Centraliser les données pour les mettre à disposition, c'est courir le risque de les voir toutes disparaître d'un seul coup. Plus que jamais, l'avènement de la copie numérique invite au contraire à les partager au maximum afin d'augmenter les chances de leur pérennité.

Roberto di Cosmo

Organiser le partage pour préserver les données

L'un des traits humains est de tendre à transmettre son expérience. Qu'il s'agisse d'histoire, de philosophie, d'avancées scientifiques, d'œuvre littéraires ou artistiques, ou tout simplement de vécu, l'humanité a toujours cherché à préserver ses savoirs et à les transmettre aux générations futures. Pour cela, elle a d'abord employé les récits oraux et les mythes, puis l'invention de l'écriture a entraîné celle des livres, et tout a changé le jour où ces « recueils » ont pu être multipliés efficacement grâce à l'imprimerie, puis partagés en masse dans des bibliothèques. Aujourd'hui, ce progrès majeur accompli pendant la Renaissance est démultiplié de façon explosive par la copie numérique et l'internet. Dans ce qui suit, nous allons analyser les modalités, les risques et les promesses de la transmission numérique, cette nouvelle forme de transmission de l'expérience humaine qui, une nouvelle fois, change tout !

Dupliquer des informations a longtemps été difficile. Pensons par exemple au coût prohibitif de la recopie d'un livre par un copiste avant l'invention de l'imprimerie. La seule stratégie possible consistait à réunir les précieux recueils en un endroit à la fois sûr et accessible, où l'on pouvait les protéger tout en les rendant disponibles à la consultation à l'aide d'index centralisés. La Grande bibliothèque d'Alexandrie, où, pendant l'époque hellénistique (336-30), on tenta de rassembler « tout le savoir du monde », illustre l'ancienneté de cette stratégie ; sa destruction à une époque indéterminée montre le risque bien connu par les informaticiens que courent les systèmes centralisés, puisque la destruction de leur centre signifie leur disparition…

La copie parfaite…

Grâce aux avancées techniques du dernier demi-siècle, nous vivons aujourd'hui une nouvelle rupture dans l'histoire culturelle de l'Humanité. Pour la première fois, nous disposons d'instruments pour répliquer parfaitement un nombre toujours plus grand de biens immatériels. Les moyens informatiques rendent en effet possible de numériser – c'est-à-dire d'encoder sous la forme d'une suite de chiffres binaires – du texte, du son, des images et de la vidéo, puis de répliquer à volonté ces informations sans perte de qualité et pour un coût

négligeable. Tout cela est devenu si banal que nous l'accomplissons plusieurs fois par jour sans même y penser ; et le domaine des biens duplicables à l'identique ne cesse de s'élargir, comme l'illustre l'arrivée des scanners et imprimantes 3D.

Dès lors, on peut être tenté de penser que le problème ancien de la préservation des données a été 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 ?

… oblige à recopier

Lorsqu'on y regarde de plus près, on s'aperçoit vite qu'il n'en est rien. Non seulement les techniques actuelles de recopie n'ont pas réglé le problème, mais elles le rendent en fait plus difficile à résoudre. 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 d'information dépassant l'imagination. Un chiffre en donne la notion : pour l'année 2011, on a calculé que 1,8 zettaoctets de données ont été produites *(voir la figure 2)*, et on nous en annonce 8 zettaoctets pour 2015 ! Précisons que la plus grande partie de cette information n'est pas créée directement par nous, mais par nos machines, qui gardent non seulement les traces numériques de nos actions informatiques, telles la transmission ou la conversion de fichiers, mais conservent 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'a rien d'aisé. Ainsi, la Librairie du congrès des États-Unis archive désormais les messages envoyés par les citoyens américains sur le réseau *Twitter*, alors que ces commentaires plus ou moins fins 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 est bien plus problématique qu'elle 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, comportant d'une part des supports de stockage de l'information sur des supports matériels susceptibles de se dégrader, et d'autre part des logiciels compliqués, utilisant des formats définis et tournant seulement sur certains processeurs, qui, eux, sont susceptibles d'être vite obsolètes. À moyen et à long terme, les risques de pertes de données numériques sont donc tout à fait considérables.

