

■ L'AUTEUR



Frédéric BLIN
est directeur
de la conservation
et du patrimoine
de la Bibliothèque
nationale
et universitaire de Strasbourg.

En réalité, si l'on considère les établissements n'ayant pas répondu à l'enquête, cette proportion doit être bien inférieure...

Dans le monde scientifique aussi, la numérisation est devenue courante depuis une petite quinzaine d'années. En France, aux programmes de numérisation de revues scientifiques, tels que NUMDAM en mathématiques et PERSEE pour les sciences humaines et sociales, ont succédé les projets soumis à l'Agence nationale pour la recherche, puis aux Investissements d'excellence. Une

science – les humanités numériques – est en fort développement aux États-Unis, en Europe et partout dans le monde, désignant l'utilisation d'outils informatiques avancés pour exploiter des contenus numériques de manière innovante. En France, la Bibliothèque virtuelle humaniste (BVH), pilotée par le Centre d'études supérieures de la Renaissance de Tours, et *Bibliissima*, de l'Institut de recherche et d'histoire des textes du CNRS, sont deux des projets les plus emblématiques de cette nouvelle science, qui a aujourd'hui ses propres Très grandes infrastructures de recherche (TGIR): *Humanum* en France et *Dariah* en Europe.

La Bibliothèque scientifique numérique (BSN), TGIR lancée en 2008 par le ministère de l'Enseignement supérieur et de la Recherche, vise à structurer le paysage et à coordonner les initiatives liées à l'information scientifique et technique à l'échelle nationale, en particulier celles mettant en œuvre la numérisation du patrimoine scientifique national. L'appel à projets lancé en avril 2013 par son segment 5 (BSN5-Numérisation du patrimoine scientifique) a ainsi fait remonter près de 75 réponses – des projets issus de disciplines aussi diverses que la paléontologie, la chimie, l'histoire de l'art ou encore les sciences économiques, témoignant non seulement de la richesse des initiatives, mais aussi des potentialités infinies que la numérisation ouvre à la recherche.

Le paradoxe du droit d'auteur

La plupart de ces projets se sont heurtés, à des degrés divers, à la question du droit d'auteur. Courant en général sur 70 années après la mort de l'auteur, ce droit contraint les porteurs de projets à s'assurer de l'accord des ayants droit avant de numériser une œuvre encore protégée. La recherche de ces ayants droit peut se révéler longue et difficile, et donc coûteuse. Deux solutions sont alors possibles: soit numériser, attendre que les ayants droit se manifestent et, si besoin, retirer ensuite les ouvrages concernés, soit se contenter de numériser des ouvrages libres de droit. La première alternative a été le choix de la société Google. Elle l'a entraînée dans des procédures juridiques avec les éditeurs de multiples pays. La seconde est celle adoptée par les bibliothèques.

Celles-ci ne disposant pas des ressources pour effectuer les recherches des ayants droit, un registre, ARROW (*Accessible Regis-*

Numériser un document

La numérisation d'un document se déroule en quatre étapes principales: la prise de vue (ou l'enregistrement en cas de matériel audiovisuel), la structuration, l'enrichissement et la description du document numérique, l'archivage des fichiers et, enfin, la mise à disposition du public.

La prise de vue est l'étape la plus simple. Dans le cas d'un document imprimé, une résolution de numérisation à 300 pixels par pouce est suffisante pour la majorité des usages, tout en limitant le poids relatif des fichiers. Mais le matériel utilisé détermine aussi la qualité de l'image et la bonne conservation du document physique. Des modèles de scanners patrimoniaux ont ainsi été développés en partenariat avec des bibliothèques pour garantir une conservation optimale de l'original.

La deuxième étape peut être assez complexe, selon les moyens et les objectifs du projet. Le premier type d'enrichissement est la reconnaissance optique de caractères (OCR), pour transformer une simple image en un texte reconnaissable par un logiciel informatique. Les logiciels de reconnaissance les plus performants atteignent aujourd'hui plus de 99 pour cent de réussite, pour des textes imprimés des XIX^e et XX^e siècles en caractères latins. Des projets de recherche sont en cours pour les documents plus anciens (projet européen IMPACT) ou les manuscrits, beaucoup plus difficiles à traiter.

La structuration informatique des documents permet d'identi-



Une page de l'*Autobiographie du bienheureux Henri Suso*, un livre manuscrit sur parchemin du XIV^e siècle. Des projets de recherche visent à mettre au point des logiciels capables de reconnaître les textes de tels manuscrits anciens.

fier des parties des ouvrages, offrant par exemple un accès direct aux chapitres. Divers schémas de structuration existent. Les plus courants pour l'imprimé sont le METS et la TEI, ou l'EAD pour les manuscrits et archives. Enfin, la création de métadonnées permet d'identifier et de repérer le document dans l'océan de données du Web. Le standard de métadonnées minimal aujourd'hui, en bibliothèque, est le *Dublin Core*.

L'archivage doit être pensé en amont du projet: la résolution des images doit optimiser le poids des fichiers par rapport aux usages attendus, les méta-

données doivent être assez précises pour que l'on puisse retracer l'historique et l'identité des fichiers numériques, le cycle de vie des archives doit être aux normes, et le coût de la conservation doit être pris en compte.

Enfin, la diffusion doit répondre aux attentes des utilisateurs en s'accompagnant de services et de données telles que les conditions d'utilisation. Le statut (*Creative Commons*, licence ouverte, etc.) sous lequel l'institution décide de placer ses images et ses données numérisées conditionnera les usages qui pourront en être faits.

