

Bonnes pratiques de numérisation et de mise en ligne

On essaiera de dire en quelques mots l'essentiel de ce qu'on peut tirer d'une quinzaine d'années d'expérimentations puis de réalisations patrimoniales et scientifiques dans le domaine, au moment où les solutions technologiques atteignent un niveau de maturité et de performance vraiment intéressant, où les initiatives foisonnent et où les exigences augmentent.

1. Numériser

Commençons par donner une **définition** : **numériser**, ce sera produire des substituts numériques (fichiers images, enregistrements sonores ou audio, textes...) à partir des sources primaires de la recherche, et des travaux des chercheurs. Le document numérique est composé fondamentalement d'une séquence de zéros et de uns, les seuls signaux (bits) qu'un ordinateur puisse comprendre, à condition qu'on lui donne les clés lui permettant d'interpréter cette séquence (type de fichier, système de codage). On peut - on doit - réintroduire dans ce champ les **documents** primaires et travaux de recherche **nativement numériques** ; il y en a de plus en plus avec l'explosion de la photographie numérique, ce qui ne veut pas dire que ceux-ci ne doivent pas être transformés pour satisfaire les critères définis pour un projet.

Tout est d'abord affaire d'**objectifs**. Trois d'entre eux sont souvent présents pour orienter un projet : **conserver** pour le moyen ou le long terme, **communiquer** au plus grand nombre, **exploiter** efficacement, éventuellement autrement.

Veiller au respect de quelques principes simples peut aider à la réussite d'un projet :

- **bien sélectionner les objets documentaires à traiter**. Il pourra s'agir d'un gisement organique ou du moins préexistant (fonds, collection) ou d'un corpus "inventé" par le chercheur, ou des deux. Dans les deux cas, expliciter le choix sera utile ; il faudra veiller à la cohérence de l'ensemble et au respect de sa nature intrinsèque, ce qui pourra orienter les choix en terme de traitements et de solutions de diffusion ; il faudra aussi s'assurer qu'on est en **droit** de reproduire et de diffuser ces documents ;
- définir les **paramètres de numérisation**, en fonction des objectifs donc des critères de qualité à satisfaire (pour l'image, ce sera choisir notamment un modèle chromatique, une définition et/ou une résolution) ;
- définir les **modes opératoires et les niveaux d'intervention**, ce qui dépendra notamment de la présentation matérielle, de l'homogénéité et de l'état initial des documents (s'agit-il de produire une copie fidèle de la source primaire, sans corriger ses défauts, ou d'appliquer des traitements visant à améliorer la lecture ou l'écoute ?) ;
- choisir des **formats d'enregistrement** et/ou des algorithmes de compression **non propriétaires**, standards de fait ou officiels, permettant de manipuler les fichiers sans dépendre d'un outil ou d'une plate-forme (ex. : TIFF LZW ; JPEG ; PNG). En outre, **pour le texte** résultant d'une reconnaissance de caractères ou d'une conversion, choisir **un format cible structuré**, ce qui en permettra l'exploitation, à condition que le modèle de données ou la documentation soient associés au texte (ex. : fichier tabulaire au format CSV, avec des noms de champs et un dictionnaire de données ; document XML¹ conforme à un modèle documentaire international comme TEI² ou Docbook³) ;
- définir un **plan de nommage et de classement** des fichiers permettant un dépôt sur n'importe quel

1 Extensible Markup Language : un métalangage simple et flexible pour le balisage, donc la structuration du texte, dérivé de SGML. Les spécifications de la version 1.0 ont été publiées par le W3C en 1998. XML a donné naissance à un grand nombre de normes et de langages, comme XSLT ; son utilisation s'est répandue dans tous les domaines informatiques, notamment ceux où l'échange et la transformation de données ou documents sont essentiels ; voir <<http://www.w3.org/XML/>>.

