arbido

2016/3 Détruire pour conserver?

Gaudinat Arnaud, adjoint scientifique HEG Genève

Le plaisir de tout conserver sans modération: une question de taille?

Pourquoi jeter lorsqu'on peut conserver? Exit le papier physique et les mètres linéaires. L'ère du numérique a tout chamboulé. L'espace dans les nuages est infini, c'est la promesse de la loi Kryder qui prédit empiriquement le doublement de la densité de stockage tous les ans depuis 60 ans. Mais conserver l'information c'est bien, la retrouver c'est encore mieux et indispensable. Google trouve plutôt bien son chemin parmi plus de 1000 milliards de documents décentralisés. Alors pourquoi devrions-nous perdre du temps à trier, archiver, sélectionner, effacer nos centaines d'emails, de photos et autres documents? Ici sont présentés quelques idées, repères et exemples relatifs à la problématique de la conservation de toute l'information numérique plutôt que de son élimination.

Tous archivistes

Dans nos sociétés modernes, nous sommes tous des archivistes … Des archivistes plus ou moins efficaces. Nous sommes bien entendu tous confrontés à la gestion de nos nombreux documents administratifs. Et tout comme des archivistes chevronnés, nous nous devons de faire des classements et, de temps en temps, de décider de nous séparer de certains documents pour retrouver un peu d'espace dans nos rangements. Nous sommes encore des archivistes lorsque nous décidons de mettre nos photos numériques sur un support optique, de les étiqueter, d'en imprimer certaines ou carrément d'en effacer d'autres, car elles sont légèrement floues. Bien entendu, l'analogie s'arrête ici, car l'archivistique professionnelle s'occupera au sein d'une entreprise ou d'une institution, entre autres, du cycle de vie du document, mais aussi d'archiver les documents de façon pérenne sur du très long terme. Cependant, cette question de l'élimination qui touche l'archiviste professionnel est universelle dans notre monde moderne. Elle se doit d'être mise en perspective, par rapport à l'état de l'art du stockage et du traitement des données. À noter que l'élimination de données pour des aspects légaux, de confidentialité et de droit à l'oubli, ainsi que la conservation à long terme des documents numériques ne seront pas considérées ici. Cependant, en ce qui concerne la conservation à long terme, les principes de bases présentés dans cette revue semblent aussi valables (pour des coûts supérieurs). Et certaines solutions flexibles liées au cloud computing semblent aussi prometteuses [1], [2]pour traiter ce problème.

Taille de l'information, de quoi parle-t-on?

Dans le monde numérique, l'espace se compte en bits, la plus petite unité d'information. Avec un simple bit par exemple, on peut indiquer que l'on rend visible des millions de documents dans une interface ou pas, plutôt que de les effacer définitivement. L'octet, l'unité de base de l'informatique, qui représente 8 bits, permet lui de coder 256 informations. Mais que représente par exemple 1 To (téraoctet) de données, taille moyenne en 2016 des disques durs dans nos ordinateurs portables? Selon le Tableau 1, il est possible actuellement de stocker dans 1 To de 1 à 4 millions de livres, 250 DVD, approximativement 10% des ouvrages présents à la Bibliothèque nationale de France (BNF), ou 0,6% de Common Crawl. Et ainsi de conclure qu'un ordinateur actuel permettrait de stocker au format numérique la plupart de nos bibliothèques, sauf les plus grandes. Et qu'internet une fois de plus pose de nouveaux défis.

Tableau 1 : La taille de la donnée en 2016 et comparée à 1 To

Type de données	Approximation de la taille	Par rapport à 1 To
Tweet[3]	2.5 Ko	400 millions (moins d'un jour de tweets)

Livre	Entre 256 Ko et 1 Mo	1 à 4 millions de livres
DVD	4 Go	250 DVDs
Encyclopédie Web, Wikipedia en anglais, en 2016	50 Go	20 encyclopédies Web
11 millions d'ouvrages présents à la BNF en 2015	10 To	10% de la BNF
Capture vidéo pendant un an (selon Gurrin 2014)[4]	32.8 To	3>#/span###
Capacité du plus gros serveur de l'équipe fouille de données serveur à la HEG en 2013	100 To	1%
CommonCrawl[5] (moissonnage publique du Web en 2015)	168 To	0.6% de Common Crawl
Archive.org[6], site web, images, et vidéo en 2014 selon archive.org	50'000 To	0.002% de archive.org
Estimation du trafic Internet en 1 jour, en juillet 2016 selon internetlivestats[7]	1 556 000 To (1.5 Eo)	0.00006% du trafic mondial

