structurées. Ainsi, les outils d'exploitation des documents que nous utilisons aujourd'hui, tel les moteurs de recherche, fonctionnent sur des réservoirs de documents encore trop cloisonnés. Ainsi, construire une page web d'information sur l'historien Georges Duby nécessite toujours d'adresser plusieurs questions à plusieurs moteurs de recherche (généralistes et spécialisés), à plusieurs formulaires de bases de données et cela même si depuis dix ans les techniques de l'interopérabilité ont fait de très grand progrès. Actuellement les données numériques sont bien rangées dans de multiples bases de données, mais nous n'avons construit que de simples petits « judas » afin de les regarder. Pouvons-nous réellement les exploiter ? Pour aller plus loin, pour construire de nouvelles façon de faire de la recherche, il nous construire un web de données scientifique dans le web de données. Le web de données est une extension du web actuel, ou les données numériques sont structurées et liées entre elles. Il nous faut dépasser la simple interopérabilité des documents en eux. Comment faire ? Par ou commencer ?

A la différence des sciences exactes, le patrimoine des sciences humaines et sociales (SHS) reste encore largement inexploité. Seules les productions scientifiques, les publications au sens large, ont suscité des questionnements et des réalisations numériques (HALSHS, Revues.org, CAIRN, etc) et peu d'initiatives ont été lancées, à très grande échelle, concernant les fonds documentaires (carnets de terrains, photographies, données d'enquêtes, cartes et plans, croquis, campagnes de capture de données 2D, 3D, 4D). Ces fonds peuvent être collectés, classés et déposés : il s'agit alors de fonds d'archives. Mais la plupart n'entrent pas dans le circuit « archives », ils restent dans les bureaux des chercheurs, aujourd'hui dans leurs disques durs, dans les centres de documentation et bibliothèques de recherche : beaucoup sont toujours utilisés par les chercheurs, comme sources de recherches actuelles. Dans ce cas, ils composent des fonds documentaires scientifiques à forte dimension patrimoniale qui, dans le futur, représentera une richesse pour l'institution qui les aura élaborés, financés, numérisés, et diffusés[1]. Les départs en retraites massifs, qui sont en cours, vont avoir pour effet, au mieux une dispersion de ces fonds dans les familles des chercheurs, au pire une destruction des données.

1 Nous renvoyons le lecteur à la note d'Isabelle de Lamberterie : Identification, traitement, conservation et mise à disposition des archives « scientifiques » en SHS, 10 avril 2009 et au document des Centres de ressources numériques du CNRS : *Travaux de prospective 2008 des centres de ressources numériques CN2SV, CRDO et TELMA* : http://www.cn2sv.cnrs.fr/spip.php?article101

A ces fonds documentaires anciens s'ajoutent aujourd'hui des données nativement numériques (photos, textes, emails, données 2D, 3D, sons, flux RSS, etc.), également constituées en fonds, mais dont l'existence et l'accessibilité ne tient qu'aux outils actuels, qui les ont – la plupart du temps – générés. L'obsolescence des formats, supports, codage de nos données numériques est une réalité qui entraine actuellement une perte de données, une altération des fonds, mémoire futur et sources de nos recherches actuelles. Les fonds de données nativement numériques sont de plus en plus importants, fréquents, et cela dans de très nombreuses disciplines des SHS. Ces données numériques sont parfois l'unique capture d'un objet archéologique dont la conservation hors de son contexte naturel ne peut être totalement garantie[2].

Ces fonds documentaires, « archives de la science », ont fait l'objet depuis 2005 d'études : une campagne de réalisation d'un inventaire (projet ARSHS, GIS MSH) a été lancé[3], avec le soutien, en 2008, du Très grand équipement ADONIS, des plateformes technologiques – dans les unités mixtes de recherche, dans les maisons de sciences de l'homme, etc – ont lancé des projets de numérisation, mais l'effort n'est pas encore suffisant pour atteindre une masse critique, offrant un méta-corpus suffisant pour faire entrer les données des SHS dans l'extension du web que sera dans les années qui viennent le web de données. La mémoire de la recherche, autrefois analogique et sur support physique, est aujourd'hui massivement numérique et dématérialisés et la courbe de progression est exponentielle.

