Web 2.0 et indexation collaborative au Centre National pour la numérisation de sources visuelles (CN2SV)

Stéphane Pouyllau

Web 2.0 et indexation collaborative au Centre National pour la numérisation de sources visuelles (CN2SV)

par M. Stéphane POUYLLAU,
Ingénieur d'études au CNRS.

***

Avec la collaboration de l'équipe du CN2SV

Ce document présente le poster réalisé par le CN2SV dans le cadre
des Rencontres des professionnels de l'information scientifique et technique (RPIST)
organisé par l'INIST-CNRS au Palais des congrès de Nancy (France) du 18 au 20 juin 2007.

This document presents the CN2SV's poster about web 2.0 presented in the RPIST meeting
in Nancy (France) – 2007- 06 – 18/20

Mots-clés : archives de science ; informatisation des données ; plateformes web 2.0 ; numérisation ; EAD ; XML ; traitements documentaires ; documents numériques
Keywords : science archives ; data processing ; web 2.0 plateform ; digitalization ; EAD ; XML ; information management ; e-documents

I. Introduction

Le centre national pour la numérisation de sources visuelles[1] (CN2SV) est un centre de ressources numériques créé en février 2006 par le Département SHS[2] et Direction de l'information scientifique[3] du CNRS. Adossé au Centre Alexandre Koyré/CRHST[4] et à son pôle informatique[5], le CN2SV a pour vocation l'expertise et l'aide à l'informatisation des données scientifiques pour la recherche en Sciences Humaines et Sociales. Il réalise et propose des plateformes web d'archives de documents numériques (ou numérisés) s'appuyant sur des réservoirs de sources primaires visuelles (manuscrits scientifiques, carnets de laboratoire, photographies de terrain, carnets de notes, cartes et plans, etc.). Il développe aussi des outils génériques permettant l'exploitation de ces réservoirs dans un environnements de travail de type web 2.0.

Son périmètre d'action, définit par une lettre de mission, couvre plusieurs domaines académiques : l'histoire des sciences et des techniques (autour des projets du Centre A. Koyré/CRHST, Paris), la géographie et la cartographie historique (autour des fonds cartographiques patrimoniaux du Centre de Documentation REGARDS-CNRS[6], Bordeaux) et les aires culturelles (autour des projets du CEMAf[7] et de la bibliothèque Eric-de-Dampierre / Université de Paris X).

Le CN2SV s'intègrera, courant 2007, au dispositif du Très Grand Équipement ADONIS (TGE du Département SHS du CNRS) et ces missions évoluerons. Dans cette première phase, il s'est appuyé sur le modèle de gestion de données numériques OAIS[8] pour mettre au point ses méthodologies et sa chaîne d'informatisation des données. La première année – qui fut expérimentale – a permis :

- De définir les briques nécessaires au fonctionnement d'une archive de fonds scientifiques[9] accessible un site web sécurisé (extranet),

1) Voir : http://www.cn2sv.cnrs.fr d.c. : 15/06/2007.
2) Voir : http://www.cnrs.fr/shs d.c. : 15/06/2007.
3) Voir : http://www.cnrs.fr/dis d.c. : 15/06/2007.
4) Unité Mixte de Recherche du CNRS, de l'Ecole des Hautes Etudes en Sciences Sociales, de la Cité des Sciences et de l'Industrie et du Muséum National d'Histoire Naturelle (http://www.koyre.cnrs.fr : d.c. 15/06/2007).
5) Le pôle informatique du Centre A. Koyré/CRHST a été créé en 2001 par Pietro Corsi (Professeur d'histoire des science à l'Université d'Oxford (UK), Directeur d'études cumulant à l'EHESS, membre du Centre A. Koyré et ancien directeur du CRHST.
   Depuis 2006, le pôle informatique est placé sous la responsabilité scientifique de Christine Blondel (Chargée de recherche au CNRS) et il oriente aujourd'hui ses actions autour des questions d'informatisations des données de sources primaires numériques et mène des actions de réalisation de projets, de veille et d'expertise sur le plan national au travers du CN2SV (http://www.hstl.crhst.cnrs.fr d.c. : 15/06/2007).
6) Voir : http://www.regards.cnrs.fr d.c. : 15/06/2007.
7) Voir : http://www.cemaf.cnrs.fr d.c. : 15/06/2007.
8) *Open Archival Information System* : Voir : http://en.wikipedia.org/wiki/OAIS d.c. : 19/06/2006
9) Voir le rapport d'activité 2006 disponible dans la rubrique « rapports » du site web du CN2SV.

