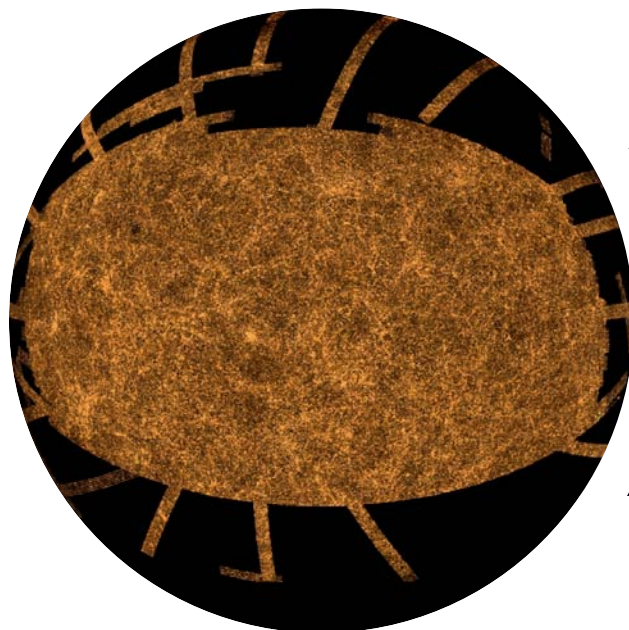
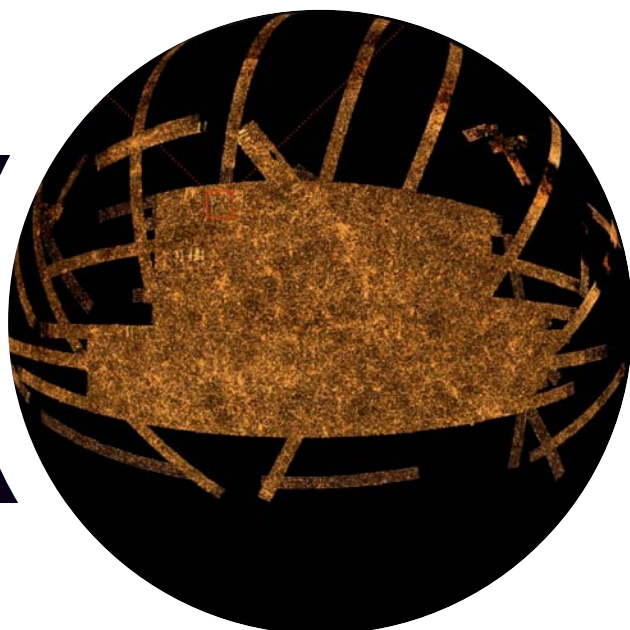




d'une centaine de nouvelles exoplanètes géantes, ainsi qu'un nombre similaire de naines brunes, astres intermédiaires entre les plus petites étoiles et les planètes gazeuses.

Enfin, le quatrième volet de SDSS, l'expérience APOGEE (*Apache Point Observatory Galactic Evolution Experiment*), consiste à observer la Voie lactée dans l'infrarouge proche. Ce domaine du spectre électromagnétique permet de pénétrer à travers la poussière qui obscurcit la plupart des étoiles situées dans le disque et le bulbe galactiques.

APOGEE utilise un spectrographe à haute résolution très performant. Les astronomes mesurent grâce à lui la composition chimique et la vitesse (le long de la ligne de visée) de 100 000 étoiles géantes, ce qui fournit un aperçu sans précédent sur la structure dynamique et l'histoire chimique de la Voie lactée.

Ainsi, SDSS constitue un projet astronomique de grande ampleur, qui a déjà apporté de nombreuses découvertes. Cette immense observation du ciel, qui couvre 14 500 degrés carrés sur la voûte céleste, a répertorié 470 millions d'astres. Et parmi eux, près de deux millions de galaxies et 500 000 étoiles sont la cible d'études spectroscopiques.

La quantité d'informations produites par SDSS est énorme. Le volume total des données actuellement accessibles en ligne est de 71 téraoctets, ce qui équivaut à l'information contenue dans une bibliothèque comptant 71 millions de livres de 1 000 pages, soit bien plus qu'il n'y en a à la Bibliothèque nationale de France...

Un travail colossal est nécessaire pour traiter les données brutes provenant des instruments d'observation et en extraire les informations sur les paramètres physiques des objets astronomiques. Bien sûr, une grande partie de ce travail est effectuée par des ordinateurs. Pour BOSS, par exemple, le traitement est réalisé au Laboratoire américain Lawrence Berkeley par une grappe de 50 machines, chacune comportant huit cœurs (huit processeurs). Les capacités de stockage doivent être également à la hauteur du projet : les différents nœuds de stockage des données brutes et traitées totalisent une capacité de 250 téraoctets...

Il reste qu'une centaine de scientifiques doivent travailler des mois entiers pour préparer les données et les rendre publiques. En astronomie, on a pour tradition de rendre les données d'observation accessibles à tous. Celles du SDSS renferment sans aucun doute de nombreuses découvertes pour les années à venir. ■

CES DEUX IMAGES DE LA VOÛTE CÉLESTE

(au-dessous du plan de notre galaxie, à gauche ; au-dessus du plan galactique, à droite) ont été construites avec

les données du relevé astronomique SDSS.

Chaque point lumineux est une source détectée. Les régions du ciel qui n'ont pas été observées sont en noir.

■ L'AUTEUR



Céline REYLÉ est astronome à l'Institut UTINAM, à Besançon, et directrice de l'Observatoire des sciences de l'Univers THETA Franche-Comté/Bourgogne. Elle participe au projet SDSS.

■ SUR LE WEB

Site du projet SDSS : www.sdss.org

Francis Daumas et Marion Massol

Au cœur d'un centre d'archivage

Pour stocker le déluge récent de données numériques, notamment scientifiques, de grands centres d'archivage ont été mis en place. Comment fonctionnent-ils ? Exemple avec le CINES, un centre national à Montpellier.

