

La numérisation des bibliothèques

Frédéric Blin

Depuis une vingtaine d'années, les bibliothèques du monde entier numérisent leurs collections et les rendent accessibles sur Internet. Ce nouveau rôle bouleverse les modes d'accès au savoir.

On trouve tout sur Internet, on n'a plus besoin de bibliothèques : cette phrase est devenue si banale qu'elle ne choque plus, ou si peu. Et pourtant. Peut-on tout trouver sur Internet ? Les centaines de millions de livres, journaux, revues, images, objets conservés par nos bibliothèques ont-ils été déjà tous numérisés ? Et ce qui existe sous forme numérique est-il accessible sur Internet, où que l'on se trouve ? Quel est le rôle des bibliothèques désormais, dans ce monde où l'information et le savoir sont de plus en plus exclusivement numériques ?

Au cœur des réflexions professionnelles des bibliothécaires, ces questions devraient intéresser un public bien plus large, notamment au sein des communautés universitaires et scientifiques. Le numérique a entraîné un bouleversement des modes d'accès aux savoirs. Toutefois, la facilité relative avec laquelle cet accès se fait masque divers dangers. Tout d'abord, les savoirs et informations non numérisés, délaissés, risquent de disparaître ou d'être oubliés. Ensuite, la

médiation entre les contenus informatifs et les utilisateurs de ces contenus, réalisée par les bibliothécaires, est remise en cause par la communication directe sur Internet, mais aussi par le contrôle que d'autres acteurs – ceux qui tirent un profit économique de l'information – exercent de plus en plus sur la possibilité même d'y accéder. Enfin, le déluge d'informations et de données, qui s'accroît chaque jour, soulève la question de nos capacités à y naviguer de façon pertinente et à tout conserver ou, du moins, à préserver ce qui sera utile aux générations futures, en particulier pour les besoins scientifiques et mémoriels.

Du projet Gutenberg à Google Books

La première initiative de numérisation de livres, le projet *Gutenberg* piloté par l'Américain Michael Hart, de l'Université de l'Illinois, aux États-Unis, remonte

à 1971. Ce projet, qui se poursuit, repose non

pas sur des bibliothèques, mais sur les contributions de volontaires. Ceux-ci numérisent des ouvrages du domaine public et participent à leur transformation en textes numériques. Les principales bibliothèques du monde, quant à elles, ont commencé à numériser leurs collections au début des années 1990, pour deux raisons principales. D'une part, il s'agissait d'assurer une meilleure préservation des documents : la numérisation était alors vue comme un moyen de remplacer à terme le microfilm. D'autre part, l'objectif était de faciliter l'accès du public à des collections fragiles et précieuses.

En France, *Gallica*, la bibliothèque numérique de la Bibliothèque nationale, a ainsi ouvert en 1997, donnant alors accès à une sélection de ses trésors les plus précieux. D'autres grandes bibliothèques lui ont emboîté le pas dans les années suivantes : la



Bibliothèque interuniversitaire de médecine avec *Medic@*, le Conservatoire national des arts et métiers avec le Conservatoire numérique des arts et métiers (CNUM), la Bibliothèque municipale de Lyon... La numérisation des fonds anciens est devenue une activité quotidienne, si ce n'est banale, des bibliothèques françaises, et les ouvertures de bibliothèques numériques modernes se sont multipliées au cours des dernières années: *NumeLyo* à Lyon, *Rosalis* à Toulouse, *Numistral* à Strasbourg, ou les bibliothèques numériques régionales d'Aquitaine, de Bretagne, de Lorraine, du Limousin, ou encore *Manioc* en Antilles-Guyane... Même les plus petites mettent en valeur leurs fonds anciens grâce à des outils partagés, telle la bibliothèque numérique *e-corpus*, développée par le Centre de conservation du livre d'Arles, qui réunit aujourd'hui les fonds de plus de 250 établissements.