Que font les enseignants-chercheurs en archives et sur le terrain aujourd'hui ? Ils utilisent le numérique dans presque toute la chaine de la recherche : photographient en numérique les manuscrits, les transcrivent à l'aide d'un traitement de texte, en font des dossiers pour leurs étudiants via des ENT[4], certains commencent à diffuser des corpus de données via le web, les bibliothèques publiques le font déjà[5], etc.

Ainsi, Il s'agit de faire entre les fonds documentaires et les fonds d'archive des scientifiques dans le web de données.

2) Enjeux et objectifs : mise en œuvre et animation d'un projet de numérisation, de re-documentarisation, d'enrichissement des données sources de la recherche, dans le cadre des SHS et en perceptive de l'émergence du web de données (*web of data*).

Il s'agira de définir l'architecture principale, combinant veille technologiques, recherche et développement sur la problématique du web de données, veille patrimoniale, conservation et traitement des fonds documentaires, archives et de collections, services liés à la recherche et diffusion des ressources numériques dans le cadre du web de données.

La place des fonds documentaires scientifiques dans le web de données est stratégique, elle positionne le patrimoine scientifique français au cœur des dispositifs d'interopérabilité et des échanges de données de la connaissance, moteur de la recherche. Actuellement, les données sont le plus souvent stockées dans les bases de données relationnelles qui peinent à

2 Cf. : Virtual Retrospect 2003, *the bucket chain of Barzan (Charente-Maritime): mechanical engineering*, S. Coadic, L. Espinasse (Institut Ausonius, Institut de Recherche sur l'Antiquité et le Moyen Age. Unité Mixte de Recherche, CNRS – Université de Bordeaux 3, France), http://archeovision.cnrs.fr/fr/publication.htm

3 Projet ARSHS : http://constel07.u-bourgogne.fr:8080/sdx/pl/generic-subset.xsp?type=collections&id=cat

4 Environnements numériques de travail.

5 Voir : http://www.flickr.com/commons

s'ouvrir : si *l'open source* s'impose peu à peu, les choix de modélisation et les formats de méta-données sont peu documentés, les référentiels sont encore encapsulés dans les outils et les finalités principales de ces outils de gestion de BDD restent très souvent exclusivement tournée vers la publication scientifique. Malgré les politiques d'*open access*, les techniques, pourtant mûres, et principe de l'interopérabilité (OAI-PMH, Dublin Core, METS, etc.) les données sont souvent « enfermées » dans un carcan documentaire reproduisant les méthodes analogiques de classement, de catalogage, d'indexation dans l'univers du numérique[6]. Le web de données ouvre une nouvelle voie : celle où les données sont liées entre elles (*linked data*) et lisibles entre machines grâce à une grammaire commune, le RDF (*Ressource description format,* standard du W3C) qui exprime les données sous forme d'un triplet (sujet, prédicat, objet). Les données sont aussi liées à des référentiels, publics, ouverts et exprimés eux même en RDF. Le développement du web de données pour les données patrimoniales scientifiques nécessite :

- Un grand projet de numérisation et redocumentarisation des fonds documentaires par production de métadonnées s'appuyant sur des standards internationaux ;
- Une volonté politique d'ouverture des bases de données constituées sur fonds publics ;
- Une formation des personnels aux méthodes, techniques, outils, de l'interopérabilité des données ;
- Une organisation des projets en réseau avec un pilotage national scientifique et technique (ces deux notions sont fondamentales et indissociables).