La plupart du temps, quand nous souhaitons conserver un document numérique, nous le copions sur un support externe, que nous stockons dans un placard. Mais sa récupération longtemps après peut se révéler compliquée. D'abord, on note rarement, ou de façon imprécise, la localisation du support et les informations pertinentes sur le contenu. Si l'on parvient à retrouver le support, encore faut-il pouvoir le lire : imaginez une photo de mariage stockée sur une vieille disquette des années 1980 ! Les équipements de lecture de ce support ont parfois disparu ou sont difficiles à trouver après tant d'années. Supposons que vous ayez trouvé le bon équipement : vous devez aussi disposer du logiciel capable de reconnaître le format du

fichier et de restituer le contenu sous une forme intelligible. Or les formats évoluent et les logiciels peuvent disparaître ou créer des fichiers qui ne seront pas lisibles par leurs futures versions. Enfin, le fichier initial ne doit pas comporter trop d'erreurs et son support doit avoir été bien conservé.

Cet exemple illustre les difficultés posées par la préservation à long terme des données : l'obsolescence et la détérioration du support, la disparition de l'outil logiciel de lecture, les évolutions du format du fichier et la perte de la localisation et de la signification de son contenu. Pour répondre à ces difficultés, des techniques d'archivage ont été développées. Nous les présenterons à travers l'exemple d'un des principaux sites d'archivage français : celui du CINES (Centre

informatique national de l'enseignement supérieur), à Montpellier, un établissement public placé sous la tutelle du ministère de l'Enseignement supérieur et de la Recherche.

Des données variées

Aujourd'hui, l'information numérique est omniprésente et protéiforme. Les données à conserver sur le long terme sont de natures variées : administratives, culturelles, scientifiques (issues d'observations ou de calculs), etc. Les livres des bibliothèques, les thèses, les articles scientifiques, les encyclopédies sont numérisés (*voir le dossier « Le savoir numérisé » dans ce numéro*). En résultent une multitude de documents à préserver, sous divers formats : textes, sons, images, vidéos...

Leur quantité croît de façon exponentielle, comme les volumes de données produites.

Des centres d'archivage ont alors été développés dans plusieurs pays depuis une quinzaine d'années. Ils ont d'abord été l'œuvre d'organismes publics, tels des agences spatiales, des instituts de recherche, des bibliothèques nationales ou des services administratifs. Un marché privé s'est aussi mis en place, mais avec des qualités de service inégales et des garanties diverses sur la pérennité des données, assurée plus souvent à court ou moyen terme qu'à long terme.

En outre, de grandes infrastructures ont été créées, à des échelles nationales ou internationales. Par exemple, les projets européens PRACE (*Partnership for advanced computing in Europe*, ou Partenariat pour le calcul avancé en Europe) et Eudat (*European data infrastructure*, ou Infrastructure de données européenne) visent à se doter respectivement d'importants moyens de calcul (et de stockage des données associées) et d'une plate-forme collaborative de partage et de conservation des données scientifiques.

Plus de huit millions de fichiers archivés

En France, le pionnier de l'archivage numérique est le CNES (Centre national d'études spatiales), qui archive depuis 1998 des images prises par des satellites et de grands observatoires, ainsi que diverses données de calcul. Le CINES, quant à lui, dispose d'un service d'archivage numérique opérationnel depuis 2006. Il a une capacité de stockage de plusieurs pétaoctets (10^{15} octets), soit l'équivalent d'une pile de CD d'environ dix kilomètres de hauteur. Plus de huit millions de fichiers sont aujourd'hui archivés, occupant une centaine de téraoctets (10^{12} octets). Le CINES a reçu un agrément (une certification) délivré par le SIAF (Service interministériel des archives de France) pour les archives dites intermédiaires, c'est-à-dire les documents qui ne sont pas nécessaires à l'activité quotidienne des services, mais qui doivent être conservés un certain temps pour des raisons administratives ou juridiques. La Commission européenne lui a aussi délivré une accréditation validant la qualité de son système d'archivage électronique.

Le CINES est chargé d'archiver toutes les thèses de doctorat françaises au format numérique. Il participe au projet de Bibliothèque scientifique numérique nationale, ou BSN (voir l'article de Frédéric Blin dans

L'ESSENTIEL

- D'énormes quantités de données numériques sont stockées dans de grands centres d'archivage dotés d'infrastructures sécurisées.
- Leur préservation à long terme pose plusieurs difficultés : détérioration et évolution des supports, formats de fichiers qui se périment, etc.
- On y répond le plus souvent par une gestion active des archives, incluant la migration des documents vers de nouveaux supports et l'actualisation de leur format.

ce numéro), où il pilote les questions liées à l'archivage du patrimoine scientifique public. Il héberge une partie des données de nombreux projets et organismes : la Très grande infrastructure de recherche (TGIR) HumaNum du CNRS (composée de moyens humains et numériques destinés aux sciences humaines), plusieurs bibliothèques universitaires, divers organismes de recherche (INSERM, IRSTEA, etc.), la cour des comptes... Il héberge aussi la première plate-forme numérique française de cours en ligne ouverts et massifs (CLOM, ou MOOC en anglais), dont le lancement a été annoncé début octobre 2013. En outre, il mène des actions de sensibilisation et de formation aux problématiques de l'archivage auprès de la communauté scientifique.

Outre l'archivage, le CINES propose des moyens de calcul numérique intensif à l'ensemble de la communauté scientifique. Le centre héberge notamment l'un des plus puissants supercalculateurs européens, financé par la société civile GENCI (Grand équipement national de calcul intensif) et utilisé pour des simulations complexes dans de nombreux domaines (climatologie, mécanique des fluides, physique des plasmas, astrophysique, géophysique, biologie, etc.). Archivage et calcul sont complémentaires, car les données servent d'entrées aux calculs qui, eux-mêmes, produisent de grandes quantités de données, dont une partie doit être conservée. Cependant, les volumes de données étant colossaux, on ne leur applique pas certaines procédures d'archivage à long terme (telle la copie du document en cinq exemplaires, sur laquelle nous reviendrons).

Une cinquantaine d'ingénieurs et de techniciens travaillent au CINES. Une douzaine d'entre eux se consacrent aux projets d'archivage, qui font aussi parfois appel à des intervenants extérieurs. Les profils sont variés : archivistes (spécialistes de la gestion des documents), informaticiens, juristes (garants du respect de la réglementation en vigueur), etc.

