arbido

2008/2 Die Wirtschaft der neuen Medien

Villard Hubert,

Numérisation à la BCU Lausanne

Le sur-mesure n’exclut pas la chaîne ... et réciproquement

Il faut d’emblée préciser que cet article, qui présente les travaux de numérisation à la Bibliothèque cantonale et universitaire de Lausanne, ne traite que de rétronumérisation – à savoir du transfert sous forme numérique de l’information imprimée contenue dans les livres, journaux et autres documents – et n’a pas pour objet d’embrasser toute la problématique de la gestion de l’information sous forme numérique dans les bibliothèques, notamment des documents «nés numériques».

Une suite logique de l’automatisation des années 1970

Cette rétronumérisation découle de plusieurs réflexions conduites au cours du temps dans le cadre plus général de l'automatisation des bibliothèques. C'est au début des années 1970 que Lausanne avait lancé son projet d'automatisation, en vue du transfert de la BCU sur le nouveau site universitaire de Dorigny; il en est résulté le système SIBIL, qui a connu une renommée bien au-delà des frontières du canton. Au début des années 1990, l'automatisation des bibliothèques avait atteint sa forme achevée en ce sens que toutes les fonctions traditionnelles de gestion et d'accès aux catalogues avaient trouvé solution. L'innovation – si innovation il devait y avoir – était ailleurs, et quand les utilisateurs de SIBIL rassemblés au sein du réseau REBUS se mirent à réfléchir à une refonte du système actuel, ou au système informatique qui, un jour, devrait prendre le relais de la solution développée 20 ans plus tôt, ils engagèrent une réflexion intitulée SIBIL 2000, confiée à quelques responsables encadrant un chef de projet. Celui-ci, visionnaire, ne mit guère de temps à faire comprendre que le défi, désormais, résidait dans la mise en ligne du contenu intégral des documents, et non plus seulement de leur référence bibliographique et catalographique. S'élaborèrent alors les premiers scénarios de bibliothèque numérique autour d'une «vision»: tout le savoir du monde en ligne, à portée de clic. Mais tout cela était trop neuf, trop avant-gardiste pour que le projet ait une chance de passer sur le plan politique, et, partant, d'obtenir les moyens financiers nécessaires. Rien n'était encore prêt, d'ailleurs, ni sur le plan technique (manque de scanners performants, d'unités de stockage à très haute capacité, de puissance de traitement des images, etc.), ni sur le plan conceptuel (absence de format normalisé d'archivage et de présentation en ligne, métadonnées, etc.)

On y va!

10 ans plus tard cependant, passé le cap de l'an 2000, les choses avaient rapidement évolué, et peu à peu les sourires narquois et entendus qu'on avait coutume de déclencher chez les bibliothécaires bon teint lorsqu'on parlait de bibliothèque numérique s'étaient estompés et faits plus gênés. S'il avait fallu du temps au cours des siècles pour réaliser que l'invention de l'imprimerie par Gutenberg avait eu un impact colossal sur l'évolution de la civilisation occidentale, on réalise aujourd'hui déjà que la numérisation de masse de l'écrit est susceptible d'avoir des conséquences encore plus significatives.

Alors, attendre pour être certain de faire juste, ou avancer en prenant quelques risques? Nous avons décidé d'avancer, persuadés d'une part, comme Valéry, que, contents ou pas, «nous entrons dans l'avenir à reculons», et d'autre part, comme Machado, que «caminando no hay camino; el camino se hace al andar». L'important étant de s'appuyer sur les meilleures pratiques, telles qu'elles se dégageaient des premiers projets lancés dans ce domaine à l'étranger (par exemple telles qu'elles étaient recensées dans le précieux didacticiel élaboré dès l'an 2000 par la Cornell University: www.library.cornell.edu/preser...).