Illustrons cette réalité mal 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 1/2 contenant un fichier *WordStar*. À supposer que cette grosse disquette soit toujours lisible, ce qui est peu probable (voir l'article de Franck Laloë et coll. dans ce numéro), nous nous heurterions ensuite au problème de la disparition quasi totale de lecteurs de tels supports. De même, le logiciel *WordStar* n'est plus employé et n'est plus disponible lui aussi, de sorte

L'ESSENTIEL

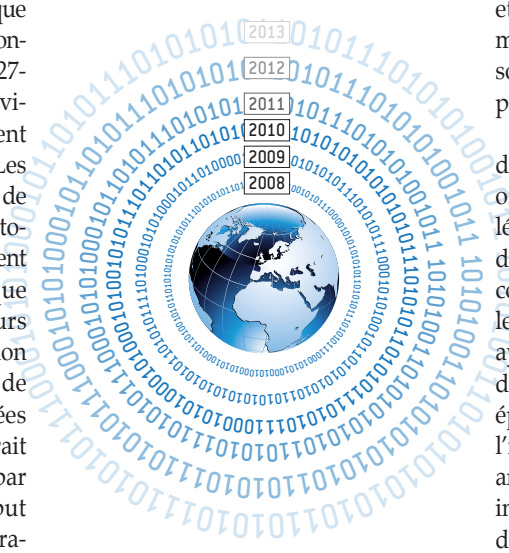
- La réplication numérique de données et leur diffusion par l'internet semblent faciliter la conservation de la mémoire humaine.
- En réalité, elle la risque car les moyens techniques de cette réplication se périment, et rendent les données inaccessibles.
- Les offres d'hébergement gratuit des grands opérateurs de l'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 !

1. LA MÉMOIRE DE L'HUMANITÉ est depuis toujours répartie entre ses membres ; il semble que dans le monde numérique aussi, cette structure soit la plus prometteuse pour conserver des données à long terme.

que le format dans lequel il enregistrait les données n'est plus reconnu par les traitements de texte actuels *(voir la figure 3)*.

Le cas du projet *Domesday* de la BBC fournit un exemple concret de ce genre de difficultés techniques. En 1986, soit quelque 900 ans après une entreprise similaire ordonnée par Guillaume le Conquérant (1027-1087), la société nationale de radiotélévision britannique a lancé un recensement et sondage général 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 cliché numérique des nationaux britanniques et de leurs pensées ; 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 (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 ce genre de situations, qu'aujourd'hui, on reconnaît de plus en plus largement que l'usage de formats de données ouvertes (dont les spécifications sont publiques et sans restriction d'accès et qui peuvent être librement utilisés et modifiés) 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 l'acte technique de reproduire l'information est devenu banal, l'imposant arsenal juridique des droits d'auteurs nous interdit nombre de réplications pour des durées devenant de plus en plus longues… Si le droit protégeant la propriété intellectuelle sert aujourd'hui essentiellement à protéger les modèles économiques prédatant l'internet, il a souvent pour effet de bloquer le développement des modèles nécessaires à l'économie numérique…

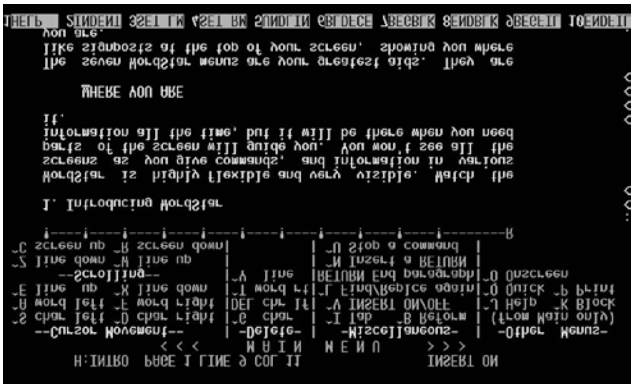


2. EN 2011, **1,8 ZETTAOCTET** et diverses estimations en annoncent 8 pour 2015. Un zettaoctet ? Cela représente un milliard de milliards de gigaoctets… Ainsi, en quelques jours, l'humanité accumule aujourd'hui plus d'information qu'elle n'en a produite en deux millions d'années avant l'invention de l'internet. Mais jusqu'où irons-nous?

Copier, 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

3. **LES SUPPORTS DE DONNÉES** se suivent et ne se ressemblent pas, tant en terme de capacité que de format d'enregistrement ! Ci-dessus, quelques-uns des formats des 30 dernières années : à gauche, une disquette 5"1/2, au milieu une disquette 3"1/2 à haute densité, ensuite un CD-ROM, et finalement à droite une clé USB actuelle.



4. **WORDSTAR** est un 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 d'en devenir inutilisable aujourd'hui.