2 Text Encoding Initiative, un projet international né en 1987, qui a produit un modèle XML pour l'édition structurée de tout type de texte, exprimé aujourd'hui sous la forme de schéma ; voir <<http://www.tei-c.org/Guidelines/P5/>>.

3 Docbook, un projet né en 1991, qui a produit un schéma pour la structuration en XML de la documentation technique ; voir <<http://www.docbook.org/>>.

ordinateur et système d'exploitation, ou sur n'importe quel support de stockage externe (norme ISO 9660) ;

- **numériser**, ou faire numériser ; cela pourra parfois se faire en plusieurs filières et solutions parallèles si les objets à traiter sont hétérogènes, ou en plusieurs étapes. Ainsi, on ne numérise pas des cartes topographiques ou des manuscrits avec les mêmes outils qu'une collection d'ouvrages reliés de format invariable ou un objet à restituer en trois dimensions. Il faudra établir un cahier des charges, et s'équiper au moins pour contrôler le travail ;
- enregistrer les **métadonnées de gestion** (techniques, juridiques, administratives) des objets numériques ;
- **indexer ou, mieux, décrire les fichiers** (plus précisément ici, les sources primaires numérisées). Au minimum, un tableau de récolement donnant la liste des fichiers produits et la concordance avec les sources traitées ; mieux, une indexation du contenu, mieux encore, une base de données ou d'informations de description et d'indexation. Là aussi, des normes conceptuelles et techniques existent, établies au niveau international (ISBD⁴, ISAD(G)⁵, ISAAR(CPF)⁶, FRBR⁷, CIDOC CRM⁸, etc., pour les concepts ; UNIMARC⁹ et MARCXML¹⁰ ou BiblioML¹¹, EAD¹², EAC¹³, TEI, X3D¹⁴, etc., pour leur traduction informatique ; normes ISO et AFNOR pour l'indexation, fichiers d'autorité, thésaurus). Souvent cette étape demande un investissement humain important, même s'il préexiste une base de connaissances il peut y avoir à revoir et normaliser les informations ;
- établir les **relations entre fichiers résultat de la numérisation des sources primaires et métadonnées de gestion et de description**. Diverses solutions existent ; si on a choisi XML et qu'on veut gérer les relations hors des fichiers, on pourra choisir le standard international METS¹⁵ ;
- **archiver** les résultats, c'est-à-dire choisir un support pour enregistrer les paquets d'information et équiper un lieu pour les stocker dans les meilleures conditions (sécurité, climat, etc.), et se doter des moyens de surveiller ces supports dans le temps, de détecter les altérations et de procéder à des migrations régulières.