Au moment où d'autres bibliothèques du pays lançaient des projets analogues (la StUB, St Gall, la Bibliothèque nationale, Fribourg, etc.), la BCU a envisagé ses opérations de numérisation comme un prolongement naturel de sa politique de préservation et de sauvegarde des collections (PAC), mise en place au début des années 1990. Dès cette époque en effet, par besoin de créer des copies de sécurité à long terme, on avait procédé au microfilmage de «trésors» de la BCU, que ce fussent des ouvrages de la Réserve précieuse, des manuscrits, des fonds photographiques ou des quotidiens. Le microfilm n'étant pas le moyen préféré d'accéder à l'information pour le public, il est très vite apparu que la seule façon de donner accès à des sources délicates ou rares sans les mettre en danger était de les numériser et de les présenter en ligne sur un serveur adéquat. Ainsi fut fait.

La bibliothèque lança alors plusieurs projets de constitution de collections numériques, laissant libre champ à l'initiative de collaborateurs enthousiastes, mais avec des moyens financiers pour le moins limités, voire inexistants. On numérisa quelques livres précieux sur les installations de la Bibliothèque nationale et auprès d'une entreprise de la place; ils sont accessibles sur le serveur RERO DOC. Mais aussi à l'atelier photo de la bibliothèque ou sur les installations de l'Université, tels des manuscrits (la *Bible Porta,* le *Répertoire biographique des pasteurs vaudois,* de Henri Vuilleumier, le *Journal* de Philippe Secrétan, etc.), et des fonds photographiques (C. F. Ramuz, R.-Th. Bosshard, A. Conan-Doyle, Géa Augsbourg, etc.), archivés pour l'instant sur des DVD, en attendant un concept d'archivage plus rassurant. La disponibilité de *scanners* tourne-pages automatiques, auprès d'une entreprise située à quelques pas de Lausanne et leader mondial dans le développement de ce genre de matériel, a permis de tester leurs performances sur des collections de quotidiens anciens (les 40000 pages de la collection complète de *L'Estafette*, les sept volumes du *Journal littéraire de Lausanne*).

Des tests entrepris par l'atelier photo de la BCU pour évaluer la qualité d'une numérisation effectuée à partir de microfilms de sécurité réalisés auparavant donnèrent des résultats encourageants: une numérisation complète du *Messager boiteux de Lausanne et Vevey* à partir d'un microfilm de sécurité 35 mm réalisé par l'atelier photo de la BCU se révèle d'excellente qualité. En parallèle, on a conduit avec la maison SUPAG quelques essais de numérisation à partir des microfilms 16 mm de quotidiens vaudois réalisés 15 ans auparavant par la maison Infoprint, qui four- nissent des résultats tout à fait satisfaisants pour une consultation en ligne qui s'avère bien plus conviviale que celle des microfilms. Reste maintenant à trouver les budgets nécessaires pour procéder à la numérisation définitive des quotidiens *24 Heures* et *Le Matin,* sans toucher aux originaux ... Enfin, le dernier projet en date résulte d'une collaboration entre *Le Temps,* la Bibliothèque nationale et la BCU pour la numérisation de la collection complète de la *Gazette de Lausanne:* un contrat est en passe d'être signé, qui prévoit un financement partagé pour la numérisation de ce quotidien important pour la vie intellectuelle, économique et culturelle du canton.

A côté de ces projets qui portent sur des documents qui appartiennent à la bibliothèque, la numérisation contribue aussi à la mission de conservation de l'héritage culturel de l'institution: ainsi, pour des livres «vaudois» qui sont absents pour une raison ou une autre des collections de la bibliothèque cantonale, mais dont on pourrait s'attendre à ce qu'ils s'y trouvent, on procède par repérage et emprunt des originaux auprès d'autres bibliothèques. Les ouvrages trouvés sont confiés en numérisation, notamment sur les installations de la Bibliothèque nationale suisse qui dispose du savoir-faire adéquat dans le traitement de documents précieux. Le fichier numérique résultant sert alors à produire un exemplaire papier selon le principe du «livre à la carte», ce qui permet de compléter après-coup les collections de la bibliothèque avec des copies d'ouvrages introuvables sur le marché.

Et pourquoi pas Google?