6 Voir : http://www.lespetitescases.net/carcans-de-la-pensee-hierarchique-et-documentaire-1 et http://blog.stephanepouyllau.org/diffusion-et-edition-de-bases-de-donnees-partie-1

Nous présentons ci-dessous des éléments issus du projet de fondation du CN2SV dont les auteurs sont Fabrice Melka, assistant ingénieur au CNRS, Daniel Pouyllau, Ingénieur de recherche au CNRS, Stéphane Pouyllau, Ingénieur d'étude au CNRS. Ces éléments ont été réactualisé et sont toujours d'actualités.

3) Contexte

3.1) Des fonds documentaires dans un monde numérique

Il n'existe ainsi ni réglementation, ni réel dispositif permettant la collecte et la conservation des matériaux documentaires (données brutes, corpus, collections, reportages) issus de l'activité de recherche proprement dite[7]. Ces fonds sont beaucoup plus fréquemment conservés dans les laboratoires, les centres de documentation et les bibliothèques de recherche, voire dans les familles, que dans des services d'archives. Il en va de même pour les données numériques.

Ces fonds sont de deux natures principales : les données collectées par les enseignants-chercheurs dans le cadre de leurs activités et les données collectées lors de programme de recherches d'équipes et dont la collecte a été le plus souvent financée par les laboratoires de recherche eux-même et ce depuis les années 60.

Ils constituent des réservoirs de données dont certains sont en cours de numérisation et d'informatisation mais dont le nombre est encore trop faible pour atteindre une masse critique permettant de lancer de nouveaux projets de recherche et des actions de redocumentarisation scientifiques de masse associant scientifiques et métiers de la documentation, des bibliothèques et de l'information scientifique et technique. En effet, l'accent a été plutôt donné dans la mise en place d'inventaires numériques signalant les données, qui sont nécessaires mais ne donnant pas réellement accès (même de façon limité et sécurisée pour les données ayant des restrictions d'accès) aux données elles-mêmes.

Cependant, entre 2005 et 2009, plusieurs structures ont pris en charge le développement des conditions de la mise en place dans le web de données des fonds documentaires SHS. Ainsi, le CNRS dispose de l'ensemble du support pour lancer une opération de masse.

3.2) Le réseau du Très grand équipement ADONIS, le savoir-faire des Centres de ressources numériques (CRN), les plateformes technologiques des unités mixtes de recherche et Maisons des sciences de l'Homme.

L'action du TGE ADONIS depuis 2007 sous l'impulsion de Yannick Maignien (directeur du TGE de mars 2007 à août 2010), relayée par les CRN – lancés en 2005 par le département SHS et la Direction de l'information scientifique – a permis de créer une dynamique, de mettre en place des outils et des procédures offrant un environnement stable pour supporter, coordonner, réaliser des opérations de redocumentarisation des fonds documentaires scientifiques des SHS. Ce réseau, informel pour le moment, a lancé plusieurs actions de formation et diffusion des bonnes pratiques en matière de production, d'informatisation, d'enrichissement des méta-données, d'édition de corpus et d'archivage

7 Voir la note : Identification, traitement, conservation et mise à disposition des archives « scientifiques » en SHS, 10 avril 2009.

pérenne des données numériques. Ces actions de formation ont deux ambitions : l'appropriation des méthodes, outils, bonnes pratiques par les personnels de l'enseignement supérieur et de la recherche et assurer des échanges réguliers entre les membres des CRN, des plateformes technologiques, des réseaux de documentalistes dans le cadre des humanités numériques. Une école thématique CNRS/SHS/CRN et une Université d'été du TGE ADONIS en 2008, une université d'été sur l'édition électronique en 2009, portée par le Centre pour l'édition électronique ouverte.

Depuis 2005, plusieurs disciplines se sont engagées massivement dans la numérisation de données de fonds documentaires, d'archives de science et de scientifiques. Nous prendrons l'exemple de l'histoire des sciences, de l'archéologie, des aires culturelles et de la géographie[8].

a) L'histoire des sciences et l'archéologie

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 spécialisée 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 bibliothèque numérique de la Cité des sciences et de l'industrie, proposent des ressources numériques en ligne. C'est dans ce domaine qu'a été l'un des 5 centres de ressources numériques : le Centre national pour la numérisation de sources visuelles[9] qui, avec le centre TELMA[10] (autre centre de ressources, IRHT-ENC) développe des corpus de sources numérisées (www.lamarck.science.gouv.fr, www.ampere.science.gouv.fr, www.buffon.science.gouv.fr, www.lavoisier.science.gouv.fr etc.), des plateformes de diffusion de fonds documentaires et d'archives scientifiques (www.cn2sv.fr/corpus) et (http://www.cn-telma.fr).

