Applications composites et digital humanities : méthodes et outils du CN2SV au service de la recherche

Stéphane Pouyllau, Daniel Pouyllau, Shadia Kilouchi, Mona Huerta

Applications composites et *digital humanities* : méthodes et outils du CN2SV au service de la recherche

Stéphane Pouyllau, ingénieur d'études au CNRS[1].

Le développement des *digital humanities* (DH) dans la recherche en sciences humaines et sociale est en progression constante depuis quelques années en France. Les centres nationaux de ressources numériques (CNRN), créés par le CNRS en 2006, ont labelisé des équipes qui ont pris le virage des humanités numériques au tournant des années 2000. Les CNRN, mais également un grand nombre d'équipes (créées autour de bibliothèques, de centres de documentation, ou d'équipe faisant de l'informatisation des données) couvre de territoire et réalisent ou coordonnent des opérations de numérisation et de traitement de données numériques natives ; elles décrivent, classent, créent des outils et diffusent des corpus de sources numérisées ou numériques. De nombreux acteurs de l'information scientifique et technique et des DH utilisent, depuis quelques années, les recommendations du modèle de gestion des données OAIS. De nombreux chercheurs, aujourd'hui réunis autour de projets collectifs, utilisent les méthodes, techniques et outils issus des DH : revues électroniques, collections de document en ligne, outils collaboratifs d'aide à la recherche (du simple forum à l'outil d'annotation d'images ou de séquences vidéo).

Le Centre National pour la Numérisation des Sources Visuelles (CN2SV), qui est un CNRN, travaille depuis 2006 à la mise en place d'une plateforme de publication d'inventaires de fonds d'archives scientifiques : Arch@CN2SV. Une première version de cette plateforme fonctionne et s'appuie sur les normes en vigeur dans ce domaine : L'*Encoded Archival Description* (EAD) pour la structuration des inventaires électroniques, schéma XML lui-même conforme à la norme des archives ISAD(G), DublinCore pour l'exposition de métadonnées sur le web et OAI-PMH pour l'interopérabilité. Arch@CN2SV utilise exclusivement des briques technologiques libres et dont le code est ouvert (*open source*) : PLEADE 2 pour l'application web, SDX pour le cadre de développement (*framework*), MySQL pour le stockage des fragments XML issus de la publication des inventaires. L'ensemble de ces travaux s'inscrit dans une dynamique nationale de mise en place du modèle de gestion des données OAIS et dont le très grand équipement ADONIS[2] assure la mise en oeuvre pour les DH au CNRS.

Cette plateforme est l'une des briques technologique du CN2SV. Elle permet d'assister les équipes de recherche, les bibliothèques et les centres de documentation en matière d'informatisation des données pour la recherche, ce que les chercheurs appellent plus simplement : « les sources ». Les documents d'archives, une fois traités (par un archiviste ou un documentaliste) deviennent des sources « utilisables » par les chercheurs, elle peuvent être citées et éditées. Le CN2SV s'attache à l'informatisation de sources à caractère iconographique : photographies, diapositives, manuscrits illustrés, carnets de terrain illustrés, cartes et plans. Dans ce cadre, deux fonds ont été traité en 2007 : le fonds d'archives scientifique d'Alexandre Koyré (déposé à la bibliothèque du Centre A.-Koyré/CRHST[3]) et un fonds de cartes historiques couvrant Madagascar issue de la cartothèque du centre de documentation REGARDS[4].

---

1 Responsable de la plateforme technologique du Centre de recherche en histoire des sciences et des techniques. Cette plateforme assure la maîtrise d'oeuvre du Centre national pour la numérisation de sources visuelles.

2 Voir : http://www.tge-adonis.fr [d.c. : 05/06/08]

3 Unité mixte de recherche n°8560 du Centre national de la recherche scientifique, de l'Ecole des hautes études en sciences sociales, de la Cité des sciences et des sciences et de l'industrie et du Muséum national d'histoire naturelle.

4 Le Centre de Documentation REGARDS a été créé en 1968, il comprend une Bibliothèque de recherche, une Cartothèque-Photothèque et un système de Bases de données. REGARDS assure aussi le développement technologique des bases de données et sites web du GIS RAFID-Formation et Information pour le Développement, du GIS Réseau Amérique Latine et du réseau européen REDIAL. Il dépend de l'unité mixte de recherche ADES regroupant des équipes du Centre national de la recherche scientifique, de l'Université de Bordeaux 3, l'Université de Bordeaux 2.