Des infrastructures sécurisées

Le centre s'est doté d'une infrastructure et d'une plate-forme logicielle adaptées à l'archivage. Les données y sont accessibles en permanence *via* le réseau Internet. L'alimentation électrique des équipements liés à l'archivage est redondante, afin de

prévenir les coupures de courant : elle est garantie par deux lignes électriques d'ERDF indépendantes et par des groupes électrogènes ; en outre, divers dispositifs de protection sont installés. Les conditions climatiques (température, humidité) des salles hébergeant les machines sont contrôlées. L'accès au site est sécurisé (filtrage des visiteurs, accès restreint à certaines zones), de même que les connexions *via* le réseau (grâce à des pare-feux, des protocoles sécurisés, des antivirus). En outre, toutes les données informatiques du centre (archives et systèmes) sont sauvegardées quotidiennement. Cette sauvegarde s'effectue de façon incrémentaire, c'est-à-dire qu'on n'enregistre que les changements.

mentale, c'est-à-dire qu'on n'enregistre que les changements.

La plate-forme d'archivage du CINES, conforme aux normes internationales du domaine, est fondée sur trois « serveurs logiques », respectivement dits de transfert, de stockage et d'accès. Le premier reçoit les documents à archiver, le deuxième les conserve et le troisième permet leur consultation. Décrivons-les plus en détail.

Un paquet d'information à archiver contient l'objet à conserver et un ensemble de métadonnées (des données permettant de comprendre d'autres données). Ces dernières sont élaborées en partie par l'organisme

dit versant, qui fournit les documents à archiver (c'est en général l'organisme qui les produit) et en partie par l'archiviste. Elles décrivent le contexte (organisme producteur, organisme versant), le document lui-même (titre, auteur, format), les paramètres de gestion (localisation physique, dispositif de stockage, etc.). Un autre type de métadonnées, qui ne sont pas associées à un document précis mais à la plate-forme d'archivage, décrit les informations dites de représentations, essentielles à la lecture de l'archive : les caractéristiques des formats, les structures des métadonnées, la description de la plate-forme d'archivage et des normes utilisées...

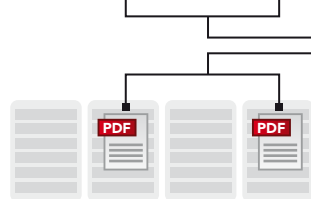
PLUSIEURS COPIES SUR DES SUPPORTS VARIÉS

Les supports physiques sont vulnérables et les estimations de leur durée de vie varient selon les études. Pour parer cette menace, cinq copies du document à archiver sont réalisées sur des types de supports différents : deux sur des bandes magnétiques et trois sur des disques durs (dont une située à plus de 300 kilomètres des autres, afin de préserver le document d'une catastrophe majeure, tel un séisme). Chaque document est accompagné d'un ensemble de métadonnées qui permettent de l'identifier, de le localiser et de faire évoluer son format.

Bandes magnétiques



Disques durs



PDF



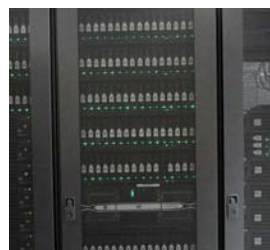
Disque dur dans un centre de données éloigné de plus de 300 km

Métadonnées :

Titre : L'archivage électronique
Auteur : Daumas, Francis
Identifiant : ark://85579/123456789
Date d'archivage : 30-09-2013
Format : PDF
Version du format : 1.7
...



Les robotiques gèrent de façon automatique un ensemble de bandes magnétiques (ci-dessus, une robotique vue de l'extérieur et, ci-contre, à gauche, vue de l'intérieur). Deux copies du document y sont enregistrées. Lorsqu'une archive est demandée, un bras robot préleve la bande où elle est stockée et la place dans un dispositif de lecture.



Les baies de stockage renferment de multiples disques durs. Deux copies du document y sont stockées, dans des salles distinctes.



Le serveur de transfert contrôle les documents reçus, en vérifiant notamment la conformité des métadonnées et des formats de fichiers. Il peut aussi effectuer diverses transformations, telle la conversion d'un format non libre (par exemple *Word*) en un autre (tel *Open office*). Il enrichit ensuite les métadonnées par des paramètres d'archivage, tel un identifiant international unique (une succession de lettres et de chiffres caractéristique du document). Enfin, il transmet le document au serveur de stockage.

Ce dernier en réalise cinq copies. Deux d'entre elles sont enregistrées dans deux baies de stockage distinctes (des armoires comprenant un certain nombre de disques durs), situées dans deux salles distantes. Deux autres sont inscrites sur des bandes magnétiques gérées par des dispositifs automatisés nommés robothèques. Ces robothèques, de deux pétaoctets chacune, sont aussi dans deux salles différentes. Une dernière copie est réalisée à plus de 300 kilomètres du site, dans un autre centre informatique, pour garantir les données contre les risques de catastrophe naturelle majeure.

Enfin, le serveur d'accès reçoit les demandes de consultation des archives, vérifie les autorisations de leurs émetteurs et leur transmet ou non les documents. Il peut également effectuer diverses modifications, comme flouter des visages pour préserver l'anonymat des personnes lors de l'envoi d'une photographie.

Des métadonnées bien renseignées répondent aux problèmes de localisation et de référencement des documents soulevés au début de cet article. Qu'en est-il de l'obsolescence et de la détérioration des supports et des formats ? Deux options se présentent : conserver en état de marche tous les dispositifs matériels et logiciels permettant la lecture de l'archive, afin de les réutiliser plus tard ; effectuer une migration des supports et des formats à mesure que les techniques évoluent (là encore, les métadonnées sont essentielles). Le CINES a choisi la seconde option.

Ainsi, les documents sont transférés vers les types de supports les plus récents – par exemple, aucune donnée n'est conservée sur disquette aujourd'hui. En outre, la durée de vie des supports est prise en compte pour assurer un renouvellement adapté. Les documents stockés sur des disques durs sont ainsi transférés tous les cinq ans, et ceux stockés sur bande magnétique tous les dix ans.