Cote BNU: Ms. 2.929. Photo et coll. BNU Strasbourg

tries of Rights information and Orphan Works towards Europeana), financé par la Commission européenne, a été démarré en 2008 pour faciliter cette recherche. Cependant, ARROW ne peut à lui seul identifier l'ensemble des œuvres sous droits et leurs ayants droit légitimes. Aussi, en raison de la réglementation sur le droit d'auteur, il est plus facile aujourd'hui de trouver sur Internet un ouvrage de 1860 que de 1960. Un pan entier de la littérature, celle du XX^e siècle, est en grande part inaccessible en ligne, alors même que la valeur commerciale des documents concernés est faible, voire nulle, et que ceux-ci sont souvent indisponibles – dans le commerce, sous format papier voire même en format électronique – ou orphelins – sans ayant droit connu.

Le programme de numérisation des œuvres indisponibles lancé par l'État en 2013 *via* la Bibliothèque nationale de France vise justement à numériser 500 000 œuvres du XX^e siècle pour combler ce vide, mais les modalités – notamment de financement – de ce programme sont très critiquées par une partie importante des cercles professionnels. Une loi sur les œuvres indisponibles est en préparation au niveau européen, bénéficiant du soutien des associations de bibliothèques. Celles-ci sont aussi moteur d'un projet de traité international sur les exceptions et les limitations au droit d'auteur en discussion au sein de l'Organisation mondiale de la propriété intellectuelle.

De nouvelles archives à préserver

La pérennité des documents numérisés est aussi au cœur des réflexions. Une erreur fondamentale serait de croire que la préservation de la version physique originale d'un document numérisé serait inutile. D'une part, l'original a une valeur patrimoniale en soi et, d'autre part, la durée de la pérennité des données numériques n'est aujourd'hui de l'ordre que d'une trentaine d'années. Numériser pour faire de la place sur les rayonnages est pour l'heure un danger réel pour la transmission du patrimoine et de la mémoire. Les exemples de pertes des données conservées sur des CD-ROM sont légion, sans parler de la qualité technique des numérisations réalisées il y a quelques années, insuffisante par rapport aux besoins d'aujourd'hui.

Quelques chiffres

130 millions de livres ont été publiés dans le monde depuis l'invention de l'imprimerie, selon *Google*, dont 20 % sont dans le domaine public.

20 % des collections des établissements culturels européens étaient numérisés en 2012, selon l'enquête *Enumerate*.

30 millions de documents sont accessibles *via Google Books*, 26 millions *via Europeana*, 11 millions *via Hathi Trust* et 2,2 millions *via Gallica*.

5 000 euros : c'est le coût annuel moyen de la préservation à long terme d'un téraoctet de données numérisées.

Les bibliothèques ne se sont intéressées à la préservation à long terme des documents numérisés qu'assez récemment, à partir de 2000 pour les plus avancées, et la norme en vigueur – l'OAIS (Système ouvert d'archivage d'information) – n'a été définie qu'en 2002. Si les bibliothèques nationales ont développé leurs propres infrastructures d'archivage (SPAR pour la Bibliothèque nationale de France, E-Depot pour la Bibliothèque royale des Pays-Bas, etc.), les autres bibliothèques recourent à des prestations extérieures d'archivage pour leurs propres collections, soit auprès de leur bibliothèque nationale, soit auprès d'autres organismes, privés ou publics.

En France, dès 2008, le Centre informatique national de l'enseignement supérieur (CINES) a été missionné pour l'archivage pérenne des bibliothèques universitaires par le ministère de l'Enseignement supérieur et de la Recherche (voir l'article de F. Daumas et M. Massol dans ce numéro). Le CINES est notamment au cœur du dispositif envisagé par la Bibliothèque scientifique numérique pour archiver les contenus numériques produits par l'enseignement supérieur national.

Néanmoins, avec un coût estimé aujourd'hui à 5 000 euros par téraoctet et par an, l'archivage numérique représente une charge financière considérable pour les établissements. Ainsi, des politiques documentaires spécifiques ont été définies par les bibliothèques, déterminant des priorités parmi les natures de données à préserver. Toutefois, la question de la préservation à long terme des données numériques reste largement ignorée des porteurs de projets de numérisation. Et l'ampleur des coûts fait peser un risque fort sur la pérennité des résultats de ces projets.

Avec le numérique, les bibliothèques – notamment les bibliothèques universitaires et de recherche – ont été bouleversées en profondeur. Pourtant, leurs missions fondamentales – préserver le patrimoine documentaire, rendre accessible le savoir, accompagner leurs usagers dans la découverte et l'usage de ce savoir noyé dans l'immensité – restent intactes : le numérique leur permet d'accomplir ces missions avec davantage encore de pertinence et d'efficacité. Les professionnels de l'information numérique sont devenus, et deviendront encore davantage, des partenaires incontournables des projets de recherche, ainsi que de l'enseignement à distance, amené à révolutionner l'université dans les prochaines années. ■

■ BIBLIOGRAPHIE

A. Jacquesson, *Google Livres et le futur des bibliothèques numériques. Historique du projet, techniques, documentaires, alternatives et controverses*, Cercle de la librairie, 2012.

Th. Claerr et I. Westeel (sous la dir.), *Manuel de la numérisation*, Cercle de la librairie, 2012.

J.-Ph. Accart et A. Rivier, *Mémento de l'information numérique*, Cercle de la librairie, 2012.

D. Baker et W. Evans (éd.), *Digital Library Economics. An academic perspective*, Chandos Publishing, 2009.

Bibliothèque scientifique numérique : bibliothequescientifique-numerique.fr

Gallica : gallica.bnf.fr