1) Une plateforme de publication d'inventaires numérique d'archives : la naissance d'un instrument de recherche

Dans le monde des archives, les instruments de recherche sont des objets très utilisés : par les chercheurs et par les archivistes eux même car ils permettent aussi le plus souvant la gestion d'un fonds d'archives ou de manuscrits. Pour les chercheurs ces objets sont souvent assez transparents, il est normal d'avoir accès à ces instruments de recherche au même titre qu'il est normal, aujourd'hui, de faire une recherche à l'aide de Google, Yahoo, d'un moteur de recherche spécialisé d'une revue électronique ou dans le moteur de recherche plein texte de HAL[5]. Depuis plusieurs années maintenant, le monde des archives[6] a développé, en France notamment et dans le monde universitaire[7], des instruments de recherche électroniques utilisant EAD qui s'est rapidement imposé comme une grammaire XML puissante, permettant de créer des inventaires électroniques de fonds d'archives possédant plusieurs entrées (matérielles, intellectuelles, etc.) et ayant, potentiellement, plusieurs niveaux de description. Un fichier EAD doit alors être considéré comme une grille (dans le sens informatique de *grid*), car il peut être lié à un ou plusieurs autres fichiers EAD ou EAC[8]. L'exploitation d'un inventaire écrit en EAD peut se faire de plusieurs façon :

- Avec une publication statique par l'utilisation de feuille de style XSL et de processeur XSLT. Il est possible d'utiliser pour cela des langages de programmation puissants tel que PHP (avec SAX ou autres)
- Avec une plateforme logicielle en ligne : un wogiciel (de *w*eb et l*ogiciel*) prenant la forme d'une application web adossée à un système de gestion de bases de données relationnelles.

Le CN2SV a choisi la seconde solution en raison des possibilités de couplage avec d'autres applications web qui ont pu être développées par le passé et sont déjà “communicantes” : c'est à dire inter-opérables grâce à des connecteurs XML.

La plateforme Arch@CN2SV s'intègre aussi dans une chaîne de publication d'inventaires EAD. Plus largement, le CN2SV aide les équipes ayant des fonds d'archives a produire des inventaires EAD tant sur le plan méthodologique que sur l'assistance à maîtrise d'oeuvre. Ces inventaires peuvent ensuite rejoindre le dépôt du CN2SV et/ou être publiés sur Arch@CN2SV après validation par l'équipe du CN2SV suivant le schéma suivant :

5 Voir : http://hal.archives-ouvertes.fr [d.c. : 03/06/08]
6 Voir : http://www.archivesnationales.culture.gouv.fr/chan/ [d.c. : 05/06/08]
7 Voir : http://www.calames.abes.fr/pub/ [d.c. : 05/06/08]
8 Pour *Encoded Archival Context*, schéma XML permettant de rédiger des notices bio/bibliographiques.