Les migrations concernent aussi les formats (PDF, JPEG...). La plate-forme du CINES accepte une quinzaine de formats, couvrant les différents types de données. Le choix est fait en privilégiant ceux dont les caractéristiques sont publiées et disponibles, et dont on pense que les versions successives seront compatibles avec les documents anciens. Une procédure de migration se déclenche lorsqu'un format est obsolète. Elle consiste en général à adapter un document aux versions successives d'un même format, mais il arrive aussi qu'un format disparaisse (tel celui des fichiers créés par le logiciel de bureautique *Claris*, abandonné au début des années 2000) et qu'il faille en changer. Elle se déroule en étroite collaboration avec l'organisme versant des données archivées, qui valide la nouvelle forme du document.

ENVIRON 40 RISQUES SONT IDENTIFIÉS aujourd'hui, tels qu'un incendie ou un format de fichier qui se périmé.

La gestion des archives à long terme nécessite également une « démarche qualité » – un ensemble de bonnes pratiques techniques et organisationnelles. Le respect des normes (assurant notamment l'interopérabilité avec d'autres systèmes d'archivage ou de collecte de données), la vérification du bon encodage du document selon le format choisi, les copies multiples et le contrôle de leur état participent à cette démarche. Au niveau organisationnel, tous les processus mis en œuvre (par exemple, l'ensemble des actions à réaliser pour faire migrer un document d'un format à un autre) sont répertoriés et conservés par écrit. Cela permet de transmettre aux générations suivantes le savoir-faire et l'historique du service, mais aussi de mieux analyser et d'améliorer les procédures. En outre, les risques potentiels et les actions à mener pour y parer sont actualisés tous les six mois. Environ 40 risques sont identifiés, tels qu'un incendie ou un format de fichier qui se périmé.

Offrir un service d'archivage numérique ne se résume donc pas à acquérir des dispositifs de stockage et à y copier des données. La pérennisation de documents nécessite une gestion active et une surveillance régulière, aussi bien des logiciels que du matériel. ■

■ LES AUTEURS



Francis DAUMAS est directeur du CINES [Centre informatique national de l'enseignement supérieur], à Montpellier.



Marion MASSOL est responsable du Département d'archivage du CINES.

■ SUR LE WEB

Groupe PIN (pérennisation de l'information numérique) de l'association Aristote : <http://pin.association-aristote.fr>

■ BIBLIOGRAPHIE

Collectif, *Archivage numérique pérenne*, *La Gazette du CINES*, février 2013 (www.cines.fr).

J.-C. Hourcade et al., *Longévité de l'information numérique : Les données que nous voulons garder vont-elles s'effacer ?*, EDP Sciences, 2010.

Norme iso 14721, ou OAIS [*Open archival information system*, soit *Système ouvert d'archivage d'information*]. www.iso.org

L'information scientifique de référence maintenant sur tablette et smartphone !



Téléchargez gratuitement
l'application sur App Store et Google Play.
Le 1^{er} numéro de *Pour la Science* est offert !



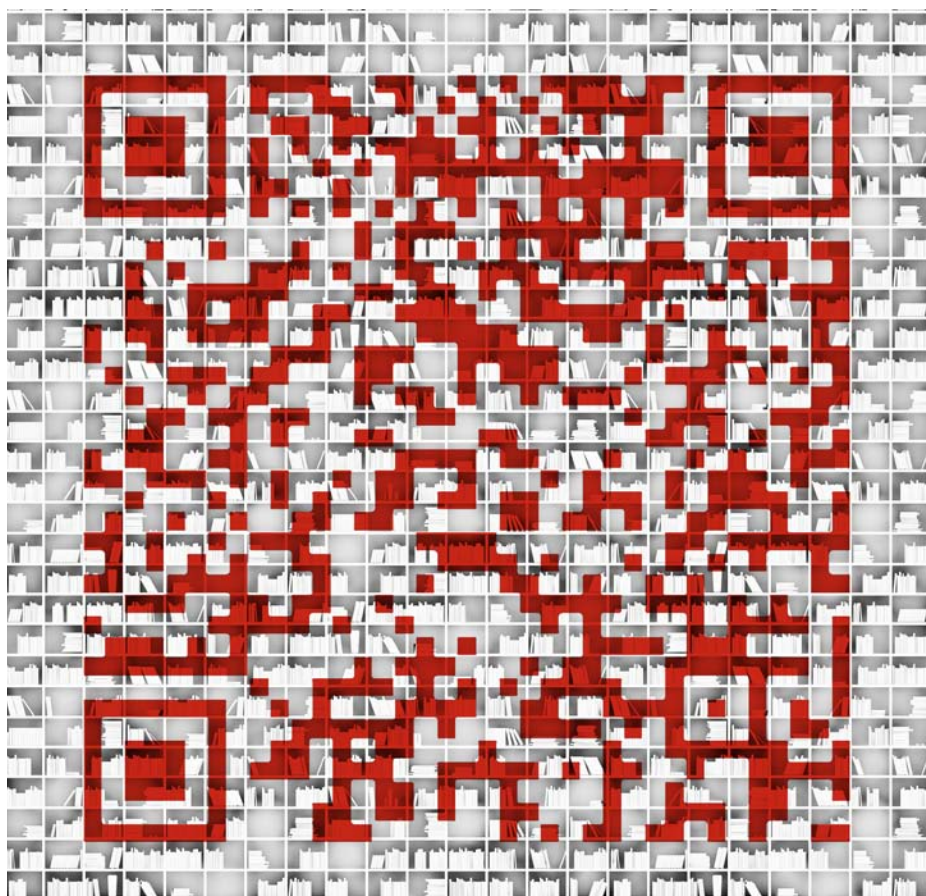
Avec l'application « Pour la Science », retrouvez dès leur sortie le mensuel *Pour la Science* à 5,49 € et son hors-série trimestriel *Dossier Pour la Science* à 5,99 € en version numérique optimisée pour tablette : lecture intuitive, sommaire interactif, contenus enrichis, feuilletage hors connexion, etc.



Flashez ce QR code avec votre mobile ou votre tablette pour télécharger immédiatement l'application.