Mais il est très vite devenu évident qu'avec quelques dizaines de milliers de francs prélevés sur des réserves et alloués chaque année à la numérisation, on ne pouvait espérer réaliser plus de quelques dizaines d'ouvrages par an, ce qui ne nous rapprochait pas beaucoup de la nouvelle Alexandrie interactive espérée ... Et c'est alors avec l'intime conviction qu'il fallait absolument changer d'échelle et rechercher des collaborations ailleurs qu'auprès des institutions étatiques – toutes soumises au même régime minceur – que nous avons approché Google, qui venait de lancer, deux ans auparavant, un gigantesque et ambitieux programme de numérisation. Quelques contacts prometteurs, quelques mois de négociation, un argumentaire convaincant à l'intention des autorités de tutelle, et le contrat était signé au printemps 2007, qui prévoyait de numériser en moins de deux ans 100 000 ouvrages libres de droit des collections anciennes de la BCU. En conjuguant l'étonnant pragmatisme de nos partenaires américains et une gestion de projet sur place qui bousculait quelque peu le rythme usuel des bibliothèques, l'opération démarrait peu après, et en fin d'année c'étaient déjà près de 15 000 livres qui avaient été numérisés et se trouvaient en ligne sur Google Recherche de Livres (GBS). Et parmi ceux-ci des ouvrages qui n'avaient jamais été empruntés ni consultés, preuve en est qu'ils n'étaient pas coupés! Qu'on le veuille ou non, on était soudain passé d'un travail artisanal, à la pièce, garantissant une très haute qualité, à une production plus industrielle, à la chaîne. Ce qui ne veut pas dire que la qualité du travail laisse à désirer, bien au contraire: Google prend très à cœur d'améliorer constamment la qualité de son produit Recherche de Livres, par le biais de développements permanents réalisés sur les techniques de numérisation, l'interprétation OCR, les métadonnées, etc. (des représentants des bibliothèques partenaires au projet actifs et motivés sont invités chaque année à des ateliers de travail et d'échange d'expérience, qui permettent de faire évoluer le produit).

Comme résultat de tous ces projets bien différenciés et juxtaposés, nous nous retrouvons avec un nombre croissant de documents significatifs de la bibliothèque déclinés sous forme numérisée, mais dispersés sur divers types de supports, et dans différents formats: en ligne sur des serveurs à l'étranger et en Suisse, sur des disques externes autonomes, sur des DVD ou des CDROM. L'urgence est maintenant double: d'une part les mettre en ligne à disposition des intéressés, et d'autre part les préserver pour le long terme. Deux objectifs qui sont à l'origine de nouveaux projets de grande ampleur.

Quelle vie après la numérisation?

Concernant l'accès en ligne, il semble que l'on ait bien compris aujourd'hui que ce n'est pas la localisation physique des documents numérisés qui importe, mais bien plutôt celle des métadonnées qui leur donnent accès. Ainsi peu importe que les fichiers issus des numérisations de la BCU se trouvent sur les serveurs de Google en Californie, sur le serveur RERO DOC à Martigny, sur le serveur académique lausannois SERVAL ou ailleurs; l'essentiel est qu'ils soient atteignables là où ils se trouvent depuis plusieurs portails – depuis n'importe quel portail si on veut bien – grâce à des identifiants de localisation permanents de type URN. Et ce sont ces identifiants URN, présents dans GBS pour les livres numérisés par Google, que nous remontons après-coup dans le catalogue collectif de RERO, dans le portail SERVAL si nécessaire, bientôt dans d'autres portails suisses en préparation comme SWISSBIB ou dans la base de données européenne EROMM, et pourquoi pas dans EUROPEANA, si nos collections peuvent intéresser un jour le portail européen en gestation. L'essentiel en fait est que ces données se trouvent désormais impérativement sur les outils de recherche privilégiés et prioritaires des intéressés, outils qui seront nécessairement très divers en raison des nombreuses communautés d'intérêt représentées (le portail «MyUni» pour les lecteurs d'une bibliothèque universitaire précise, des portails thématiques pour des chercheurs en sciences humaines disséminés dans toute l'Europe, une base comme EROMM pour des bibliothécaires à la recherche de copies déjà numérisées d'ouvrages de leurs collections, EUROPEANA pour le grand public curieux, etc.). Il est sans doute prématuré de s'imaginer que le catalogue d'OCLC pourrait être un jour le point d'entrée commun utilisé par tous les usagers des bibliothèques de la planète, où qu'ils se trouvent, mais les signes qui pointent dans cette direction méritent attention: il faut suivre très attentivement le projet que vient d'initier OCLC d'enrichir considérablement son catalogue par agrégation de catalogues d'autres grands réseaux et bibliothèques de par le monde, de même que le partenariat qu'il vient de conclure en mai 2008 avec Google pour faciliter l'accès direct à plus d'un million de documents en texte intégral à partir de son formidable réservoir de métadonnées Worldcat. Alors, en attendant, mettons notre énergie à multiplier intelligemment les points d'accès aux bons endroits plutôt que de nous obstiner à dupliquer sur des serveurs «à nous» nos contenus numériques gourmands en espace de stockage!