-
- 4 International Standard Bibliographic Description : un ensemble de normes internationales publiées à partir de 1970 par l'IFLA (Fédération internationale des associations de bibliothécaires et de bibliothèques) définissant les informations nécessaires pour cataloguer les monographies, publications en série, et documents spéciaux tels qu'images fixes, documents cartographiques et audiovisuels. Voir par ex. : <<http://www.bnf.fr/pages/infopro/normes/no-isbd.htm>>.
 - 5 ISAD(G) : Norme générale et internationale de description archivistique, publiée par le Conseil international des archives. Deuxième édition parue en 2000 ; voir <<http://www.ica.org/fr/node/30001>>.
 - 6 ISAAR(CPF) : Norme internationale sur les notices d'autorité archivistiques relatives aux collectivités, aux personnes et aux familles, publiée par le Conseil international des archives. Deuxième édition parue en 2004 ; voir <<http://www.ica.org/fr/node/30231>>.
 - 7 Le modèle FRBR (Functional Requirements for Bibliographic Records / Spécifications fonctionnelles des notices bibliographiques) est un modèle conceptuel élaboré par un groupe d'experts de l'IFLA (Fédération internationale des associations de bibliothécaires et de bibliothèques) de 1992 à 1997. Voir notamment <<http://www.bnf.fr/pages/infopro/normes/no-acFRBR.htm>>.
 - 8 Le modèle CIDOC CRM est un modèle sémantique de référence élaboré depuis 1994 par le Groupe de normalisation documentaire du Comité international pour la documentation du Conseil international des musées (ICOM-CIDOC). Voir : <<http://www.bnf.fr/pages/infopro/normes/no-acCRM.htm>>.
 - 9 En France, UNIMARC (acronyme pour UNiversal MARC) est le format officiel d'échange de l'information bibliographique et le format de travail du Sudoc (Système universitaire de documentation) et de la plupart des bibliothèques publiques. Voir notamment : <<http://www.bnf.fr/pages/infopro/normes/no-acuni.htm>>.
 - 10 MARCXML est un schéma XML, transposition en XML du format MARC 21. Voir <<http://www.loc.gov/standards/marcxml/>>.
 - 11 BiblioML est une application XML pour des références bibliographiques et des données d'autorités, basée sur les formats bibliographique et d'autorités UNIMARC. BiblioML et AuthoritiesML sont des formats basés sur XML pour l'échange d'enregistrements bibliographiques et d'autorités entre applications. Ils ont la même granularité que le format UNIMARC d'origine c'est à dire les mêmes éléments de base. Mais le marquage est utilisé pour indiquer de manière lisible, non seulement la structure de l'enregistrement bibliographique, mais aussi la sémantique précise de chaque élément. Voir : <<http://90plan.ovh.net/~adnx/biblioml/doku.php>>.
 - 12 EAD : Encoded Archival Description 2002, une DTD XML pour l'encodage des instruments de recherche archivistiques, norme internationale compatible avec ISAD(G). Voir <<http://www.loc.gov/ead/>> et <<http://www.archivesdefrance.culture.gouv.fr/fr/archivistique/DAFlangage.html>>.
 - 13 EAC : Encoded Archival Context, une DTD XML pour l'encodage de notices d'autorité, compatible avec la norme ISAAR(CPF). Voir <<http://www.iath.virginia.edu/eac>> et <<http://www.archivesdefrance.culture.gouv.fr/fr/archivistique/DAFdescnoms.html>>.
 - 14 X3D (<<http://www.web3d.org/>>) : une initiative internationale pour développer les technologies permettant de produire, manipuler et diffuser facilement les objets numériques en trois dimensions. Le projet a notamment produit un schéma XML (<<http://www.web3d.org/x3d/specifications/>>).
 - 15 METS : Metadata Encoding and Transmission Standard, un schéma XML pour encoder des métadonnées de description, de gestion et de structure pour une collection de documents numériques. Voir le site de ce projet : <<http://www.loc.gov/standards/mets/>>.

2. Mettre en ligne

Ici on sous-entend qu'on a choisi un moyen parmi d'autres pour la diffusion des sources et travaux de la recherche : **Internet**, un système d'information extraordinaire, dont la nature (un réseau de réseaux interconnectés) et la relative facilité d'accès suffisent à expliquer le choix.

Comment construire un système d'information pour le projet ou la structure porteuse de l'opération ?

D'abord le définir, **savoir ce qu'on veut qu'il offre** en termes de fonctionnalités (moyens de navigation et de recherche) et de services (consultation, extractions, annotations privées ou publiques...).

Les fonctionnalités de recherche et les services utilisateurs devront être l'objet d'un soin particulier si on s'adresse à une ou plusieurs communautés de chercheurs, sans illusion sur l'ambition de servir aussi bien l'historien que le linguiste, le doctorant se posant des questions extrêmement pointues que l'étudiant débutant. Donner le moyen d'**extraire tout ou partie des documents ou données** produits dans le cadre d'un projet pour pouvoir les intégrer à son propre corpus et les exploiter à l'aide de ses outils, de pouvoir **annoter et commenter** lesdits documents en ligne, pour soi-même ou pour tous, constituent aujourd'hui des objectifs réalistes dans de nombreux cas, que divers projets patrimoniaux ou scientifiques, parmi lesquels certains réalisés par les centres de ressources numériques TELMA¹⁶, CN2SV¹⁷ et CRDO¹⁸ avec leurs partenaires, ont d'ailleurs atteints. Travailler à construire des **outils de production et d'exploitation de corpus** qui soient facilement pris en main par le chercheur non informaticien chez lui, et adaptables à ses besoins, est un autre sujet ; on peut imaginer pour l'avenir que les deux logiques se rejoignent pour obtenir des systèmes d'information en ligne dont certaines parties au moins soient générées par l'utilisateur final.