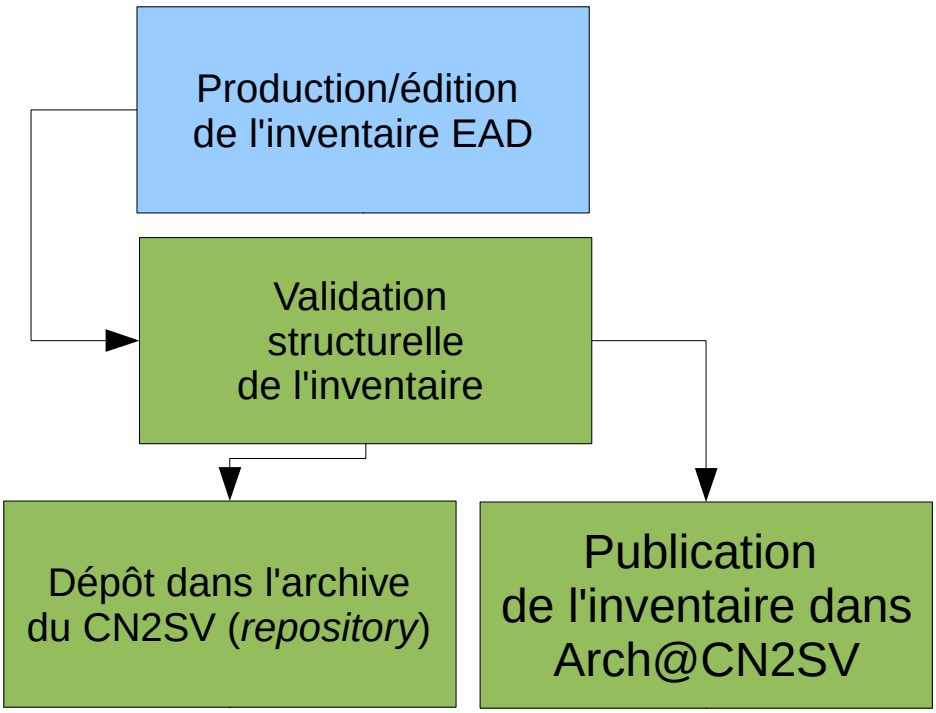

fig. 1 : schéma fonctionnel du versement dans la plateforme du CN2SV

La production/édition de l'inventaire peut être réalisée par les équipes possédant les fonds. Il existe de nombreux éditeurs XML permettant l'encodage EAD. Certains, comme XMLMind peut être équipé de feuilles de style XSL permettant une saisie au travers de formulaires de type web. L'EAD est cependant une grammaire complexe nécesistant une formation minimale. Il existe d'ailleurs des formations spécialisées, tel que celle proposée par l'Ecole nationale des chartes : le master 2 « Nouvelles technologies appliquées à l'histoire » qui comprend des enseignements spécifique au XML dans une optique de gestion de fonds d'archives[9].

2) Production d'inventaires de recherche à partir d'inventaires EAD : des applications composites (*mashup*) au service de la recherche.

En 2007 et dans le cadre d'un partenariat avec le CN2SV, le Centre de documentation REGARDS-ADES a réalisé un instrument de recherche[10], publié sur Arch@CN2SV, donnant accès à un inventaire EAD de cartes anciennes de Madagascar. Ces cartes ont été numérisées et des connecteurs XML, génériques et offrant la possibilité de géo-localiser ces cartes numérisées sur Google Maps, ont été développés :

9 Voir : http://www.enc.sorbonne.fr/master-nouvelles-technologies-appliquees-a-l-histoire.html [d.c. : 04/06/08]
10 Voir : http://www.cn2sv.cnrs.fr/corpus/ [d.c. 05/06/08]

