



de s'élargir, comme l'illustre l'arrivée des scanners et des imprimantes 3D.

Dès lors, on est tenté de penser que le problème ancien de la préservation des données est résolu... Après tout, quand il est possible de dupliquer, sans perte de qualité, et à coût quasi nul, autant de fois qu'on le souhaite, comment pourrait-on encore perdre de l'information ?

... n'existe pas !

Mais à y regarder de plus près, on s'aperçoit vite qu'il n'en est rien. Non seulement les techniques actuelles de reproduction n'ont pas résolu le problème, mais elles l'ont en fait compliqué. En premier lieu, les mêmes outils qui rendent facile la duplication de l'information ont aussi pour conséquence d'engendrer dans la pratique une quantité gigantesque d'information : pour l'année 2011, on a calculé que 1,8 zettaoctet de données ont été produites (voir la figure 2), et huit zettaoctets sont annoncés pour 2015 ! Précisons que la plus grande partie de cette information n'est pas créée directement par l'homme, mais par nos machines, qui gardent non seulement

les traces numériques de nos opérations informatiques, telles la transmission ou la conversion de fichiers, mais aussi les données et les métadonnées (données sur les données) produites par nos échanges sur les réseaux sociaux...

Décider de ce qui doit être préservé dans cette gigantesque masse et de ce qui peut tomber dans l'oubli n'est pas aisé. Ainsi, la Bibliothèque du Congrès des États-Unis archive désormais les messages envoyés par les citoyens américains sur le réseau *Twitter*, alors que ces commentaires et autres micro-informations devaient à l'origine n'être qu'éphémères...

Même si l'on disposait d'une méthode pour décider exactement de ce qui doit être préservé, la conservation à long terme des objets numériques serait bien plus problématique qu'il n'y paraît. À la différence des livres et des tableaux, qui livrent immédiatement leur contenu, l'accès aux objets numériques suppose un environnement technique complexe. Ce dernier comporte, d'une part, des supports matériels de stockage de l'information susceptibles de se dégrader et, d'autre part, des logiciels complexes, utilisant des formats définis et fonctionnant seulement sur certains processeurs qui, eux, sont susceptibles de devenir rapidement obsolètes. À moyen et long termes, les risques de pertes de données numériques sont donc considérables.

Illustrons cette réalité peu connue par un exemple : si l'on retrouvait aujourd'hui un document électronique rédigé il y a seulement... 30 ans, sans doute aurait-il la forme d'une disquette magnétique de 5 pouces 1/4 contenant un fichier *WordStar*. À supposer que cette grosse disquette ne soit pas trop dégradée, ce qui est fort peu probable (voir l'article de Franck Laloë et Erich Spitz dans ce numéro), nous nous heurterions à une autre difficulté : les lecteurs de tels supports ont presque tous disparu. De même, le logiciel *WordStar* n'est plus employé (car il n'est plus disponible), de sorte que le format dans lequel il enregistrait les données n'est plus reconnu par les traitements de texte actuels (voir la figure 4).

L'ESSENTIEL

- La réplication numérique de données et leur diffusion par Internet semblent faciliter la conservation de la mémoire humaine.
- En réalité, elle la menace, car les moyens techniques de cette réplication se périment et rendent les données inaccessibles.
- Les offres d'hébergement gratuit des grands opérateurs d'Internet, pour utiles qu'elles soient, ne constituent pas non plus une solution sans risques.
- Il reste à refaire sur le réseau ce que l'humanité a toujours fait : répartir sa mémoire entre tous !

© Shutterstock.com/festa

1. LE RÉSEAU HUMAIN est depuis toujours le dépositaire de l'expérience de l'humanité. Pour en faire aussi le dépositaire de sa mémoire numérique, encore faut-il faciliter les échanges de données entre pairs...