Le plus souvent on choisira un système d'information de type **Web dynamique**, qui intègre notamment un moteur de recherche et retourne à l'utilisateur ce qu'il a demandé. Il faudra aussi pouvoir en mettre à jour facilement les contenus.

Le système d'information public devrait être chargé de servir ou de produire des paquets d'information destinés à la diffusion, à distinguer du paquet de versement ou du paquet d'archivage. Ceci pour employer le vocabulaire fonctionnel du **modèle OAIS**¹⁹.

Des critères de choix importants :

- **performance et stabilité** : des composants et une architecture adaptés aux besoins, et éprouvés ailleurs, ou testés, y compris en production ;
- **évolutivité, durabilité** : composants compatibles avec de nombreux systèmes d'exploitation, bien soutenus par les structures ou communautés qui les développent ; langages de programmation normés et populaires ; maintenance correctrice et évolutive prévue et facile ; modularité ; si possible, code ouvert ;

16 TELMA (Traitement Electronique des Manuscrits et des Archives) est un centre de ressources numériques (CRN) du CNRS adossé à l'École nationale des Chartres d'une part, à l'Institut de Recherche et d'Histoire des Textes d'autre part. Il œuvre pour la pérennisation, la diffusion, la mise en valeur et partant, l'exploitation électroniques des sources anciennes. Il a organisé, archivé et publié sur Internet une quinzaine de corpus documentaires - éditions de textes et répertoires de ressources, le plus souvent associés aux substituts numériques des textes décrits et recensés -, après les avoir convertis en XML/TEI. Cette activité de publication va se poursuivre dans les années qui viennent à partir d'autres corpus de nature parfois différente (instruments de recherche archivistiques, textes de l'époque moderne, glossaire de latin médiéval...). Le site Web de TELMA <<http://www.cn-telma.fr/>> a été développé par l'ENC et l'IRHT et est actuellement hébergé et administré par l'ENC, sur une plate-forme technique elle aussi nommée TELMA.

17 Le CNS2V (Centre National pour la Numérisation de Sources Visuelles, <<http://www.cn2sv.fr/>>) est un CRN du CNRS créé en 2005 par le Département Sciences Humaines et Sociales et par la direction de l'information scientifique du CNRS. C'est une plate-forme technologique en réseau spécialisée dans l'informatisation de données visuelles (photos, diapos, carnets de terrains, cartes, planches, dessins, croquis, etc.) pour des activités de recherche (projets d'équipe, projets ANR, etc.). Dans ce cadre, il met en place des outils de diffusion de corpus numériques scientifiques au travers d'outils Web 1.0 ou 2.0, d'extranets, d'entrepôts OAI-PMH, etc. Le CN2SV est adossé pour son fonctionnement au Centre Alexandre Koyré/CRHST.

18 Le CRDO (Centre de Ressources sur la Description de l'Oral, <<http://www.crdo.fr/>>) est le CRN centré sur les ressources orales. Il a été composé à partir de deux propositions portées respectivement par les laboratoires Lacito et LPL. Il a pour objectif de permettre le partage et l'échange de données linguistiques. De nombreuses données, réparties en trois catégories, sont disponible sur son site : des corpus oraux et vidéos accompagnés d'échantillons ; des outils dédiés à la linguistique ; des ressources, ou plus précisément des lexiques, bases de référence, systèmes de représentation, grammaires....