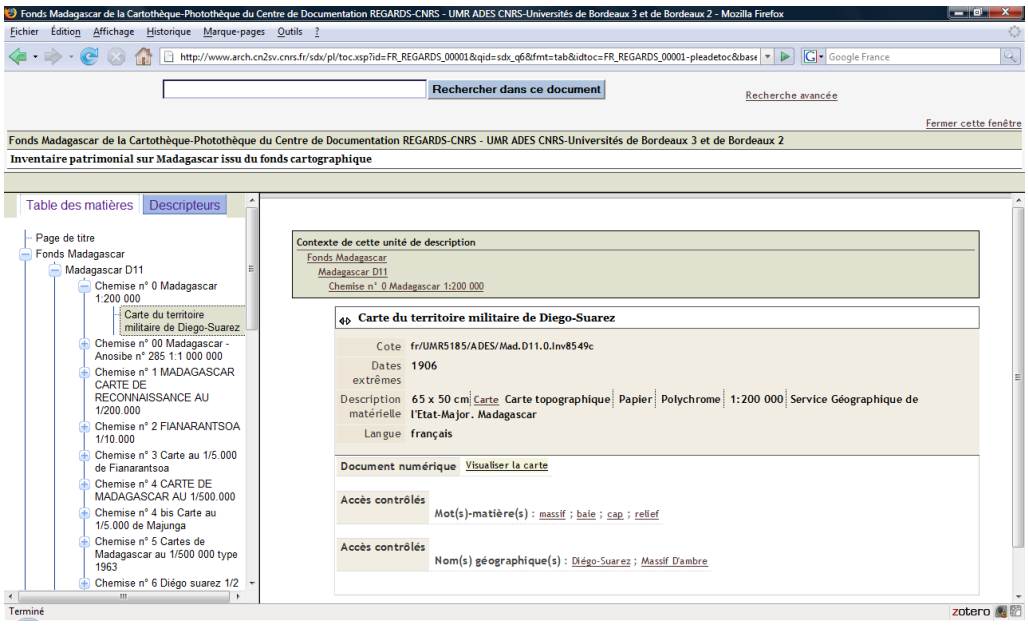

fig. 2 : instrument de recherche sous Arch@CN2SV

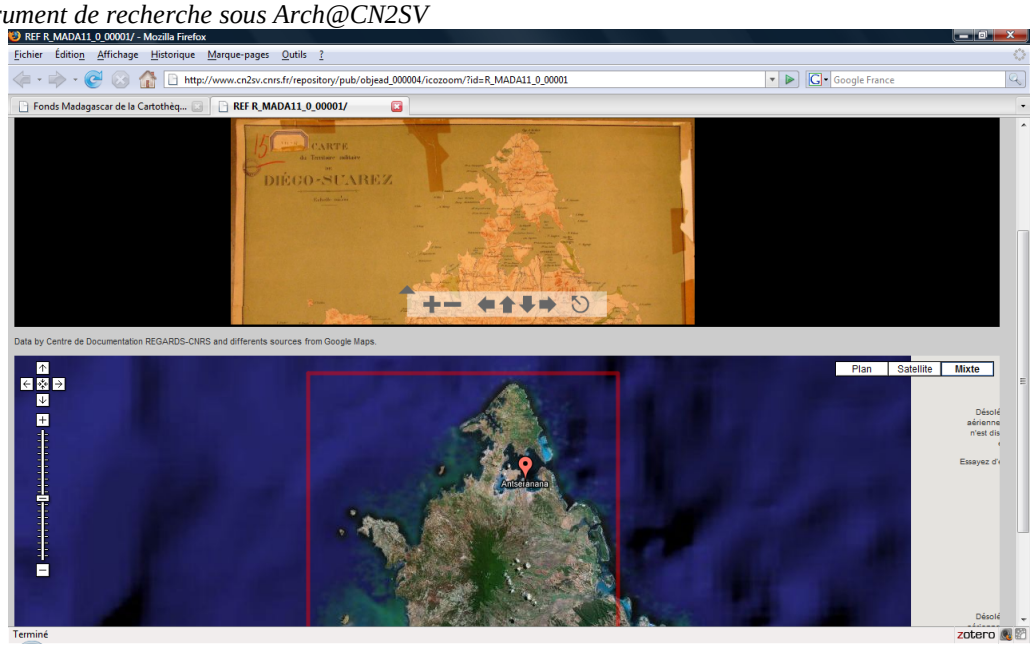

fig. 3 : géo-localisation de la carte ancienne sous Google Maps. Il est possible d'utiliser les marqueurs Google pour interroger des bases de données documentaires ou bibliographiques ayant des données en liaison avec la carte.

Cet instrument de recherche, qui sera complété en 2008-2009 utilise les schémas XML suivant :

- L'EAD pour l'inventaire des cartes
- GeoDataXml, schéma XML permettant le stockage des coordonnées géographiques
- KML, schéma XML lisible dans Google Earth et Google Maps (pour certaines balises)

Les notices EAD sont connectées aux cartes numérisées à l'aide d'une application PHP créée par le CN2SV. Cette application[11] affiche la carte ancienne (traitée en tuiles XML avec Zoomify[12]) et positionne les données GeoDataXml[13] dans une fenêtre Google Maps au travers de l'API Google à l'aide d'un processeur XSLT (*XSLTProcessor* sous PHP5). Les données

11 Le code est disponble, en version alpha, sur : http://www.hstl.crhst.cnrs.fr/doc/wiki/index.php/GeoDataXml [d.c. 05/06/08]

12 Voir : http://www.zoomify.com/ [d.c. 05/06/08]

13 Les données permettent de dessiner un cadre de couleur, représentant la surface de la carte ancienne sur Google Map. Il serait également possible de « couvrir » Google Map avec la carte. Mais le manque de précision de Google Map en matière de géo-localisation entrainerait des décalages trop important.

GeoDataXML offre la possibilité de placer des marqueurs dans la fenêtre Google Maps et de les connectes eux-même à une ou plusieurs bases de données documentaires ou bibliographiques. Nous sommes là dans une chaîne logicielle utilisant toute la puissance d'XML. Basée sur PHP, cette application sera disponible en ligne avec un guide de mise en place sur le site du CN2SV dans quelques mois.

D'autres bibliothèques ont fait appel au CN2SV pour la réalisation et la publication d'instruments de recherche EAD. La bibliothèque du Centre Alexandre-KOYRE (EHESS) a réalisé un inventaire électronique des archives de l'historien et philosophe Alexandre Koyré (1892-1964)[14] et propose une sélection de documents numérisés en haute définition. Le CN2SV a créé des connecteurs génériques et réutilisables, une fois encore à base d'XML et permettant de lier les documents numérisés à l'inventaire. Cet inventaire est disponible en ligne[15] sur le site du Centre A.-KOYRE.