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 de quelles données et dans quelles conditions ? Plusieurs réponses sont possibles. Chacune a ses avantages et ses inconvénients.

Une première possibilité serait de déléguer la préservation des données à quelques grands opérateurs chargés de les copier régulièrement sur les nouveaux supports avant que les anciens ne se dégradent, et, éventuellement, de les convertir dans les nouveaux formats. À cet égard, les offres d'hébergement gratuit ne manquent pas, qu'il s'agisse d'abriter nos photographies, nos vidéos, nos documents bureautiques ou encore notre courrier. Le plus souvent, elles sont le fait de grands fournisseurs de services sur l'internet, tels *Google, Amazon, Facebook, Apple, Microsoft, etc.* Toutefois, les récentes révélations sur le programme de surveillance PRISM de l'Agence américaine nationale de sécurité soulignent un phénomène général : nous payons ces services en fournissant involontairement en échange des informations très détaillées sur nos activités, nos préférences, nos liens personnels, nos courses, nos voyages, etc., lesquelles sont souvent employées à notre insu, ce qui se traduit non seulement par une surveillance occulte, mais aussi de plus en plus souvent par des violations à grande échelle de notre vie privée, par exemple par des marchands de toutes sortes.

D'autres grands acteurs ont lancé un gigantesque effort de préservation des données disponibles sur internet ou sur certains réseaux sociaux : fondée en 1996, l'*Internet Archive* (archives de l'internet) fait régulièrement des copies des sites disponibles sur la toile et conserve une grande collection d'images, de vidéos, de livres et de logiciels. S'agissant des œuvres littéraires, musicales ou cinématographiques, les grandes sociétés qui en détiennent les droits se réservent le droit exclusif de les archiver, et interdisent au public de les copier ou de les échanger ; en outre, elles mettent parfois en place des mesures de protection contre la copie qui entravent sérieusement la conservation de ces données à long terme.

Aussi, tous ces acteurs, aussi puissants soient-ils, finissent par constituer des points uniques de défaillances comparables à la grande bibliothèque d'Alexandrie des temps hellénistiques. Malgré les investissements faits pour créer des centrales de serveurs plus sûres et mieux organisées, la concentration de grandes masses de données sous le contrôle d'une seule organisation est toujours source d'une certaine fragilité. Et au-delà du problème technique de la sauvegarde des données contre les pertes accidentelles, il s'agit aussi de trier entre les données qui méritent d'être préservées, et celles qui peuvent être effacées. Or le fait qu'une seule organisation, ayant sa logique et ses intérêts particuliers, en décide, introduit manifestement un énorme risque de destruction. Aurions-nous aujourd'hui des éditions complètes initiales des *Fleurs du mal* de Charles Baudelaire (1821-1867), ou de *Madame Bovary* de Gustave Flaubert (1821-1880), si, il y a plus d'un siècle, alors que ces chefs-d'œuvre étaient censurés, leur conservation avait été confiée à une organisation étatique centralisée ?

5. **L'AGENCE NATIONALE DE SÉCURITÉ AMÉRICAINE** (la NSA ou *National security agency*) exploite la concentration des données personnelles chez les grands fournisseurs de services sur internet pour surveiller la population. Dans le projet PRISM, révélé par le lanceur d'alerte Edward Snowden, elle se fonde sur la loi américaine régulant les activités d'espionnage pour se faire fournir en données en masse à décrypter chez les gros opérateurs, tels *Google*, *Amazon*, *Facebook*, etc.

■ **L'AUTEUR**

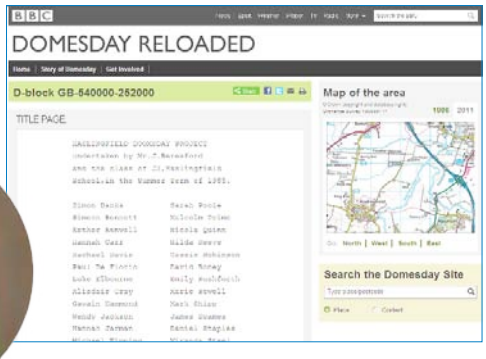
Roberto DI COSMO, professeur d'informatique de l'Université Paris VII, travaille dans le Laboratoire Preuves, Programmes et Systèmes de la même université et du CNRS.