Garder les moules

Surtout que de l’espace de stockage, il en faut pour archiver convenablement et de manière pérenne les résultats bruts de la numérisation. Et là c’est une autre paire de manches! Car si les ver- sions accessibles en ligne sont volontai- rement compactées dans des formats de présentation allégés (PDF la plupart du temps), et occupent de ce fait relati- vement peu de place sur les serveurs, ce sont les résultats bruts de numérisa- tion en format riche (des «master» en TIFF généralement) qu’il faut pérenni- ser, eux qui permettront, si nécessaire, de refaire des versions allégées en cas d’accident. Et encore ne parle-t-on ici que de documents à prédominance tex- tuelle, numérisés en nuances de gris; les fichiers qui résultent de numérisa- tions couleur affichent pour leur part une corpulence tout à fait impression- nante. Maintenant que de nombreux

projets de numérisations sont lancés quasi partout, maintenant que les bon- nes pratiques semblent établies, que les formats sont maîtrisés et stabilisés, bref que ce genre d’exercice ne relève plus de l’intrépidité de quelques pionniers, le problème prioritaire qui se pose est de ne pas perdre le fruit de tous ces ef- forts, soit par accident (destruction bru- tale des supports), soit par négligence (obsolescence imperceptible des sup- ports et appareils). L’archivage pérenne, qui devrait être garanti au sein de dé- pôts sécurisés («trustworthy reposito- ries»), fait à juste titre l’objet de toutes les attentions dans la profession, sans toutefois parvenir encore à attirer l’at- tention du politique qui peut seul déga- ger les ressources financières nécessai- res, et qui sont considérables. Saluons cependant l’existence d’une norme ISO (14721) qui définit les conditions ca- dres, sur les plans organisationnel, po- litique et financier, auxquelles un tel archivage qui se veut pérenne doit sa- tisfaire (là encore, voir notamment l’autre excellent didacticiel – en fran- çais! – préparé par Cornell: www.library. cornell.edu/iris/tutorial/dpm-french/ foundation/tdr/compliance.html). Dans ce domaine également, OCLC propose désormais des espaces d’archivage pé- renne communs pour les bibliothèques qui manquent des moyens nécessaires à la mise en œuvre de leur propre ar- chive.

Autres problèmes rencontrés

Ces pérégrinations au travers de plusieurs projets de numérisation touchant des types de documents assez variés ont permis aux bibliothécaires de la BCU de se familiariser avec les points qui font problème, et qui interdisent pour l’instant une forte extension des projets de numérisation.