Découvrez aussi l'application « Cerveau & Psycho » pour retrouver le magazine *Cerveau & Psycho* et son hors-série *L'Essentiel Cerveau & Psycho*



© Zastolskiy Victor / Shutterstock.com / Pour la Science

Le savoir numérisé

Nouvelles pratiques, nouveaux enjeux

Bibliothèques numériques, fonds numérisés, encyclopédies et dictionnaires en ligne, portails d'archives scientifiques en libre accès... En quelques années, Internet a permis un développement extraordinaire de l'accès au savoir, offrant à tout un chacun la possibilité de consulter, en quelques clics, des manuscrits de Descartes ou les dernières recherches en physique, ou encore de découvrir par le menu l'histoire du transistor.

Les bouleversements introduits par ces nouveaux modes de stockage et de diffusion de la connaissance sont cependant bien plus profonds. Qu'il s'agisse d'ouvrages numériques natifs, de sites Web ou de documents numérisés, les œuvres et les connaissances sous forme numérique s'accumulent. Comment les préserver, alors qu'elles sont stockées sur des supports dont la pérennité ne dépasse pas quelques dizaines d'années ? Et comment concilier ces nouvelles pratiques avec le droit d'auteur ? Les bibliothèques sont au cœur de ces réflexions (voir La numérisation des biblio-

thèques, page 48). Le modèle économique de diffusion du savoir est aussi ébranlé par le succès, d'une part, de l'encyclopédie en ligne gratuite *Wikipédia* et, d'autre part, du libre accès aux publications scientifiques : si certains éditeurs d'encyclopédies semblent tirer leur épingle du jeu (voir Dictionnaires & encyclopédies à l'ère numérique, page 52), la situation est tendue entre les chercheurs partisans du libre accès et les éditeurs de revues scientifiques (voir La science accessible à tous, page 54).

Mais le changement le plus drastique concerne les contenus eux-mêmes. Quelle fiabilité accorder aux entrées des encyclopédies et dictionnaires en ligne, qui ne sont pas estampillés par une autorité intellectuelle bien définie, ou aux articles scientifiques proposés en accès libre par leurs auteurs avant publication ? Plus que jamais, il est indispensable d'aborder ses lectures avec un œil critique. L'accès libre au savoir est à ce prix.

Marie-Neige Cordonnier

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 numérisées conditionnera les usages qui pourront en être faits.

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

Dictionnaires & encyclopédies à l'ère numérique

Rémi Mathis

L'encyclopédie en ligne *Wikipédia* et le *Wiktionnaire* modifient en profondeur les définitions mêmes de l'encyclopédie et du dictionnaire.



Alors que le livre numérique se développe et se diffuse à une vitesse mesurée, les encyclopédies ont connu, sous l'effet du numérique, des bouleversements majeurs. Et pourraient en subir de nouveaux. L'intérêt d'un support permettant d'accompagner les textes d'un grand nombre d'images, voire de films et de sons, a vite été compris des éditeurs. Entre 1985 et 1995, ils proposent tour à tour des versions sur CD-ROM de leur encyclopédie. L'avantage est évident : plus besoin de manipuler de lourds volumes, ni de les stocker chez soi, possibilité de rechercher un terme ou un concept dans le contenu, ou de rebondir d'un article à l'autre au sein d'une navigation simple et efficace, sans compter l'extrême réduction du coût de fabrication de l'objet en lui-même.

Jusqu'alors, le marché de l'encyclopédie était bien balisé, avec des marques de référence dans chaque pays : *Britannica* en Angleterre, *Universalis* en France, *Brockhaus* en Allemagne. Le changement de support rebat les cartes en attirant sur le marché de nouveaux venus : dès 1993, le géant de l'informatique *Microsoft* propose sa propre

■ L'AUTEUR



Rémi MATHIS est conservateur au Département des estampes de la Bibliothèque nationale de France. Il est aussi président de *Wikimédia France*.

■ BIBLIOGRAPHIE

R. Mathis, *Wikipédia. Une somme originale de copies*, *Medium*, n° 32-33, pp. 356-377, 2012.

R. Mathis, *Wikipédia, espace fluide, espace à parcourir*, *Revue de la BNU*, n° 6, pp. 52-61, 2012.

R. Mathis, *Défiances et production. Les bibliothèques françaises et Wikipédia*, *Bulletin des bibliothèques de France*, vol. 1, pp. 10-13, 2011.

encyclopédie multimédia, *Encarta*, disponible en sept langues. L'entreprise se fonde sur des technologies développées pour d'autres usages (par exemple son logiciel *MapPoint* pour les cartes) afin de rendre la consultation plus efficace et agréable. Elle change aussi complètement le modèle intellectuel de l'encyclopédie.

Une encyclopédie était alors un produit éditorial de luxe, au coût élevé, dont les articles étaient rédigés par les meilleurs spécialistes de la discipline. *Encarta* a une approche différente : *Microsoft* achète des droits non exclusifs de diffusion à des encyclopédies existantes (*Funk&Wagnalls*) et rachète d'autres encyclopédies (*Collier*, *New Merit Scholar*). Le contenu est traduit, mais varie selon les pays, car des partenariats spéciaux sont signés : la version en néerlandais reprend, par exemple, des contenus de l'encyclopédie locale *Winkler Prins*.

Le résultat est un corpus d'environ 40 000 articles (plus ou moins selon les langues), mélanges de textes divers qui ne sont pas signés et ne font pas référence à des sources externes. Le texte est d'abord



Plusieurs centaines de versions linguistiques se développent (285 en 2013), dans des langues qui n'avaient jamais connu d'encyclopédie auparavant, telles que le wolof ou des langues indiennes parlées par plusieurs dizaines de millions de personnes.

Le modèle étonne et est parfois remis en cause : des erreurs grossières sont relevées par des journalistes et des polémiques naissent sur la fiabilité de *Wikipédia*. Surtout, *Wikipédia* place son lecteur dans une situation nouvelle. Il ne lui est plus demandé de croire ce qui est écrit en partant du principe qu'un spécialiste a peu de chance de faire des erreurs, mais il lui est fourni un texte, de qualité variable, en transformation perpétuelle. Au lecteur d'adapter sa lecture à ce qu'il trouve et de faire preuve d'esprit critique. Autre différence : *Wikipédia* met en valeur la « neutralité de point de vue ». Ses contributeurs ne prennent pas parti sur des polémiques ou des débats ; ils se contentent de rapporter (sources à l'appui) les différents points de vue jugés pertinents par la communauté des spécialistes.