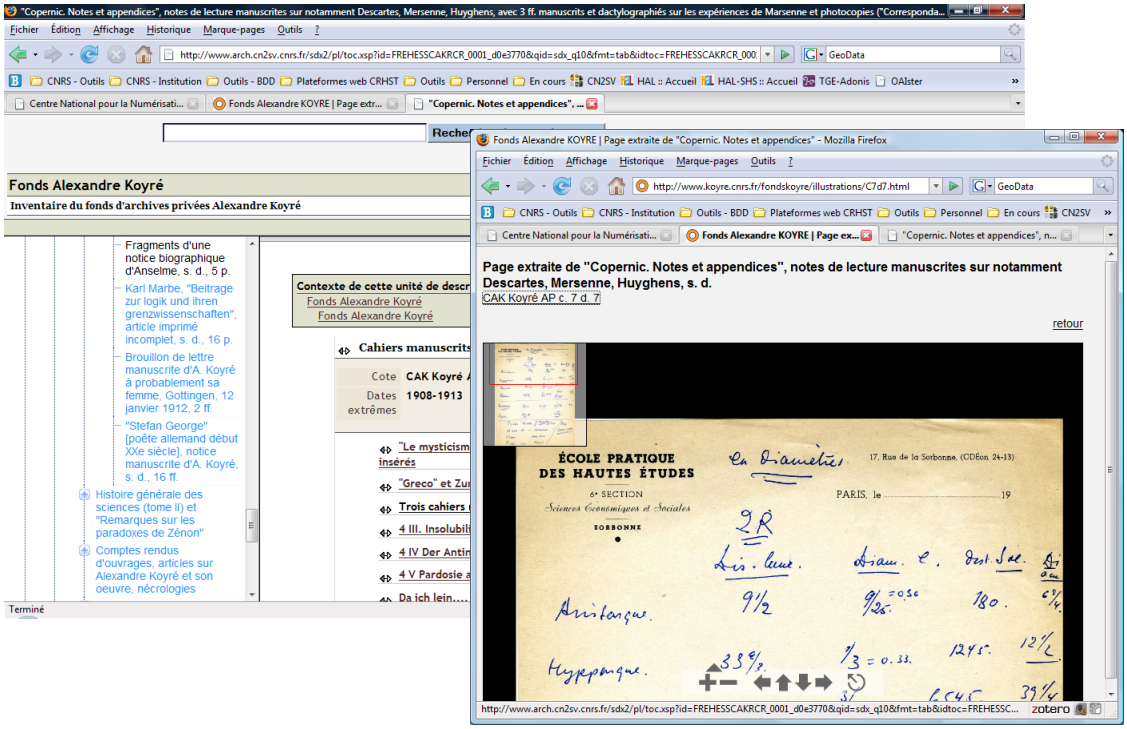

Fig. 4 : le fonds des archives d'Alexandre Koyré en liaison avec une sélection de documents numérisés.

Au travers de ces deux exemples, réalisés en 2007-2008, il nous semble important de souligner que ces outils forgent des instruments d'aide à la recherche réalisés et utilisés par les communautés scientifiques (chercheurs, ingénieurs, techniciens). L'utilisation du XML et le respect des standards et/ou normes internationales, qu'elles soient issues des réseaux de professionnels de tel ou tel domaines (EAD, EAC, etc.) ou adossées à des API dont le code et la documentation sont mis à disposition par les grands éditeurs d'applications (Google, etc.) pour des applications de recherche scientifique, seront au coeur des futures applications permettant la mise en ligne des données du patrimoine scientifique et technique. Ces applications, aujourd'hui innovantes, seront remplacées dans un futur proche par de nouveaux systèmes toujours plus simples d'utilisation et plus robustes sur le plan fonctionnel et informatique. Ainsi, ce sont les données qu'il nous incombe de pérenniser et non les applications qui les exploitent : ces dernières seront tôt ou tard supplantées par de nouvelles générations d'outils. A ce titre, la pérennisation des données (textuelles, iconographiques, sonores, 3D, etc.) est un enjeu majeur, d'aujoud'hui et de demain, pour les organises de recherche scientifique et les Universités.

14 L'inventaire et les recherches documentaires ont été réalisées par Fabien Cardoni et Margerite Vasen, bibliothécaire au CNRS et responsable de Bibliothèque de centre A.-KOYRE, à l'automne 2007.

15 Voir : http://www.koyre.cnrs.fr/fondskoyre [d.c. 05/06/08]