- La création d'une plateforme de documents numériques regroupant des instruments de recherche créés à partir de fonds d'archives,
- D'organiser un réseau de producteurs de fonds et d'utilisateurs (bibliothèques, centres de documentation, services d'archives scientifiques, centre de recherche, chercheurs) et la création d'un atelier technologique qui a rassemblé en octobre 2006 une vingtaine de personnes (organisé par la formation continue de la DR1 du CNRS),
- D'organiser un réseau de compétences, qui pourrait préfigurer un réseau métier national, autour de la problématique du CN2SV et plus largement de celle des centres de ressources numériques du CNRS (TELMA ; CRDO),
- De développer une expertise autour des questions de mise en ligne d'archives numériques de fonds scientifiques.

Ce projet n'aurait pas pu voir le jour sans le soutien institutionnel et financier de la Direction de l'information scientifique, du Département SHS du CNRS, des partenaires producteurs, du Centre A. Koyré/CRHST et du centre de calcul de l'IN2P3-CNRS[10] qui a mis en place l'infrastructure serveurs nécessaire au développement du projet[11].

En premier lieu, la création d'une chaîne de traitement numérique des documents était primordiale : il était nécessaire de bien comprendre les objets traités (tant sur le plan des formats de fichiers, des contraintes d'informatisation que de la façon de procéder au traitement archivistique de ces derniers).

II. Une chaîne numérique : des métiers et des outils au service de la recherche.

Les travaux du CN2SV s'articulent le long d'une chaîne de traitement de l'information. Ils répondent à un besoin de chercheurs : exploiter au mieux, dans un environnement de travail informatisé, des sources primaires numérisées[12] et souvent peu ou mal référencées ; principalement quand elles sont hébergées dans les sites personnels de chercheurs, de doctorants ou de laboratoires. Depuis quelques années, l'accès en ligne à ces sources numériques est rendu plus facile : de grandes

10) Voir : http://cc.in2p3.fr d.c. : 15/06/2007.

11) Nous tenons à remercier plus particulièrement Messieurs Laurent Romary et Gérard Sabah (DIS), Denis Peschansky (SHS). Mesdames Jacqueline Carroy, Christine Blondel, Thérèse Charmasson, Lucie Secchiaroli, Caroline Lefranc, Anne-Laure Pierre, Laurence Camous, Nathalie Queyroux, messieurs Dominique Pestre, Alain Michel, Robert Vergnieux. Je tiens à remercier, en tant qu'auteur de ce document, mes collaborateurs du CN2SV (Marie-Dominique Mouton, Catherine Morel-Pair, Fabrice Melka, Daniel Pouyllau) qui ont été de près ou de loin associés à la création du poster présenté le 19 juin 2007 aux Rencontres des Professionnels de l'Information Scientifique et Technique (Palais des congrès de Nancy, 18-20 juin 2007), voir : http://rpist.inist.fr d.c. : 18/06/2007.

12) Dont nous donnons une typologie non exhaustive dans l'introduction.

institutions (bibliothèques, archives nationales, laboratoires[13], etc.) ont investi le domaine. Cependant, il existe peu d'outils d'exploitation pour ces sources. Le protocole d'échange OAI-PMH, s'appuyant sur XML, offre la base aujourd'hui pour la création d'entrepôts ou réservoirs moissonnables assurant une partie de la diffusion des données (des publications scientifiques aux sources primaires). Les catalogues de bibliothèques sont très nombreux à être accessibles en ligne mais ils ne donnent pas accès au texte intégral ou à « l'image source » (principalement pour des raisons juridiques). Les inventaires de fonds d'archives scientifiques disponibles en ligne sont très rares[14] et, quand ils existent, peu maniables. Certains sont réalisés pourtant lors de mémoires de masters ou de thèses mais ils restent non-publié. Autre élément, l'explosion de la photo numérique entraîne une révolution dans la création de corpus visuel. De nombreux chercheurs photographient leurs sources, qu'ils soient en archives ou sur le terrain. Ces photos, qui ne sont que numériques et qu'il n'est pas question d'imprimer pour des raisons de coût, se retrouvent stockées dans nos fragiles disques durs d'ordinateurs portables, nos « mémoires flash » ou dans des CD/DVD dont on sait aujourd'hui qu'ils ne sont pas réellement pérennes. Mais surtout, cette « mémoire numérique » qui permet la recherche aujourd'hui n'est pas assez pérennisée pour une exploitation par les générations futures[15] : il nous faut repérer, préparer, stocker, transcoder et enrichir ces « e-documents » pour qu'ils soient ré-utilisables un jour. Nous sommes là à la frontière entre plusieurs métiers : la recherche (production des savoirs), la documentation et l'informatique (gestion de l'information), et l'archivistique (conservation, analyse, classement, signalement).