Pour arriver à stocker autant de données dans nos ordinateurs, il s'est passé 60 ans d'évolution et de défis techniques[8] (voir Tableau 2) qui ont donné naissance à la loi de Kryder [9] (du nom de l'ingénieur chez Seagate qui identifie cette loi en 2005). Cette loi empirique (similaire à la fameuse loi de Moore) indique que la densité de stockage est multipliée par deux tous les ans pour un coût deux fois moindre. Et même si un certain ralentissement de la croissance de stockage dans les zones de stockage personnelles est constaté, la dernière colonne du tableau donne aussi la capacité qu'il est possible d'acheter pour CHF 100.–. Ainsi, si en 2015 il est possible d'obtenir 2 To pour CHF 100.–, pour le même prix en 1973 nous obtenions 33 Ko seulement.

Tableau 2: 60 ans d'évolution du stockage et capacité pour 100 CHF (source pcworld.com)

Année	1 Go	Capacité pour 100 CHF
1956	26 000 000.00 CHF	3.8 Ko
1973	3 000 000.00 CHF	33 Ko
1980	100'000.00 CHF	1 Mo
1987	40'000.00 CHF	2.5 Mo
1995	800.00 CHF	125 Mo
2002	2.00 CHF	50 Go
2007	0.28 CHF	357 Go
2015	0.05 CHF	2 To
2020 (selon loi de Kryder)		(estimation) 64 To
2025 (selon loi de Kryder)		(estimation) 2 Po

Pourquoi désherber ?

Si on reprend les six objectifs de la méthode de désherbage CREW[10], on trouve 1/ Gagner de l'espace, 2/ Gagner du temps, 3/ Rendre la collection plus attractive, 4/ Améliorer la réputation de la bibliothèque, 5/ Adapter la collection aux besoins, 6/ Avoir un retour sur les forces et faiblesses de la collection. L'espace et le temps semblent en effet essentiels. Cependant, éliminer les données permet aussi de diminuer le bruit lors de la recherche.

Ainsi, l'objectif numéro 1 du désherbage est de récupérer de l'espace. Qu'il soit physique ou numérique, l'espace est forcément fini (en tout cas avec nos connaissances scientifiques actuelles) et a de facto un coût proportionnel à sa taille. Si le coût de l'espace physique a tendance à croitre dans le temps, inversement le coût de l'espace numérique a tendance à diminuer et ce de manière importante (voir Tableau 2). C'est pourquoi on peut se poser la question sérieusement: dans le monde numérique, devrait-on tout conserver plutôt que de passer du temps (objectif 2) à sélectionner ce que l'on doit éliminer?

Le gain d'espace n'est certainement pas le seul objectif de l'élimination. Il est aussi d'éviter de se retrouver submergé par l'information lors de la recherche d'information. Car, comme mentionné précédemment, conserver l'information c'est bien, la retrouver c'est encore mieux et surtout indispensable. Google nous montre la voie en indexant plus de 1000 milliards de documents déjà en 2008. Bien entendu, nous ne sommes pas Google, mais la bonne nouvelle est que la plupart des solutions efficaces de traitement de données massives sont des logiciels « Open Source ». Par exemple, l'indexation des vingt-six millions de documents de Medline (citations de la littérature scientifique médicale) dans un logiciel tel que Elasticsearch [11] prend moins de dix heures sur un ordinateur portable datant de 2012.

Les objectifs 3 à 6 peuvent être réalisés aussi sans avoir besoin d'éliminer définitivement les documents, mais en les catégorisant comme tels pour les rendre moins visibles (mais encore accessibles) et aussi en les enrichissant automatiquement d'informations d'usages (objectif 6).

Conserver des données dans le but de les analyser : La sélection, une certaine forme d'élimination délétère ?