19 OAIS (Open Archival Information System) est un modèle conceptuel destiné à la gestion, à l'archivage et à la préservation à long terme de documents numériques. La mise au point de OAIS a été pilotée par le CCSDS (Comité Consultatif pour les Systèmes de Données Spatiales). L'OAIS est enregistré norme ISO sous la référence 14721:2002. Voir <<http://fr.wikipedia.org/wiki/OAIS>>

respect des principes et normes techniques du développement Web, essentiellement du principe de la séparation entre contenu et présentation ;

- **accessibilité** : respect des recommandations gouvernementales²⁰ et de la recommandation WAI²¹ du W3C ;
- **interopérabilité** : des moyens pour faire en sorte que les contenus de l'application soient référencés pour nourrir d'autres applications distantes de type portail. Une solution consiste à prévoir de monter un **entrepôt**, interrogeable au moyen de requêtes conformes au protocole d'échange OAI-PMH²², donc exposant des métadonnées en XML conformes au modèle Dublin Core²³.

De par sa nature et son statut, un composant logiciel libre au code ouvert peut être un bon candidat pour satisfaire de telles exigences. Sans parler de logiciels libres très connus et génériques comme Linux, Apache, MySQL, Lucene, qui entrent très souvent dans une architecture applicative, des logiciels libres spécialisés dans la publication et la consultation des bases de connaissances métier sont aujourd'hui disponibles.

3. Conclusion

Nous espérons avoir montré qu'**un projet de numérisation et de mise en ligne ne peut pas s'improviser**.

Tout d'abord, les étapes préparatoires (études préalables, cahiers des charges et spécifications, recueil, description et indexation des documents) et finales (contrôles, tests, mise en production...), en amont et en aval des opérations proprement dites de numérisation ou de développement informatique, sont essentielles. De façon plus générale, rares sont les projets de cette nature qui aboutissent en moins d'un an. Surtout, il s'agit de **mobiliser** autour d'une unité de recherche ou d'un projet contractualisé, non seulement des moyens financiers ou techniques, mais **des compétences solides dans plusieurs disciplines**, de l'archivistique à l'ingénierie système et l'ingénierie documentaire en passant par l'expert en photographie numérique ou en traitement du son. Si certaines de ces compétences n'existent pas en interne, il est sage de se faire aider et accompagner par des personnes extérieures.

Des risques et des incertitudes existent à chaque étape, inhérents à ce type de projet informatique. Orienter, coordonner, réunir et arbitrer, programmer, faire le point, évaluer, communiquer, sont des tâches indispensables, qui demandent un "manager" à la fois réaliste et ambitieux, dévoué et sans prétention, capable de parler avec le chercheur comme avec l'informaticien ou avec les grands centres de stockage des données numériques, de faire rebondir et de capitaliser.

La **dimension humaine** est donc cruciale dans de tels projets d'informatisation.

Dans de telles aventures, les **centres de ressources numériques du CNRS** jouent depuis quelques années, et souhaitent continuer à jouer, au sein de l'infrastructure **ADONIS**, le rôle de conseillers aussi bien que d'opérateurs techniques. Ils forment, avec plusieurs autres partenaires (services d'archives, bibliothèques, centres de documentation, etc.) un réseau d'expertise dans le domaine des *digital humanities* déjà bien développé à l'étranger. En fonction des attentes et des projets et selon leurs spécialités, ils proposent une **gamme variée de services**, allant du simple conseil ponctuel à la prise en charge complète d'un projet, en passant par l'accompagnement, la formation, et le développement d'outils.

Florence Clavaud

directrice des nouvelles technologies à l'École nationale des Chartes / co-responsable technique du CRN TELMA
4 février 2008

20 Notamment le référentiel accessibilité des services Internet de l'administration française, publié par la direction générale de la modernisation de l'Etat, ex. Agence pour le Développement de l'administration électronique, en 2004 et en cours de refonte actuellement.