La chaîne que nous proposons s'appuie justement sur ces trois métiers. Notre projet permet d'apporter un début de réponse aux questions des équipes de recherche ou de plusieurs chercheurs qui, préparant un projet de recherche (avec demande de financement ANR par exemple), ont besoin de savoir comment informatiser des lots de documents.

A titre d'exemple, l'équipe du CN2SV réalise des expertises en matière de numérisation : A partir de combien de documents est-il intéressant de sous-traiter ? Dois-je faire de la numérisation photographique ou avec un scanner ? Existe-t-il des normes ? Quels formats choisir ? Même si l'informatisation des données doit commencer avant ces questions sur la numérisation, lors de définition du projet de recherche, mais nous nous efforçons d'apporter des réponses claires et

13) Nous pourrions citer par exemple l'IRHT pour les textes, le LACITO pour les sources orales.

14) Les questions juridiques des fonds d'archives sont très complexes, nous envisageons ici, la diffusion de fonds scientifiques dont les documents appartiennent à des laboratoires (cartes, collections photographiques) et dont les questions juridiques de diffusion et d'accessibilité peuvent faire l'objet d'une convention. Nous travaillons en priorité sur des documents qui sont tombés dans le domaine public.

15) Cependant, depuis la fin des années 90, certains laboratoires – dans leurs domaine – archivent ces « nouveaux documents ».

pragmatiques en présentant par exemple les normes européennes MINERVA[16], en réalisant des tests d'informatisation, en aidant à la rédaction d'un cahier des charges pour une sous-traitance, en expliquant les formats de sauvegarde (TIFF) et d'exploitation web (JPG, PNG, etc), en signalant l'importance des métadonnées images de type IPTC ou XMP. Cette chaîne commence à partir du moment ou le document - la source – devient numérique. Notre chaîne aide à faire des choix, à prendre des options, à réaliser des opérations concrètes. Nous ne faisons pas de numérisation[17] directement car il y a, aujourd'hui, de nombreux partenaires qui le font très bien ; nous aidons cependant les responsables de projets à organiser leur propre chaîne de traitement numérique. Au delà, nous aidons bien évidement les équipes en matière d'initialisation de bases de données, de réalisation de cahier des charges, de mise en place d'entrepôts OAI-PMH ou de services web de diffusion de sources numériques (via des outils PHP principalement).

Notre chaîne comporte 8 briques :

1. Analyse des besoins du projet de recherche
2. Traitements documentaires, analyse des documents ou du fonds
3. Organisation de la sous-traitance en matière de numérisation
4. Sensibilisation aux conditions de stockage sécurisées tendant vers le stockage pérenne
5. Mise en oeuvre des briques logicielles permettant la gestion (utilisation du concept de *mashup*[18])
6. Mise en oeuvre d'outils collaboratifs permettant le repérage et l'indexation communautaire.
7. Appropriation par les chercheurs de ces outils et concepts.
8. Diffusion des documents via des plateformes web et/ou OAI-PMH[19] proposant ainsi un corpus « raisonné » de sources.