La sélection permet le gain de place, mais ce gain de place ne se fait-il pas au détriment de la qualité de l'information lorsque le but du stockage de données est de procéder à des analyses ? Cette question s'avère très importante lorsque l'on commence à vouloir faire parler les données éphémères. Prenons l'exemple récent du projet GEoTweet [12]. Ce projet avait comme objectif premier la cartographie des tweets géolocalisés dans le seul grand Genève pendant une période de sept mois (voir Figure 1). La collecte des tweets, limitée par l'API (Interface de programmation) offerte par Twitter, se faisait en définissant une fenêtre de capture rectangulaire. Dans GEoTweet, la fenêtre de capture utilisée pour couvrir le grand Genève alla arbitrairement de Culoz en France à Fribourg en Suisse. Le grand Genève s'arrête bien entendu avant cette zone, mais l'avantage de ce choix a permis d'étudier de façon pertinente les différences de tweets entre Genève et Lausanne, alors que ce n'était pas prévu au départ. À noter que dans ce cas, les informations en dehors du rectangle de capture sont actuellement perdues, à moins de disposer d'un budget conséquent pour utiliser les services de GNIP (entreprise permettant d'accéder à la totalité des tweets émise depuis les débuts de Twitter). Ainsi, une collecte de données plus large permettra de répondre à plus de questions, surtout si des questions additionnelles intéressantes apparaissent en cours d'analyse de données.

Le mouvement lifeloging et quantified-self

Pour certains, la question de tout conserver l'information ou non ne se pose déjà plus. Bien que singulier, le mouvement du lifelogging - le fait d'enregistrer sa vie de manière la plus continue possible - est déjà lancé et trouve de plus en plus d'adeptes grâce aux nouvelles technologies[4]. Son origine date de 1945 par Vannevar Bush et la vision Memex (une sorte de bureau qui capture l'activité de son utilisateur). Aujourd'hui, ce mouvement est représenté par Gordon Bell et son projet MyLifeBits[13]. Quant au quantified-self, il est déjà en partie démocratisé grâce aux capteurs de nos téléphones portables, aux balances connectées et surtout aux bracelets permettant entre autres de mesurer son activité physique, sa position GPS ou son sommeil.

La promesse de mémoriser sa vie mieux que sa propre mémoire fait définitivement partie du mouvement transhumaniste. L'objectif est de pouvoir conserver toutes nos interactions avec nos outils (ordinateur, portable) et objets connectés (voiture, vélo, frigidaire, verrou de porte) et surtout, tout l'environnement visuel et audio proche pour pouvoir les analyser et les consulter à postériori. Les valeurs du Tableau 3 représentent la taille réelle des captures typiques du lifelogging suivant les différentes sources[4]. Bien entendu, elles sont dépendantes de la personne et sont ici données à titre indicatif. Par ce biais, il est possible d'enregistrer un an de données audio sur une carte mémoire SD de 2016 de 512 Go. D'ici cinq ans, il sera possible en théorie d'enregistrer les données de la durée d'une vie sur le disque dur de son ordinateur portable. La capture totale des données nécessite la sauvegarde de toutes ces sources simultanément et, mise à part les données vidéo qui sont beaucoup plus lourdes, le stockage ne semble pas être une limitation. Dans le cas du stockage des données vidéo, le Tableau 2 indique que ce sera probablement possible dans une dizaine d'années.

Pour avoir une illustration du lifelogging, je ne peux que conseiller de voir l'épisode d'anticipation « The entire history of you » de la série *Black Mirror* de Charlie Brooker où les dérives d'un tel dispositif sont mises en exergue.

D'un point de vue technique, tous ces capteurs vont générer énormément de données qu'il faudra stocker, analyser, archiver, indexer afin de pouvoir les rendre utiles pour l'utilisateur final. Ceci offre de nouveaux défis en terme de traitement de données hétérogènes, de stockage, de recherche et de visualisation. Dans le cadre du lifelogging, aucune donnée n'est éliminée, car même si elle n'est jamais utilisée, toute donnée est potentiellement utile.