En 2009, face au succès public de *Wikipédia*, Microsoft ferme *Encarta*, et les édi-

numérique aura servi de catalyseur pour réinventer un modèle et se tourner vers les besoins des lecteurs.

La définition même des publications évolue. Une encyclopédie traitait autrefois du savoir légitime, à dominante littéraire et historique. Les possibilités offertes par le numérique – telle l'absence de contrainte de place – a ouvert l'encyclopédie à une autre catégorie de savoir, technique, scientifique ou populaire. Certains y retrouvent l'esprit de l'*Encyclopédie* de Diderot et d'Alembert, dont les auteurs allaient jusque dans les ateliers des artisans. D'autres, tenants de la tradition, se désolent de voir s'effacer le caractère légitimant de l'encyclopédie – et donc son rôle de guide et de juge.

Quoi qu'il en soit, les canons sont délaissés et le relativisme gagne du terrain : tout est bon à savoir, pensent de plus en plus de lecteurs, et ce n'est pas à une entreprise privée – fit-elle travailler des experts – de nous expliquer ce qui doit être su et ce qui est sans intérêt. Les dictionnaires subissent la même évolution. Si *Larousse* et le *Robert* continuent à décider quels mots peuvent être utilisés, le *Wiktionnaire*, cousin de *Wikipédia*, référence tous les mots, du moment qu'ils sont attestés dans des publications d'auteurs. La légitimité d'une instance chargée de faire le tri dans l'usage des mots s'affaiblit.

Le paysage évolue à une rapidité telle qu'il est difficile d'anticiper les dix prochaines années. L'accès mobile vient encore rebattre les cartes, et les modèles éditoriaux et économiques sont mal assurés. *Wikipédia* est de plus en plus au centre d'Internet par les données sémantiques qu'elle propose... mais peut s'effondrer en quelques années si les contributeurs se désintéressent de la rédaction et de la maintenance du projet. L'avenir dépendra des relations entre les industries d'Internet, les lecteurs – dont les pratiques évoluent sans cesse – et les éditeurs des outils de diffusion du savoir. Si *Wikipédia* révèle que le lecteur est encore à la recherche d'une connaissance organisée et centralisée, le développement du Web sémantique – un mouvement qui vise à faire émerger de nouvelles connaissances de celles déjà présentes sur Internet – et son utilisation dans les moteurs de recherche pourraient faire disparaître la notion d'encyclopédie ou de dictionnaire en tant qu'œuvre en soi. Autant dire que des surprises peuvent vite arriver. ■

LE WEB SÉMANTIQUE POURRAIT faire disparaître la notion d'encyclopédie ou de dictionnaire en tant qu'œuvre en soi.

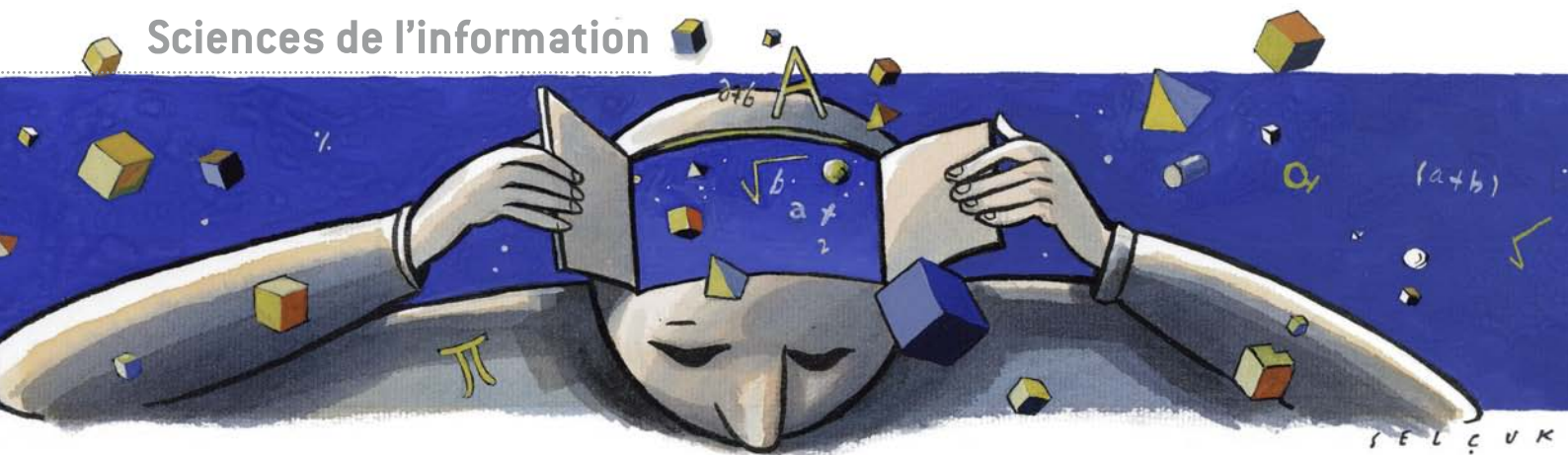
diffusé sur CD-ROM, puis sur Internet. Ce nouvel acteur, qui distribue son encyclopédie selon des canaux nouveaux, est un concurrent redoutable pour les encyclopédies traditionnelles : le chiffre d'affaires de *Britannica* est divisé par deux entre 1990 et 1996.

Créée en 2001, l'encyclopédie en ligne *Wikipédia* prend vite une place préminente sur le marché. Le changement est radical, car son modèle se fonde sur des habitudes venues d'Internet et non de l'édition. L'idée est de bâtir une encyclopédie de façon collaborative selon les méthodes du logiciel libre : chacun peut apporter sa pierre à l'édifice selon des règles mises au point par la communauté. Il n'y a pas de rédacteur en chef, tous les contributeurs sont à égalité et résolvent les éventuels conflits par la discussion et la recherche de consensus.