Cette chaîne n'est bien sur pas exhaustive, ni un modèle, elle se veut avant tout pragmatique car elle s'appuie sur notre expérience en matière de gestion du document numérique que la plupart des membres du CN2SV proposent par ailleurs dans leurs laboratoires depuis plus de 15 ans. Elle est adossée à des compétences, à des métiers et à des savoir-faire. Elle permet aux chercheurs ayant un

16) Réseau Ministériel pour la Valorisation des Activités de Numérisation : voir : http://www.minervaeurope.org/ d.c. : 15/06/2007 et : http://www.culture.gouv.fr/mrt/numerisation/fr/f_minerva.htm d.c. : 15/06/2007.
17) Excepté dans quelques cas expérimentaux qui préparent le plus souvent une numérisation professionnelle par un prestataire spécialisé.
18) Le *mash up* est un terme anglais désignant le mélange de technologies dans le but de répondre au mieux à une problématique posée.
19) Nous encourageons l'utilisation du Dublin Core (DC) comme jeu de métadonnées et cela dans le but d'être compatible avec les projets européens DRIVER, CASPAR, etc.

projet de recherche d'équipe de travailler avec les documentalistes, les ingénieurs en développements d'applications, les archivistes, les webmestres.

III. Réalisations

Le CN2SV a développé en 2006 deux outils informatiques permettant de faire de l'indexation collaborative de fonds scientifiques (EXE pour *Ead Xml Explorer*) d'une part et de la diffusion d'instruments de recherche XML de fonds d'archives d'autre part (arch.cn2sv). « EXE » offre aux chercheurs la possibilité de « travailler » une collection de documents ; « arch.cn2sv » permet de diffuser des inventaires archivistiques après encodage XML.

1. EXE est un outil fonctionnant autour d'une application de type web 2.0[20] : il propose un instrument de gestion de mots-clés individualisés et sécurisés dont la pratique s'est généralisée depuis quelques années : c'est le concept de folksonomie[21]. Chaque personne ayant un compte EXE peut affecter aux documents des mots-clés et des commentaires qui seront stockés dans une base de données propres et qui ne se mélangera pas avec le fonds original. A chaque connexion le chercheur retrouve ces mots-clés et ces commentaires. Il peut en visualiser la fréquence d'utilisation sous la forme d'un nuage de mots-clés (ou *tags cloud*). La pondération des termes et leurs répétitivités sont visualisés par l'utilisation de tailles de caractères différentes[22].
2. L'outil arch.cn2sv[23] est réalisé à l'aide du logiciel PLEADE[24] (logiciel *open source* développé par les sociétés AJLSM et Anaphore à partir du système SDX[25] et qui est utilisé massivement pour la diffusion d'instruments de recherche archivistiques par les Archives Nationales et Départementales et certains ministères). Dans notre cas, PLEADE est installé sur une plateforme web Apache/Tomcat dédiée et hébergée par le centre de calcul de l'IN2P3-CNRS à Villeurbanne (France). Il est possible de la consulter à l'adresse :

20) Nous entendons par « applications web 2.0 » celles qui sont capables d'interagir avec ses utilisateurs de façon sécurisée sur le modèle du client web / serveur web. EXE offre des profils personnels et la possibilité de conserver des paramètres entre plusieurs sessions d'utilisation. Les applications web 2.0 se développent depuis quelques années, si ce concept est encore un peu flou et fait débat, une deuxième génération d'outils web s'est bien mis en place autour des plateforme *flickr*, *YouTube*, *Scribd*, etc.
21) Voir : http://fr.wikipedia.org/wiki/Folksonomie d.c. : 19/06/2007.
22) Système très utilisé de nos jours dans les blogs.
23) Un présentation vidéo est disponible sur *YouTube* : http://www.youtube.com/watch?v=x2vCVKnmFSU d.c. : 15/06/2007
24) Voir : http://www.pleade.org d.c. : 16/06/2007
25) SDX est une plateforme de publication de documents XML construite autour du serveur d'applications *Apache/Tomcat* et de l'application *Cocoon*. SDX est une application *open source* et gratuite. Voir : http://www.sdx.org d.c. : 15/06/2007