Premier point technique, le type de numérisation: on l'a dit, il n'est pas raisonnable aujourd'hui, sauf pour quelques catégories de documents, de numériser systématiquement en couleur des objets de nature textuelle, étant donné la taille excessive des fichiers qui en résulte et qui pénalise aussi bien le stockage que l'affichage. Pour une numérisation à la pièce de documents précieux en revanche, cela se révèle indispensable, comme cela a été fait pour le somptueux DVD qui présente les richesses de la Bibliothèque de Genève. Un deuxième point technique concerne le taux de réussite de la reconnaissance automatique des caractères (OCR) pour les documents textuels, qui doit dépasser 99% pour être acceptable, mais n'atteint pas encore ce seuil pour certains ouvrages anciens ou imprimés en alphabets non latins (difficulté de distinguer le «s long» du «f» dans certaines typographies du XVIIIe siècle par exemple). De nombreux projets publics ou privés s'activent à tenter d'améliorer l'efficacité de cette catégorie de logiciels (par ex. le projet IMPACT financé par la Communauté européenne), et il y a bon espoir que les performances de ces outils s'améliorent sensiblement. Bonne surprise, par ailleurs, au niveau de l'identification et de la localisation des documents, lorsque nous avons vu que Google travaillait avec des identifiants de ressources permanents (URN): les documents référencés sont ainsi identifiés de manière univoque quelle que soit leur localisation (URL), ce qui garantit en principe qu'ils resteront accessibles même s'ils sont déplacés sur d'autres serveurs. Ce sont par conséquent ces URN que nous remonterons dans le catalogue collectif de RERO. Sous l'égide de la Bibliothèque nationale, la Conférence des bibliothèques universitaires (CBU) a pris récemment certaines décisions qui précisent les modalités d'attribution des URN en Suisse. C'est cependant sur un autre plan que l'avènement de la très grande bibliothèque numérique est entravé, à savoir celui des droits: on ne peut aujourd'hui numériser et mettre en ligne que des ouvrages libres de droits, soit des œuvres dont l'auteur est décédé depuis plus de 70 ans. C'est pour cette raison que Google, par extrême prudence, ne retient aucun ouvrage du XXe siècle parmi ceux que la BCU envoie en numérisation. Or il est bien évident que ce sont essentiellement les ouvrages les plus récents que le grand public aimerait voir en ligne. Et l'indécision persiste quant à l'évolution des choses en matière de protection des œuvres, qui peut tout aussi bien conduire à un raidissement encore plus marqué de la législation actuelle – ce qui pourrait inhiber toute nouvelle initiative de numérisation – comme à un assouplissement raisonnable, à la manière de ce qui s'est passé pour la musique en ligne. Le domaine de la gestion des droits numériques (DRM) est encore en pleine évolution.

Dans le cas d'un scénario optimiste (pour le bibliothécaire et ses usagers s'entend), c'est-à-dire où il serait dorénavant loisible à certaines conditions de numériser des œuvres actuellement sous droits, on pourrait alors proposer à Google de poursuivre, après la numérisation des collections anciennes, par les ouvrages typiquement vaudois du XXe siècle, notamment tous ceux qui parviennent à la BCU par le biais du Dépôt légal, et que la bibliothèque s'est donné pour mission de les conserver «pour l'éternité».

A titre d'illustration de cette problématique générale des droits, il faut mentionner les restrictions d'accès qui peuvent être imposées par un tiers qui aura contribué financièrement à l'effort de numérisation, et qui se sera réservé les droits d'accès dans un certain périmètre, ou durant un certain laps de temps. Les contrats en discussion avec des rédactions de quotidiens qui détiennent les droits sur leurs titres contiennent des clauses qui vont dans ce sens. Et tout l'art de la négociation consiste à trouver le meilleur compromis qui préserve à la fois l'intérêt des collectivités publiques et l'idéal du libre accès à l'information pour tous d'une part, et une rétribution équitable des bailleurs de fonds et détenteurs de droits d'autre part.

Car le manque de financement reste l'obstacle majeur pour les institutions publiques une fois que les autres problèmes ont été résolus (et la plupart sont en passe de l'être). Il y a à peine plus d'une année, la directrice de la Bibliothèque nationale suisse soulignait, dans une interview accordée au *Temps,* à quel point la dotation financière en matière de numérisation était inexistante en Suisse, à quelque échelon que ce soit. Les choses auraient-elles soudain beaucoup changé? Rien n'est moins sûr. Au point que certaines bibliothèques ont récemment tenté de solliciter des crédits extraordinaires alloués par la Confédération dans le cadre du projet de «bibliothèque numérique suisse» (E-lib.ch) pour financer la numérisation de certaines de leurs collections, alors que cet argent n'est pas prioritairement destiné à être affecté à des contenus. Quoi qu'on en pense en définitive, le partenariat signé avec Google – qui, précisons-le, prend en charge tous les frais de la numérisation proprement dite – représentait en 2007 la seule et unique alternative à l'expectative, pour qui voulait lancer une opération de numérisation d'une certaine envergure.