L'Échange entre pairs

Une solution alternative de conservation des données de l'humanité à long terme provient du très grand nombre des ordinateurs personnels et de la croissance exponentielle de la capacité de stockage des mémoires disponibles pour le grand public. Le déploiement d'un système d'échange pair-à-pair rend envisageable la transformation du parc mondial d'ordinateurs personnels en une gigantesque base de données distribuées, puis de laisser les gens y répliquer

LES DONNÉES DU JUGEMENT DERNIER

En 1085, Guillaume le conquérant ordonna un recensement général en Angleterre, qui devint ensuite la base indiscutable des taxes royales, de sorte que les Anglais nommèrent *Domesday book* (livre du jugement dernier) le volume où les résultats étaient consignés *(à gauche)*. En 1986, la BBC décide de stocker les données du nouveau *Domesday* – un nouveau recensement et sondage général organisé par elle – sur des vidéodisques de type *LaserDisk*, support alors estimé pérenne *(au milieu)*. Toutefois, il se périma si vite que, 25 ans plus tard, il était déjà devenu impossible d'en lire le contenu ! Un long et coûteux effort technique fut alors nécessaire pour imiter sur un ordinateur moderne le fonctionnement d'un lecteur de *LaserDisk*, ce qui conduisit au *Domesday reloaded* (Domesday refait), une version rafraîchie consultable *(à droite)*, qui, par prudence, a été versée aux Archives nationales britanniques…



et y partager les données. Pourvu que les ordinateurs personnels continuent à exister (!), cette solution aurait l'avantage de faire l'impasse sur une autorité centrale fixant seule de ce qui mérite d'être conservé, bref, de permettre à l'humanité de décider par elle-même et spontanément des informations qu'elle va garder… Les années fastes du téléchargement de musique sur les anciens systèmes pair-à-pair nous ont donné un aperçu de ce qu'un tel réseau peut accomplir : à côté de tubes et autres succès commerciaux plus ou moins faciles, on y retrouvait souvent des morceaux de musiques rares et, sinon, indisponibles. Il suffit en effet qu'existe un petit groupe de passionnés par tel ou tel type de données pour que celles-ci soient préservées quelque part dans le réseau social. Ce trait typiquement humain constitue un atout en faveur de la conservation de données à long terme, étant donné qu'il est difficile d'identifier quelle information pertinente aujourd'hui le sera demain.

Toutefois, pour que l'humanité profite de cette façon d'elle-même, encore faut-il que l'accès aux données dans la durée soit amplifié par l'usage de formats ouverts et de logiciels libres ; et il serait aussi utile aussi de mettre en place à cet égard une infrastructure de recherche, répartie elle aussi comme les données partout dans le réseau afin d'indexer ces informations.

Le plus grand obstacle à la réalisation de cette vision n'est pas technique, mais économique. Nombre d'acteurs des industries de la musique, du cinéma et du livre militent pour qu'on restreigne les échanges entre pairs et pénalise ceux qui y participent ; fondés sur la vente de copies individuelles, leurs modèles économiques s'écroulent si l'on laisse advenir la possibilité de réaliser à volonté un nombre illimité de copies numériques de toute œuvre. C'est pourquoi, ils poussent à l'entrée en vigueur de lois controversées, contraignant les internautes à se détourner des systèmes pair-à-pair pour passer à des échanges cryptés, lesquels ne peuvent que restreindre l'accès public aux données.

Nous ne pouvons détailler ici les propositions qui ont été faites pour concilier ces intérêts contradictoires, mais il nous paraît essentiel de souligner ici que le libre partage sans entrave est un instrument essentiel pour la préservation des données pour les futures générations. Il doit être encouragé et accompagné. De la même façon que les nombreuses collections privées de livres ont préservé des titres dont les éditeurs mêmes ont disparu, les très nombreuses copies numériques individuelles de données nouvelles ou anciennes, enregistrées dans des formats ouverts avec des logiciels libres, seront pour nos successeurs un moyen de restitution de notre culture actuelle.

■ BIBLIOGRAPHIE

Jeff Rothenberg, **Avoiding Technological Quicksand : Finding a Viable Technical Foundation for Digital Preservation,** rapport au Council on Library and Information Resources, Janvier 1999, http://www.clir.org/pubs/reports/rothenberg.

Ph. et S. Aigrain, **Sharing : Culture and the Economy in the Internet Age**, Amsterdam University Press, 2012.

Thierry Priol, **Vers le tout-en-réseau ?,** Dossier PLS L'ère d'Internet, N°66, janvier-Mars 2010, http://www.pourlascience.fr/ewb_pages/f/fiche-article-vers-le-tout-en-reseau-24085.php

Sur le parc de PC prévisible à l'échelle mondiale : http://www.gartner.com/id=1602818