L'archéologie, au sens large du terme, a également développé la numérisation de masse des données de la recherche, la plateforme ArchéoVision[11] (UMR Ausonius) par exemple développe, dans le cadre du réseau du TGE ADONIS, un conservatoire des données 3D et dans le cadre de la « grille ADONIS »[12] elle développe une plateforme de corpus iconographiques en ligne avec accès contrôlé et sécurisé aux données. Les travaux lancés par l'ENS-Ulm sur la géo-localisation des données de fouilles (SIG) permettent de développer aujourd'hui des outils de géo-localisation en ligne, ils sont très proche des problématiques SIG développé à la MSH de Besançon et à l'UMR Ausonius de Pessac (Université de Bordeaux 3, CNRS). Cependant, la masse des fonds est très importante et il reste de très nombreux fonds, certains très anciens, qui pourraient être éligibles.

Ces deux domaines, différents sur le plan scientifique, ont aujourd'hui de nombreux points communs, grâce à la dissémination des bonnes pratiques en matière d'informatisation

8 Cette note ne peut être totalement complète sur les projets en cours, il s'agit d'avoir une idée globale s'appuyant sur des exemples précis et des réalisations fortes.
9 Voir : http://www.cn2sv.fr
10 Voir : http://www.cn-telma.fr
11 Voir : http://archeovision.cnrs.fr
12 Voir : http://www.tge-adonis.fr/?TGE-ADONIS-les-SHS-se-dotent-d-une

des données et d'appropriation, par leurs équipes techniques, des pratiques de l'*open access*, des protocoles tel que OAI-PMH, des schémas de méta-données tel que *Dublin Core*, etc.

b) Aires culturelles et géographie

Le patrimoine scientifique sur les aires culturelles nous semble avoir sa propre spécificité, d'une part grâce à la place importante qu'y occupe la variété des matériaux de terrain (notes, carnets, productions sonores, iconographiques ou audiovisuelles), d'autre part, hormis un réel effort 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 de laboratoires, 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.

Plusieurs actions ont été lancées, certaines avec le soutien en 2008 du TGE ADONIS et dont certaines pourraient tout à fait être hébergées sur la « grille ADONIS » : portail plozévet, ODSAS[13], Portail des Études Africaines, Fonds cartographies du CEGET[14], etc. Ces projets ont mis en œuvre des mécanismes permettant l'interopérabilité des données, dans le respect des droits de diffusion, il s'agit là d'une première marche vers le web de données.

Le contexte est donc favorable pour le lancement d'un projet massif :
- prise de conscience de la fragilité des fonds documentaires ;
- développement des centres de ressources « au contact des fonds » et des chercheurs (y compris des projets ANR) ;
- développement d'un réseau d'opérateurs inter-institutionnel autour d'un TGE, porté par le CNRS, qui partage les méthodes, techniques, du web de données ;
- des actions de formation (Écoles thématiques, universités d'été) permettant la formation des personnels.

4) Modalités

4.1) Nécessité de la numérisation.

Ces fonds documentaires 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 « brutes » pour la recherche ;
- dépendance de ces données vis-à-vis de leurs producteurs (formats, annotations, localisation…).

