

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 !



Disponible sur
App Store

DISPONIBLE SUR
Google play

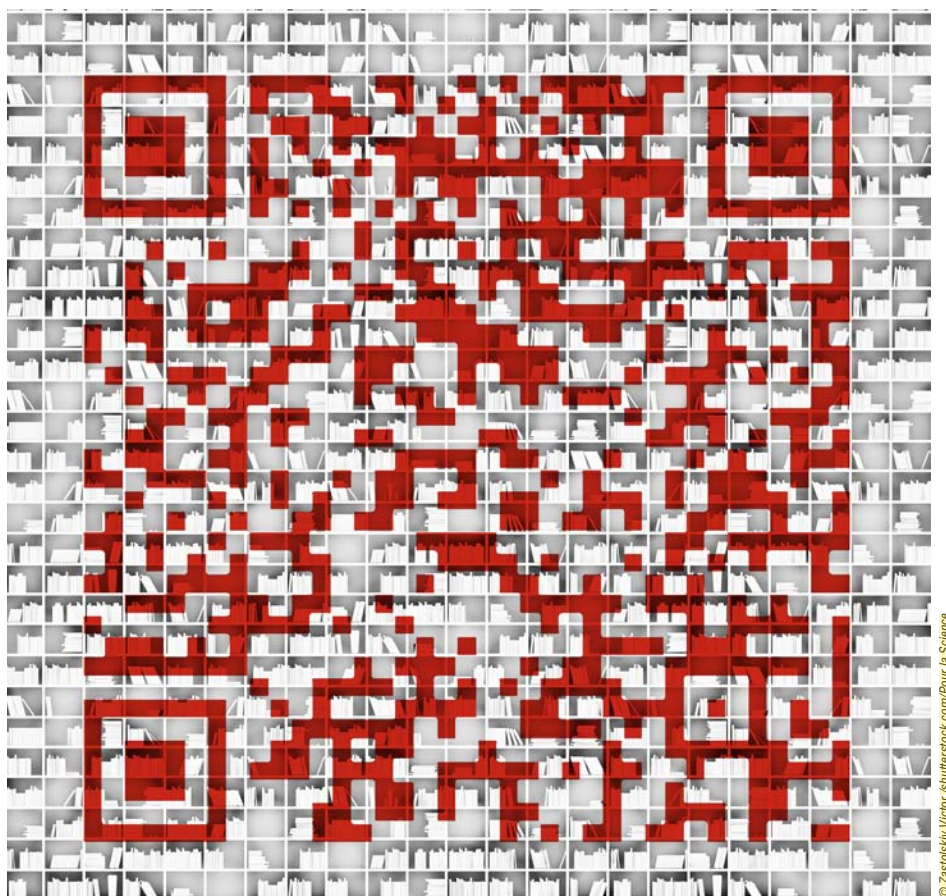


Flashez ce QR code
avec votre mobile
ou votre tablette
pour télécharger
immédiatement
l'application.

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.



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



© Zastolisky 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