En guise de bilan

En définitive, il est vrai que, plutôt que d'attendre que tout soit balisé pour se lancer, la BCU a privilégié une politique plus audacieuse et de ce fait plus risquée, à la Napoléon si l'on peut dire («à la guerre, d'abord on s'engage, et puis l'on voit»). Mais avec un peu de recul cependant, on discerne qu'une certaine cohérence se dessine dans ce faisceau d'actions lancées sans plan préalable mûrement établi, et qui permet néanmoins d'échafauder une politique de numérisation qui combine judicieusement le travail à la chaîne et le sur-mesure. Ainsi, en participant avec Google à la constitution du plus vaste réservoir d'ouvrages numérisés jamais imaginé, la BCU, par le biais de sa contribution, procède à la numérisation «industrielle» et quasi systématique de ses collections anciennes, à l'exception de quelques ouvrages qui constituent le cœur des collections précieuses: quels que soient leur degré d'utilisation, leur langue, leur sujet, leur provenance, les livres de la bibliothèque publiés jusqu'à la fin du XIXe siècle sont numérisés systématiquement au cours de cette opération à flux tendu. Le fait que l'essentiel du poids financier de cette opération soit assumé par le partenaire permet par contrecoup à la BCU d'allouer de manière plus ciblée et plus judicieuse ses maigres budgets à des opérations sur-mesure, conduites avec d'extrêmes précautions, qui vont concerner tout naturellement les ouvrages les plus précieux et les plus rares de son patrimoine, les manuscrits ou les fonds iconographiques.

Même si elle entre dans l'avenir à reculons, la BCU, au travers de ces différents projets complémentaires de rétronumérisation, progresse de manière décidée vers la réalisation de la «BCU hybride», à savoir une bibliothèque qui soit à même de gérer de manière cohérente et harmonisée la documentation imprimée ET la documentation numérique, comme deux volets complémentaires et non exclusifs d'un riche ensemble de prestations.

Hubert Villard

Ancien directeur de la Bibliothe?que cantonale et universitaire de Lausanne

Abstract

Deutsch

Digitalisierung in der Kantonal- und Universitätsbibliothek Lausanne (BCU Lausanne) Die nachträgliche Digitalisierung der Printbestände geht auf allgemeine Automatisierungsarbeiten Anfang der 1970er-Jahre zurück (Stichwort System SIBIL). Anfang der 1990er-Jahre war diese Automatisierung abgeschlossen, d.h. sämtliche traditionellen Funktionen einer Bibliothek wie Verwalten des und Suchen im Katalog waren automatisiert. Im Rahmen eines weiterführenden Projekts (SIBIL 2000) wurde erkannt, dass die Herausforderung in Zukunft nicht nur bei der Automatisierung der Kataloge liegen würde, sondern bei der Digitalisierung der gesamten Bestände. Anfang der 1990er-Jahre wurden die «Schätze» der BCU auf Mikrofilm gespeichert. Trotz fehlender Ressourcen wurden deshalb bereits erste Bestände digitalisiert. Diese wurden vorerst auf DVD gesichert. Mit der Verbreitung von bestimmten Standards wie etwa ISO 14721 (PDF für die Konsultation online, TIFF für die «digitale Konservierung») konnten einige Digitalisierungsprojekte (v.a. Zeitungsbestände) realisiert werden. Weitere Projekte sind aufgegleist – das grösste Problem ist nicht mehr die Technik, sondern das Geld ... In die Digitalisierungsprojekte einbezogen werden übrigens auch Dokumente, die zwar ausserhalb der BCU gelagert werden, aber zum kulturellen Erbe des Waadtlandes gehören. Angesichts der fehlenden Ressourcen wurden externe Partner gesucht. So wurde 2007 mit Google ein Vertrag über die Digitalisierung von 100 000 urheberrechtlich freien Büchern aus den Beständen der BCU (in einem Zeitrahmen von 2 Jahren) unterzeichnet. Hier stellten sich urheberrechtliche Probleme, Google digitalisiert keine Werke aus dem 20. Jhrdt. Angesichts der knappen Ressourcen wird die Zusammenarbeit mit neuen Partnern gesucht. Die Projekte EUROPEANA und das Projekt zur «Weltbibliothek», welches das Online Computer Library Center OCLC angerissen hat, werden aufmerksam verfolgt.