http://www.arch.cn2sv.cnrs.fr/ead. PLEADE offre la possibilité de publier des inventaires de fonds d'archives (dans notre cas issue de fonds de chercheurs, de scientifiques, de laboratoires, etc.) qui ont été encodés suivant la DTD[26] EAD (*Encoded archive description*)[27]. L'EAD est une grammaire XML, conforme à la norme ISA(G) diffusée par l'ICA[28] et qui est utilisée par les services d'archives du monde entier[29]. Cette DTD, qui devient aujourd'hui un schéma XML[30], est composée de balises XML permettant la description de tous les niveaux de classement des inventaires : du fonds, jusqu'aux « pièces archivistiques ». Pour certains fonds traités par le CN2SV, nous avons pu donner accès aux documents numériques. Dans le cadre du respect du modèle OAIS (cf. ci-dessus), nous ne stockons pas le fichier contenant le document numérique dans l'inventaire : il est stocké dans silo documentaire dédié (un *repository*) se trouvant sur un serveur dédié. Le lien est assuré – comme en OAI-PMH – par un numéro unique d'identification. Cette stratégie permet de faire évoluer l'inventaire publié d'une par sans avoir à manipuler les documents numériques. Une fois publié, un inventaire deviennent un instrument de recherche (que nous nommons « I.R. »). Un I.R. est le fruit du traitement documentaire et archivistique, de processus numériques et informatique (dont la numérisation et l'encodage) et du travail de réflexion sur les contenus effectué par le chercheur. Nous retrouvons ici nos trois métiers. Il est possible de publier plusieurs fois un inventaire si son contenu s'accroît et lors de modifications des contenus. Ces nouveaux objets, entièrement numériques, sont amenés à évoluer dans le temps, ainsi l'un des rôles du CN2SV est aussi de suivre les évolutions technologiques, les besoins des équipes de chercheurs, les bonnes pratiques afin de maintenir disponibles ces I.R. Nous touchons là au domaine de la pérennisation[31] des données numériques[32].

Ces deux produits sont accompagnés d'un site web[33] présentant le CN2SV et comportant un extranet

26) DTD : *Document Type Definition* ou « définition de type de document » ensemble de règles qui structurent un document XML

27) Cette DTD, dont la dernière révision date de 2002, est maintenue par la *Library of Congress* (http://www.loc.gov/ead/ d.c. : 15/06/2007).

28) International Council on Archives : http://www.ica.org d.c. : 15/06/2006

29) En avril 2007 s'est tenu à Berlin la 3eme conférence internationnale sur l'EAD.

30) Un schéma XML à la même fonction qu'une DTD, la différence réside dans le fait que le schéma XML est lui-même écrit en XML, ce qui n'est pas le cas de la DTD.

31) La pérennisation des données numériques n'est pas uniquement centrée sur le stockage pérenne. Elle regroupe plusieurs actions : suivit des formats, la transmission des données, l'historique de l'objet numérique, le stockage, l'anticipation des évolutions technologiques et institutionnelle, etc.

32) Sur les questions de pérennisation des données numériques – qui sont « relativement » neuves hors des « milieux informatisées » – il sera possible de consulter mon blog dans lequel il m'arrive de traiter certaines approches sous la forme de billets : http://blog.stephanepouyllau.org

33) Accès par : http://www.cn2sv.cnrs.fr d.c. : 15/06/2007 ; rubrique « extranet », une inscription est à demander par courriel.

de veille méthodologique sur les briques de notre chaîne numérique d'informatisation, sur OAIS, etc. Cet extranet prendra la forme d'un blog dans quelques semaines.

IV. Conclusion

Créé en 2006, le CN2SV est une structure souple, regroupant des ingénieurs et des chercheurs, autour de compétences technologiques communes et aujourd'hui organisée sous la forme d'un réseau et d'un groupe de travail. Ses missions sont centrées sur l'informatisation de données (fonds d'archives, de chercheurs, de scientifiques et de savants) dans le but de créer des entrepôts de données et de sources qui seront utilisés par les chercheurs et les enseignants dans quelques années ; ces premières réalisations préfigurent les outils et méthodes qui seront utilisés dans le futur. Les outils collaboratifs d'annotations, les espaces numériques de travail existent déjà, le CN2SV propose des briques simples qui peuvent s'y greffer et propose des actions et méthodes pragmatiques d'aide à l'informatisation de fonds. Ces éléments sont illustrés par le poster présenté lors des rencontres des professionnels de l'information scientifique et techniques organisées par l'INIST-CNRS à Nancy du 18 au 20 juin 2006.

Contact : Stéphane POUYLLAU, courriel : pouyllau@ivry.cnrs.fr, tél. +33 (0)1 40 05 73 92