Le cas du projet *Domesday* de la BBC (société britannique de radiotélévision) fournit un exemple de ces difficultés techniques. En 1986, soit quelque 900 ans après une entreprise similaire (voir l'encadré page 90) ordonnée par Guillaume le Conquérant (1027-1087), la BBC a lancé un recensement et un sondage sur l'ensemble du Royaume-Uni. Les données, recueillies par plus d'un million de volontaires dans les 23 000 cellules territoriales définies pour l'occasion, constituaient une sorte d'immense instantané de la société britannique; elles furent mises à la disposition du public au moyen d'un vidéodisque de type *LaserDisc*, un support des données dont on considérait à l'époque qu'il serait durable. Or, dès 1998, il fut supplanté par les DVD, de sorte qu'à partir du début des années 2000, on ne trouvait déjà pratiquement plus de lecteurs de *LaserDisc*, ni d'ordinateurs où pouvait tourner le logiciel nécessaire pour présenter le contenu du disque. Résultat: plusieurs équipes durent fournir un important effort de recherche afin d'émuler (d'imiter) le fonctionnement des lecteurs de *LaserDisc* dans un environnement informatique actuel.

Ce cas montre que pour préserver des logiciels, il importe de maintenir les connaissances nécessaires pour les faire exécuter par une machine, et donc de disposer de leur code source et de leur environnement d'exécution. Dans le cas du *LaserDisc*, quelques années avaient suffi pour que tout cela disparaisse! C'est bien à cause de telles situations qu'aujourd'hui, on reconnaît de plus en plus que l'usage de formats de données ouverts (qui peuvent être librement utilisés et modifiés, et dont les spécifications

sont publiques et sans restriction d'accès) et de logiciels libres (dont le code source est disponible pour tous) est indispensable à la pérennité des informations.

Mais une fois le problème de la lecture des *LaserDisc* du projet *Domesday* résolu, on s'aperçut que d'épineux problèmes légaux restaient à démêler. Selon le droit d'auteur en vigueur, la réalisation d'une copie exigeait que l'on recueille d'abord les autorisations des millions d'individus ayant contribué au projet, ainsi que celles des auteurs des logiciels utilisés... À une époque où reproduire l'information est devenu banal, l'imposant arsenal juridique des droits d'auteurs nous interdit nombre de répliques, et ce sur des durées de plus en plus longues... Si le droit protégeant la propriété intellectuelle sert aujourd'hui essentiellement à préserver les modèles économiques ayant précédé l'ère d'Internet, il a souvent pour effet de bloquer le développement des modèles nécessaires à l'économie numérique.

Copier est-il hors la loi ?

Si nous voulons préserver les données, nous sommes donc confrontés à un dilemme: nous devons d'un côté les recopier sur de nouveaux supports (avant leur inévitable dégradation) et sous de nouveaux formats (avant leur inévitable obsolescence); d'un autre côté, nous devons respecter des lois qui nous interdisent de réaliser ces copies... Notre civilisation est donc confrontée à la question suivante: qui doit se charger du renouvellement continu des données et dans quelles conditions? Plusieurs



2. EN 2011, 1,8 ZETTAOCTET de données a été produit, et diverses estimations en annoncent huit pour 2015. Un zettaoctet représente 1 000 milliards de gigaoctets (c'est-à-dire 10^{21} octets). Ainsi, en quelques jours, l'humanité accumule aujourd'hui plus d'information qu'elle n'en a produit en deux millions d'années avant l'invention d'Internet.



3. LES SUPPORTS DE STOCKAGE se suivent et ne se ressemblent pas, tant en termes de capacité que de format d'enregistrement. Ci-dessus, quelques-uns des formats datant des 30 dernières années: à gauche, une disquette 5 pouces 1/4, au milieu une disquette 3 pouces 1/2 de haute densité, puis un CD-ROM et, à droite, une clef USB actuelle.



4. WORDSTAR est un logiciel de traitement de texte introduit en 1978, qui fonctionnait sous DOS, le système d'exploitation pour ordinateurs personnels de Microsoft. Il occupa une position dominante au début des années 1980, mais fut supplanté par *Word* dans les années 1990, puis abandonné et oublié, au point de devenir inutilisable aujourd'hui.