Autant de raisons pour se préoccuper dès maintenant de la transmission de ces

13 Voir : http://www.odsas.fr/
14 Voir : http://www.cn2sv.cnrs.fr/spip.php?article76

données et connaissances qu'elles datent d'hier ou d'aujourd'hui. La numérisation des fonds doit être sous-traitée, le CNRS et ses unités n'ont pas vocation à numériser de grand volume de données. En revanche, ils devront assurer le suivi de cette réalisation et assurer la redocumentarisation des données. Les CRN, les plateformes des MSH pourront assurer le suivi, l'expertise et la gestion de cette phase, mais uniquement avec un renfort important en RH et/ou financier.

4.2) Nécessité d'un espace de stockage pérenne pour les données : la grille ADONIS, du stockage à l'archivage pérenne des données numériques.

La « grille ADONIS » est une infrastructure informatique mais aussi humaine, mise en œuvre par le TGE ADONIS et financée par le plan de relance gouvernemental et le CNRS qui offre :
- des espaces web pour le développement de sites web de corpus scientifiques numériques ;
- des espaces de stockage pour de grandes masses de données (200 Téra-octets disponibles) ;
- une offre d'archivage pérenne des données numériques fondée sur l'*Open Archival Information System* (norme ISO 14721:2003) réalisée avec le CINES, le CC-IN2P3. Ce projet pourrait entrer, après validation, dans le cadre de l'instruction du 15/01/2007 sur les archives cosignée entre le CNRS et la Direction des Archives de France.

Totalement opérationnelle depuis décembre 2009, la grille ADONIS offre une base de travail, une infrastructure informatique, permettant l'hébergement d'un projet de numérisation et redocumentarisation. Le CN2SV l'utilise depuis 2009.

4.3) Accessibilité

L'accessibilité des fonds numérisés ou numériques doit tenir compte pour les fonds documentaires analogiques :
- de l'environnement de production des fonds ;
- des contraintes de confidentialité en matière de diffusion.

Pour les données directement numériques, ces deux questionnements sont également présents. Avec l'apparition des licences Creative Commons[15] ou le projet Science Commons, il serait possible d'inciter fortement les chercheurs à mettre leurs corpus sous ce type de licence. Les données acquises sur fonds publics (y compris le salaire du chercheur, pour les collectes réalisées hors actions spécifiques) pourrait être par exemple sous licence Creative Commons : BY-NC-SA (http://creativecommons.org/licenses/by-nc-sa/2.0/fr/). Dans cette optique, le CN2SV a inclus ces licences dans la plateforme MédiHAL (archive ouverte de photographies scientifiques, basée sur HAL et en cours de test) réalisée avec le Centre pour la communication scientifique directe et avec le soutien du TGE ADONIS.

L'accessibilité passe par le signalement des fonds, le projet ARSHS (GIS MSH), signalant et décrivant les fonds anciens, de science et de scientifiques, pas forcément numérisés, la plateforme NUMES, signalant les projets et corpus numériques en cours mais aussi les équipes et institutions portant les projets. NUMES diffuse aussi les informations de

15 Voir : http://fr.creativecommons.org/ et http://creativecommons.org/

signalement sous la forme d'un entrepôt OAI-PMH, donc interopérable avec des portails nationaux, européen (Projet Michael), etc. Le CN2SV travaille avec ces deux programmes afin de signaler les fonds qu'il propose.

L'accès aux données doit être multiple, ouvert, exploité par les portail thématiques (thématiques, discplinaires, régionaux), par des moteurs de recherche spécialisés et des moteurs de recherche généralistes. Par l'accès, il s'agit d'introduire dans le champ des sciences humaines et sociales des matériaux numériques en masse et des savoirs « numérisés » 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 méta-données 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.

L'accessibilité nécessite la mise en place d'une recherche et développement importante et pas uniquement sectorisée sur les SHS, elle doit s'étendre aux STIC, à l'informatique appliquée, à la documentation et au monde des bibliothèques. Cette dimension R&D est stratégique car elle permettra de développer, en France, une maitrise des technologiques de diffusion massive de données numériques issue des SHS, et de faire entrer les SHS dans l'ère du web de données, dans l'ère « du tout numérique » qui est aujourd'hui incontournable. Il s'agira tout en développement des accès unifiés, utilisant les technologiques actuelles et pragmatiques, de travailler sur de nouvelles solutions d'accessibilité.