La croissance de *Wikipédia* est exponentielle. Dès 2005, la version anglophone propose 500 000 articles – le double l'année suivante. La version francophone, elle, atteint le demi-million d'articles en 2007...

teurs s'adaptent au lectorat numérique. *Britannica* met un terme à la publication sur papier en mars 2012, *Universalis* en 2013. Ils mettent l'accent sur le fait que les articles sont écrits par des spécialistes, ce qui est censé être un gage de qualité. Surtout, ils cherchent de nouveaux modèles économiques compatibles avec l'existence de *Wikipédia* : un abonnement à *Universalis* ne vaut que 50 euros par an (contre 5 000 euros sur papier, 3 000 pour la dernière édition).

Chez *Britannica*, afin de souffrir la comparaison avec la capacité de *Wikipédia* à être à jour en temps réel, les articles sont repris et rédigés par des employés (non experts). Surtout, la vente de l'encyclopédie ne représente plus que 15 pour cent du chiffre d'affaires de la société, le reste étant constitué de services liés à la formation en ligne et au marché de l'éducation. Déficitaire dans les années 1990, *Britannica* fait des bénéfices depuis 2003. Le passage au



La science accessible à tous

Christine Berthaud et Agnès Magron

Bien que certaines revues scientifiques demeurent payantes, de plus en plus d'articles de recherche sont en accès libre sur Internet. Les enjeux de ce mouvement sont tant scientifiques qu'économiques.

Connaissiez-vous Jack Andraka ? Cet Américain de 15 ans a mis au point un test de dépistage du cancer du pancréas simple et peu coûteux. Après la mort d'un proche atteint de cette maladie, l'adolescent a inventé tout seul sa méthode révolutionnaire en faisant des recherches sur Internet et en lisant des études médicales libres d'accès. La performance est surprenante. Toutefois, certains ne s'étonneront pas, à l'heure du tout gratuit sur le Web, qu'un internaute ait pu compiler gratuitement de la littérature savante. Or cette situation a été obtenue de haute lutte : elle est le fruit d'un mouvement international – le libre accès – né au sein d'une communauté scientifique déterminée à se réapproprier la diffusion des résultats qu'elle produit.

L'édition scientifique représente un marché lucratif, notamment celui des sciences dites exactes (chimie, médecine, biologie, physique), évalué à 17,3 milliards d'euros en 2011. Publier est un enjeu majeur pour les chercheurs : c'est par ce moyen qu'ils font connaître leurs travaux, sont reconnus par les collègues de leur discipline, progressent dans leur carrière, obtiennent des

financements. Ils ne perçoivent en général pas de droits d'auteur pour leurs publications. Ce sont aussi les chercheurs qui expertisent les textes soumis aux éditeurs : ce processus d'évaluation, nommé revue par les pairs, est indissociable de la publication scientifique. Il permet de valider les résultats présentés dans le texte, en particulier d'un point de vue méthodologique. Ce travail de relecture n'est pas rétribué par les éditeurs : être considéré comme « expert » constitue en quelque sorte une rémunération symbolique. Le support de diffusion le plus utilisé est la revue, même si, en sciences humaines et sociales, les livres restent importants.

L'édition scientifique : un marché contrôlé

La plupart des lecteurs sont des chercheurs, mais ce sont en majorité (68 à 75 pour cent) les bibliothèques universitaires ou de recherche qui acquièrent les ouvrages et financent les abonnements aux revues. Les prix sont si prohibitifs (en moyenne 4450 dollars l'abonnement annuel à une revue en chimie ou 2322 dollars en astronomie) que seuls des

budgets institutionnels peuvent les supporter. Et les prix ne cessent d'augmenter, surtout depuis l'arrivée du numérique.

Les éditeurs ont justifié un temps ces augmentations par les investissements qu'ils devaient faire pour s'adapter et numériser des archives (qu'ils revendent après d'ailleurs). Ils ont surtout affaire à une clientèle captive : une bibliothèque se doit de proposer à ses usagers les revues incontournables de leurs disciplines. Le numérique a renforcé cette dimension : les revues sont proposées dans des bouquets composés par des éditeurs, et les bibliothèques doivent négocier des licences d'accès pour leurs utilisateurs.

Qui vend ? Le marché de l'édition scientifique est dominé par des entreprises privées, certaines cotées en Bourse. Quatre grandes sociétés contrôlent le marché : *Reed-Elsevier* (qui affiche un portefeuille de 2000 titres de revues), *Springer-Verlag*, *Blackwell-Riley* et *Thompson Reuters*. Cette situation de monopole cristallise les critiques de la communauté scientifique : elle y voit un détournement par des opérateurs privés de fonds publics finançant la recherche aussi bien en amont (salaires des chercheurs) qu'en aval

(subventions des bibliothèques). De plus, la pression financière exercée sur les bibliothèques menace la diffusion des résultats.

Le passage au numérique a permis de s'affranchir des contraintes de la distribution papier. L'appropriation forte et rapide de ce nouveau média par certaines communautés de chercheurs a favorisé l'émergence de réalisations qui ont bouleversé l'accès à la littérature scientifique.

De ArXiv au libre accès

Les physiciens ayant besoin de partager vite les résultats de leurs recherches, l'un d'eux, Paul Ginsparg, a créé en 1991, au Laboratoire américain Los Alamos de l'Université de Californie, aux États-Unis, le premier serveur d'archives ouvertes, *hep-th* (*High Energy Physics - Theory*), aujourd'hui connu sous le nom d'*ArXiv* (prononcé *arkaïv*). Ce serveur, sur lequel les chercheurs déposent leurs articles (publiés ou non) sous la forme de fichiers, a servi de modèle à d'autres projets répondant à la définition de ce que, un peu plus tard, on a nommé l'*Open Access*, le libre accès aux publications scientifiques.

Ce nouveau concept a été formalisé en 2002 par une déclaration, la *Budapest Open Access Initiative* (BOAI), qui définit deux stratégies pour mettre en accès libre les publications des chercheurs : la voie verte, c'est-à-dire la mise en ligne des écrits sur le Web par leurs auteurs, et la voie dorée – la publication dans des revues électroniques en accès ouvert. Dès 2003, les grandes institutions de la recherche et de l'enseignement supérieur s'engagent à soutenir les politiques d'accès ouvert, notamment en incitant leurs chercheurs à déposer leurs publications sur des serveurs d'archives ouvertes au sein des bibliothèques universitaires et des institutions. En juillet 2012, la Commission européenne a demandé aux États membres de définir une politique intitulée « Pour un meilleur accès aux informations scientifiques : dynamiser les avantages des investissements publics dans le domaine de la recherche ».