Les big data et la valeur de la donnée

Aujourd’hui, tout le monde parle des big data ou données massives et de la valeur de la donnée. Les grandes entreprises d’internet (par exemple Google et Facebook) l’ont bien compris en nous offrant des services gratuits dans le cadre desquels chacun d’entre nous offre ses données « idiotes » en surfant sur internet, utilisant telle ou telle application, etc. Ces données mises bout à bout et multipliées par le nombre d’utilisateurs ont une valeur considérable pour qui sait les faire parler. Faut-il éliminer de l’information ? Ces entreprises ont déjà choisi et répondu clairement que non. Si on prend comme exemple Twitter, application dans laquelle les utilisateurs publient des messages de 140 caractères maximum (les tweets), ceux-ci contiennent en réalité vingt fois plus d’informations (2.5 ko par tweet en moyenne[3]). Chaque tweet qui est échangé dans le monde contient à chaque fois, en plus du message proprement dit (les 140 caractères), la description de l’émetteur (pseudo, langue, origine géographique, etc.), le contexte (date, géolocalisation si activée) et l’historique des retweets. Ceci a l’avantage d’offrir une grande transparence, mais montre clairement qu’on essaie de garder toutes les informations disponibles, y compris le contexte (si cher aux archivistes). Aucune information n’est éliminée à priori. Pourtant, avec 500 millions de tweets par jour, soit 1.25 To, on pourrait faire pas mal d’économies d’espace de stockage en évitant la redondance de l’information. Pour continuer avec ce fameux micro-blog, la Bibliothèque du Congrès des États-Unis semble avoir compris l’intérêt d’archiver des données sans faire de désherbage. Et ce malgré le fait que le contenu de beaucoup de tweets peut paraître à priori peu intéressant et peu pertinent (par exemple « il fait bo ce matin »). En effet, ils ont décidé en avril 2010 de s’associer avec Twitter et d’archiver la totalité des tweets[14] émis à ce jour et d’en offrir l’accès gratuit. Cependant, après six ans de bonnes intentions, les tweets ne sont toujours pas disponibles et Zimmer[15] semble indiquer aussi bien des problèmes techniques et juridiques que commerciaux.

Le critère d’élimination diachronique

Les critères d’élimination utilisés de nos jours sont adaptés à nos connaissances et usages d’aujourd’hui. Un critère d’élimination qui vaut aujourd’hui ne vaut pas forcément demain. Pour faire un parallèle avec un domaine très différent, la sauvegarde de la biodiversité a aussi son intérêt et ces critères. L’exemple du rat-taupe nu du Kenya est emblématique. Ainsi, si le seul critère de la sauvegarde de l’animal était la beauté (voir Image 1), le pauvre n’aurait aucune chance d’être sauvegardé en tant qu’espèce. Par contre, des recherches de 2005 ont montré que l’animal pouvait vivre jusqu'à 50 ans, qu’il était insensible à la douleur et surtout qu’il avait une forte résistance aux maladies cancéreuses. Cela l’a remis de facto sur le podium des animaux suscitant l’intérêt des humains. Par analogie, effacer une donnée aujourd’hui ne veut pas dire qu’elle n’aura pas d’importance à l’aulne des critères du futur. Et si les coûts sont inférieurs ou équivalents, pourquoi éliminer l’information ?

Conserver la vie numérique de la donnée

Une première tendance en informatique est de garder un historique le plus précis possible des applications et de la vie du document. La seconde tendance est la redondance et la distribution de l'information. Ainsi, même si conserver le document, c'est bien, conserver la vie du document, c'est mieux ! Pouvoir avoir la trace de la naissance du document, des premiers mots de son auteur. Pouvoir voir qui a contribué à tel ajout ou à telle élimination (mais toujours réversible) de manière non ambigüe. Pouvoir voir que plusieurs copies du document ont évolué de leur côté, alors que le document original de l'auteur reste inchangé. Les historiens l'ont rêvé, les informaticiens l'on créé : un système de gestion des versions pour le développement informatique (dit « versionning »). L'outil le plus abouti et populaire à l'heure actuelle est le dépôt GitHub, interface web collaborative basé sur Git. Il est principalement utilisé pour la gestion des codes source, mais l'est aussi pour du texte ou des sites web. Ici, c'est l'auteur du document qui décide de la granularité des versions, mais toutes ces informations et bien d'autres sont conservées et donnent une grande valeur à ce dépôt. De plus, la pérennité de ces données semble assurée, car l'INRIA vient d'annoncer l'initiative « software heritage »[16] qui a pour objectif d'être l'archive universelle de l'open source.

Un autre exemple contemporain et intéressant est celui du bitcoin, crypto-monnaie la plus populaire et controversée. Il est actuellement utilisé pour des échanges commerciaux non contrôlés par un établissement bancaire ou étatique. Le bitcoin est basé sur une base de données distribuée qui se nomme la blockchain (ou chaîne de blocs) et qui a pour particularité de conserver toutes les transactions financières effectuées depuis sa création. Elle fait actuellement plus de 70 Go[17] (juillet 2016), existe en copie sur les ordinateurs de plus de 100 000 « mineurs » - ceux qui minent les bitcoins (comme on minerait de l'or) et gère de facto les transactions. Dans le blockchain, qui est une archive des transactions, tout est conservé (et partagé) et à valeur de preuve.