4.4) Processus et démarche

Il s'agit 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. Ceci entraine et 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 et collabore dans un espace commun de référentiels, de standards et normatifs déjà en place pour la plupart. 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 actuels, autour du réseau formé au sein du TGE ADONIS[16], se connaissent et partagent le même intérêt pour la numérisation des fonds, la conservation des fonds et archives de sciences et de scientifiques. Par ailleurs ils ont déjà eu pour certains l'occasion de travailler ensemble sur des projets similaires (ACI, ANR,

16 Qui regroupe des chercheurs du CNRS, des enseignants-chercheurs des Universités, des ingénieurs et techniciens.

etc.). 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 et sur le rôle de relais des partenaires de ce projet auprès de réseaux scientifiques plus larges portés par les institutions de l'enseignement supérieur et de la recherche.

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 des fonds assurant la cohésion du traitement de ces données.

Pour cela, la production de méta-données 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 et surtout les méta-données qui explicitent le lien entre les ressources (notice, fichier numérisé et annexes d'une archive) qui seront au cœur d'une projet de numérisation suivant les règles du web de donnée. La notion de *linked data*[17] ou données inter-connectées est l'un des enjeux de ce type de projet : il est possible de construire des réservoirs de données hétérogènes – devant aller plus loin que des bases de données documentaires inter-connectées - et de tisser des liaisons (documentaires, référentielles et sémantiques) entres elles par l'utilisation de modèle de données exprimées en RDF. Ce type de projet doit également s'appuyer sur l'utilisation de schémas de méta-données standardisées tel que le *Dublin Core Terms* (plus riche que le *DC Element Set*). La notion d'interopérabilité des données[18] est le point central dans l'introduction de la problématique du web de données pour les données et fonds documentaires des SHS, elle en est la première marche, l'étape la plus importante ou le suivit et l'accompagnement doit être maximal.

5) Conclusion

Les fonds documentaires et plus largement les données sources pour la recherche dans SHS ont commencé à prendre le tournant du numérique, de plus en plus de données, servant à faire de la recherche en SHS, sont nativement numériques, il s'agit de mettre en œuvre une importante politique de conservation et de diffusion de ces fonds, acquis la plupart du temps sur fonds publics depuis plus de 40 ans. Cela implique de la numérisation, de la redocumentarisation, de développer des accès multiples tout en assurant l'interopérabilité des données et en plaçant ces fonds dans le web de données.

Depuis 2005, les choses ont évoluées dans le bon sens, mais l'informatisation des fonds documentaires reste faible, Il s'agit maintenant d'atteindre une masse critique, de gérer un passage à l'échelle supérieure afin de positionner les données des SHS dans l'extension du web que sera dans quelques années le web de données.

Le web à 20 ans, il est devenu l'espace la diffusion des publications scientifiques (revues, sites, ouvrages, colloques, etc.) qui ont été les premières à l'utiliser comme vecteur de

17 Voir, sur le site du W3C, les travaux de Tim Berners Lee : http://www.w3.org/DesignIssues/LinkedData.html

18 Voir : POUYLLAU, S., De l'interopérabilité au web de données, http://blog.stephanepouyllau.org/de-linteroperabilite-au-web-de-donnees

diffusion massive, dépassant les ambitions initiales de Tim Berners-Lee, son inventeur, qui s'engage aujourd'hui dans la mise en place du web de données[19]. Mais le web reste vide de « données brutes », il est temps d'inter-connecter les publications et les fonds documentaires, et les fonds entre eux, afin de construire des espaces de données plus larges, mondiaux, ouvert, libre d'accès (dans la mesure du possible) afin d'offre aux scientifiques de demain des corpus de données numériques, documentés, accessibles, interopérables et pérennes.

19 Voir la présentation de TBL aux conférences TED : http://www.ted.com/talks/lang/eng/tim_berners_lee_on_the_next_web.html