En France, la voie dorée est représentée notamment par le Centre pour l'édition électronique ouverte (CLEO), qui gère Open-edition, un portail de revues en sciences humaines et sociales. Pour la voie verte, le CNRS a créé en 2000 le Centre pour la communication scientifique directe (CCSD), dont la principale mission est la gestion de l'archive ouverte HAL (Hyper articles en

HAL en chiffres



2 800 nouveaux fichiers
y sont déposés par mois.

236 000 fichiers
y étaient hébergés
en septembre 2013.

1,5 téraoctet
de données y est stocké.

286 000 auteurs y ont
déposé au moins un article.

2 500 collections
(institutionnelles,
thématiques...)
y ont été créées.

■ À ÉCOUTER

Le jeudi 31 octobre de 14h à 15h, Christine Berthaud reviendra sur les enjeux du libre accès dans la partie Actualités de l'émission **La marche des sciences**, sur *France Culture*.
www.franceculture.fr

■ LES AUTEURS



Christine BERTHAUD est directrice du Centre pour la communication scientifique directe (CCSD, CNRS-UPS2275), à Villeurbanne.

Agnès MAGRON est gestionnaire de communauté au sein du CCSD.

■ SUR LE WEB

ArXiv, archive ouverte en physique et en mathématiques : arxiv.org

HAL, archive ouverte couvrant tous les domaines scientifiques : hal.archives-ouvertes.fr

Le CCSD, concepteur de HAL et chargé de sa gestion : ccsd.cnrs.fr

Le Centre pour l'édition électronique ouverte : cleo.openedition.org

ligne) interconnectée aux grands serveurs internationaux tel *Pubmed Central*, de l'Institut américain de la santé, proposant de nombreuses archives en sciences de la vie.

Contrairement à *ArXiv* dont elle s'est pourtant inspirée, HAL rassemble des publications issues de tous les domaines scientifiques, déposées par les auteurs eux-mêmes dans plus de 60 pour cent des cas ou par des professionnels de la documentation. Ces dépôts sont faits dans le respect du droit de chaque acteur de la publication. Cette plate-forme, dont l'illustration page ci-contre est l'emblème, est utilisée par l'ensemble des établissements français, universitaires ou de recherche. Faisant partie de la mise en œuvre de la politique demandée par la Commission européenne, elle est aussi soutenue par le ministère de l'Enseignement supérieur et de la Recherche.

Ce projet permet de prendre en compte des aspects techniques lourds, telle la possibilité de garantir l'accessibilité aux textes par des liens URL solides et d'apporter des solutions au problème de la pérennité à long terme des données par des mises à jour régulières des fichiers déposés.

Les éditeurs considèrent le libre accès comme une menace pour leur modèle économique. Ils tentent de différer le dépôt dans une archive ouverte en imposant aux auteurs un embargo de six mois à deux ans après la date de publication, mais leur proposent aussi de payer, avant publication, pour que leur article soit en libre accès. Ce modèle est dit hybride, car si les revues sont disponibles sur abonnement, le téléchargement de certains articles est gratuit. Les revues en libre accès bousculent les modèles économiques. Certaines reportent vers l'auteur les coûts de production de la revue. Le financement n'intervient donc plus en aval de la publication sous forme d'abonnements, mais en amont grâce au paiement par l'auteur d'une somme forfaitaire pour chaque article accepté. En général, cette somme est subventionnée et prise en charge par son institution ou par l'organisme de financement de la recherche.

Aujourd'hui, les enjeux du libre accès et de la pérennité se généralisent aux données produites au cours de la recherche, tels les résultats d'expériences ou les données statistiques. Cela concourt à la validation des résultats en facilitant leur reproduction par d'autres équipes, ainsi qu'au libre accès de l'ensemble des recherches financées par des fonds publics. ■

Orienter la société grâce aux données massives

Les traces numériques que chacun de nous laisse derrière lui chaque jour pourraient devenir un cauchemar pour notre vie privée... ou servir à construire un monde plus sain et plus prospère.

Alex Pentland

Dès le milieu du XIX^e siècle, la croissance urbaine rapide, encouragée par la révolution industrielle, avait engendré des problèmes sociaux et environnementaux majeurs. Les villes y ont répondu en construisant des systèmes centralisés pour fournir de l'eau, de la nourriture et de l'énergie, favoriser le commerce, faciliter les transports, maintenir l'ordre et donner un accès aux soins. Aujourd'hui, ces solutions plus que centenaires sont de moins en moins adaptées. Nos grandes villes sont surpeuplées et encombrées par les bouchons. Et nous sommes confrontés à de nouveaux défis, en particulier pour nourrir et loger une population mondiale qui devrait croître de deux milliards d'habitants, tout en limitant au mieux l'impact du réchauffement climatique.

Ces problèmes typiques du XXI^e siècle appellent une réflexion nouvelle. Mais de nombreux économistes et sociologues abordent encore les questions sociales avec des concepts d'un autre âge, tels que les marchés et les classes, dans des modèles simplifiés qui réduisent les interactions sociales à des règles générales et ignorent les comportements individuels. Il faut

examiner la situation plus en profondeur et prendre en compte les détails les plus fins des interactions sociales.

Les technologies numériques nous le permettent aujourd'hui. Elles offrent une vue unique sur les milliards d'interactions individuelles par lesquelles les gens échangent des idées, de l'argent, des biens ou des informations. Le Laboratoire de dynamique humaine que je dirige au MIT (l'Institut de technologie du Massachusetts) explore cette masse de données pour tenter de faire émerger les lois qui régissent ces interactions. Par cette approche, nous commençons à pouvoir prédire certains phénomènes tels que les chutes des cours de la Bourse, les résultats d'élections politiques ou les épidémies de grippe. L'analyse des données sociales peut nous aider à bâtir des systèmes financiers stables, des institutions politiques qui fonctionnent, des systèmes de soins efficaces et économiques, et bien d'autres choses. Mais nous devons d'abord prendre la pleine mesure du poten-

1. LES TRACES NUMÉRIQUES que nous laissons derrière nous permettent de retracer et de prédire nos comportements individuels.