21 WAI : Web Access Initiative, <<http://www.w3.org/WAI/>>.

22 OAI-PMH : Open Archives Initiative Protocol for Metadata Harvesting, un protocole d'échange facilitant les échanges entre fournisseurs de données (entrepôts) et fournisseurs de services (moissonneurs), élaboré par l'Open Archive Initiative à partir de 1999. Voir notamment <<http://www.openarchives.org/pmh/>>, et <<http://90plan.ovh.net/~adnx/documentation/oai/>>.

23 Dublin Core Metadata Element Set, un jeu de 15 propriétés défini par le Dublin Core Metadata Initiative, une organisation internationale, pour structurer des métadonnées. Le modèle conceptuel est associé à un schéma XML et est très utilisé sur Internet. Voir la dernière édition de la recommandation, la version 1.1, à : <<http://dublincore.org/documents/dces/>>, et un guide d'encodage en XML <<http://dublincore.org/documents/dc-xml-guidelines/>>.

Éléments de bibliographie

La littérature en ligne et imprimée sur ce sujet vaste, complexe et à multiples facettes est abondante. En sus des références données au sein des notes, nous citons ci-après, sans prétendre couvrir le domaine, quelques sites Web ou ouvrages de référence.

Conduire un projet de numérisation. Sous la direction de Charlette Buresi et Laure Cédelle-Joubert. Lyon : Presses de l'ENSSIB, 2002. 326 p. La Boîte à outils ; 13. ISBN : 2-910227-43-X : ISSN : 1259-4857.

Corpus oraux : guide des bonnes pratiques. Sous la direction d'Olivier Baude. Paris : CNRS, 2006. 192 p. ISBN : 2-271-06425-2.

DORRER, Luc. *Hommes et projets informatiques : dix commandements pour réussir.* Paris : Hermès science publications, 2004. 256 p. Collection Management et informatique. ISBN 2-7462-0877-6.

JACQUESSON, Alain et RIVIER, Alexis. *Bibliothèques et documents numériques : concepts, composantes, techniques et enjeux.* Nouvelle édition. Paris : Éditions du Cercle de la librairie, 2006. Bibliothèques. ISBN : 2-7654-0716-9.

La numérisation des textes et des images : techniques et réalisations. Actes des journées d'étude organisées par l'Université de Lille III, 16 et 17 janvier 2003. Textes réunis et édités par Isabelle Westeel et Martine Aubry. Villeneuve-d'Asq : Presses de l'Université Charles de Gaulle - Lille III, 2003. 190 p. ISBN : 2-84467-050-4

Les chercheurs et la documentation numérique : nouveaux services et usages. Sous la direction de Ghislaine Chartron, avec la collaboration de Gabriel Gallezot, Annaïg Mahé, Agnès Melot et alii. Paris : Éditions du Cercle de la Librairie, 2002. 268 p. Bibliothèques. ISBN : 2-7654-0840-8 / ISSN : 0184-886.

Les documents écrits. Texte imprimé : de la numérisation à l'indexation par le contenu. Sous la direction de Rémy Mullot. Paris : Hermès Science publications, 2006.

Mener un projet Open Source en bibliothèque, documentation et archives. Electre - Cercle de la Librairie, 2007. Bibliothèques. ISBN : 978-2-7654-0954-0.

Numérisation du patrimoine culturel. France. Ministère de la Culture et de la Communication. Site Web accessible à : <<http://www.culture.gouv.fr/culture/mrt/numerisation/index.html>>.

Pérenniser le document numérique. Séminaire INRIA, 2-6 octobre 2006, Amboise. Ouvrage coordonné par Lisette Calderan, Bernard Hidoine et Jacques Millet. Paris : ADBS Editions, 2006. 206 p. Collection "Sciences et techniques de l'information".

Recommandations techniques pour les programmes de création de contenus culturels numériques. Version révisée de mai 2004, élaborée dans le cadre du projet Minerva par l'UKOLN, Université de Bath, en collaboration avec l'agence britannique Resource. Disponible en ligne : <http://www.culture.gouv.fr/culture/mrt/numerisation/fr/eeurope/documents/guide_technique.pdf>