Conclusion

À l'heure du numérique et des données massives, il est, dans la plupart des cas, inutile d'éliminer l'information, car d'une part le coût du stockage continue de baisser énormément et d'autre part les algorithmes de recherche de données structurées et non structurées sont taillés pour gérer plusieurs milliards de documents plus ou moins hétérogènes. Lifelogging, quantified-self, blockchain et big data sont autant d'exemples où le choix de la conservation de toutes les données a déjà été effectué et ce malgré la masse considérable d'informations.

La tendance en informatique est à la conservation des données pour des raisons de traçabilité, de transparence et pour pouvoir reconstruire l'histoire du document.

Néanmoins, ne pas éliminer l'information ne veut pas dire ne pas filtrer l'information. Ainsi, sans faire disparaître l'information définitivement, l'intérêt de classer, catégoriser, voire d'enrichir, à encore tout son sens.

L'espace physique est fini. L'espace numérique, qui est *in fine* physique, l'est tout autant. Cependant, le numérique est, sans aucune comparaison ou mesure, le champion de la compression d'espace physique. Les exemples et idées développés ici montrent que les limites actuelles et surtout futures sont plutôt le fait d'utilisation extrême tel l'archivage du Web ou des expériences du CERN avec leurs tailles de stockage de plus de 200 Po (soit 200 000 To).

Paradoxalement, si la conservation et la diffusion sont facilitées par les technologies, il est urgent de trouver des moyens de mieux collecter l'information publique, voire l'information privée (courriels, réseaux sociaux fermés, photographies et films produits en masse comme autant de souvenirs qui ne dureront pourtant guère, faute de réflexion), car des millions de documents disparaissent tous les jours. Sur le web par exemple, la durée de vie d'une page est estimée entre cinquante et cent jours en moyenne selon Brewster Kahle fondateur principal de l' « Internet Archive ».

Mais la bonne nouvelle, c'est que la page Wikipédia du rat-taupe nu du Kenya est bien à l'abri parmi les Peta données d'archive.org et ne « souffre » d'aucun risque d'élimination !

L'auteur remercie Esther Bettiol pour sa relecture.

1 Steven C. Horii, « Archiving, Chapter 10: Future Storage Trends and Technologies » [en ligne], (consulté le 22.07.2016).
2 Rosenthal. David, «The Future of Storage» [en ligne], 2016, (consulté le 22.07.2016).
3 Valeski Jud, « Handling High-Volume, Realtime, Big Social Data », 2011, (consulté le 22.07.2016).
4 Gurrin, C., Smeaton, A. F., & Doherty. « Lifelogging: Personal big data ». Foundations and trends in information retrieval, 8(1), 1-125. A. R. 2014.
5 Merity, Stephen. « Common Crawl » [en ligne], 2016, (consulté le 22.07.2016).
6 « PetaBox » [en ligne]. 2014, (consulté le 22.07.2016).
7 « Internet lives stats » [en ligne]. 2016, (consulté le 22.07.2016).
8 Cocilova, Alex, «The astounding evolution of the hard drive» [en ligne]. 2013, (consulté le 22.07.2016).
9 Walter, Chip. «Kryder's law». Scientific American, 293(2), 32-33. 2001.
10 Larson, Jeanette. « Crew : A weeding Manual for Modern Libraries ». [en ligne]. 2008, (consulté le 22.07.2016).
11 « Elasticsearch » [en ligne], 2016, (consulté le 22.07.2016).
12 Banfi, E., Béguelin, F., & Gaudinat, A. « GEoTweet » (No. TRMASID 7). Haute école de gestion de Genève. 2016.

13 Bell, C. G., & Gemmell, J. « Total recall : How the e-memory revolution will change everything ». Dutton.
14 Matt Raymond. « How tweet it is! Library acquires entire Twitter archive » [en ligne], 2010, Library of Congress Blog, (consulté le 22.07.2016).
15 Zimmer, M. « The Twitter Archive at the Library of Congress: Challenges for information practice and information policy ». First Monday, 20(7). 2015.
16 « Software Heritage » [en ligne], 2016, (consulté le 22.07.2016).
17 « Taille de la blockchain » [en ligne], 2016, consulté le 22.07.2016).

Arnaud Gaudinat

Arnaud Gaudinat est ingénieur en informatique de formation. Après quinze ans d'expérience en recherche appliquée, il intègre en 2010 la filière information documentaire de la HEG. Il est actuellement adjoint scientifique et enseigne les sciences de l'information, la programmation et la gestion de contenu sur le Web. Il effectue principalement sa recherche dans le domaine de la fouille de données sur le Web et de la webométrie.