La France n'est à cet égard ni en avance ni en retard par rapport aux autres pays comparables. Depuis quelques mois, l'Allemagne dispose de sa bibliothèque numérique nationale, la *Deutsche Digitale Bibliothek*. Nourri en partie de programmes financés par la Fondation allemande pour la recherche et des réalisations des trois principaux centres de numérisation du pays, à Göttingen, Munich et Berlin, ce projet ambitieux a pour objectif la numérisation de l'ensemble des publications du pays antérieures à 1800. Dans le monde hispanophone, la bibliothèque numérique *Miguel de Cervantes*, hébergée à l'Université d'Alicante, fait figure de référence depuis 1999.

En Chine, c'est tout un réseau national qui se construit à partir de la Bibliothèque numérique nationale ouverte en 2007. D'ici 2015, ce réseau reliera les plates-formes de bibliothèques numériques développées localement par les universités ou bibliothèques publiques, donnant à chacune d'elles accès non seulement au patrimoine numérisé, mais aussi à des ressources et à des services numériques et multimédia.

En Afrique, plusieurs projets de soutien au développement de bibliothèques numériques ont vu le jour depuis une dizaine d'années. Parmi eux, le projet *African Digital Libraries Support Network*, lancé en 2010 avec le soutien de la fondation internationale EIFL (*Electronic Information for Libraries*), a permis de créer un réseau de centres d'expertise dans huit pays africains dont l'objectif est

Lexique

Humanités numériques :

Cette nouvelle science utilise les méthodes et les outils du numérique en sciences humaines et sociales.

BSN : La Bibliothèque scientifique numérique coordonne en France les initiatives favorisant l'accès à l'information scientifique.

ARROW : Ce dispositif européen vise à simplifier la gestion des droits dans les projets de numérisation.

OAIS : Ce modèle de gestion, d'archivage et de préservation des données numériques est la norme en vigueur.



l'accompagnement de projets de constructions de bibliothèques numériques.

Citons enfin la toute récente Bibliothèque numérique publique d'Amérique (*Digital Public Library of America*), dont la mission est de donner accès aux considérables patrimoines des bibliothèques, universités, archives et musées des États-Unis, afin d'éduquer et d'informer, et de transmettre ce patrimoine à l'ensemble des citoyens américains.

Dans ce paysage, *Google Books* occupe une place particulière. Le lancement, en 2004, de ce projet visant à numériser et à diffuser l'ensemble des livres du monde a marqué un tournant majeur dans l'histoire des bibliothèques numériques. Pour la première fois, une entreprise privée disposait du potentiel de contrôle sur le patrimoine documentaire mondial, concurrençant directement les bibliothèques. Celles-ci ont dû se positionner, à l'instar de la Bibliothèque nationale de France: son passage d'une logique de numérisation sélective à une logique de numérisation systématique s'explique par la volonté de constituer une masse documentaire conséquente qui lui permettrait de devenir le plus légitime vecteur du patrimoine français numérisé.

Les bibliothèques réagissent

La bibliothèque numérique *Europeana*, quant à elle, a été créée en 2008 pour redonner aux États européens l'initiative sur la mise en valeur de leur patrimoine. Grâce notamment au fort soutien de la Commission européenne aux programmes de numérisation collaboratifs visant à l'enrichir, elle compte plus de 20 millions de documents et d'objets. Aux États-Unis, *Hathi Trust*, construit par les bibliothèques ayant participé aux débuts de *Google Books* à partir du fonds d'ouvrages numérisés par *Google*, est devenu, en volume de titres numérisés, le plus proche concurrent de *Google*.

Malgré la richesse et la multiplicité des initiatives, les dizaines voire les centaines de millions de documents déjà numérisés ne représentent qu'une petite partie des richesses documentaires présentes dans les bibliothèques. Ainsi, en 2012, le projet *Enumerate*, financé par la Commission européenne, a estimé que seuls 20 pour cent des collections des institutions culturelles européennes étaient déjà numérisés.

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 numériques conditionnera les usages qui pourront en être faits.