Bac à sable hexadécimal

Jeremy Pedrazzi

Cet article est issu d'une conférence intitulée : *Pour la reconstruction des processus d'écriture numériques de Derrida grâce à la « computer forensics » : reconstruction des données et matérialité numérique historique* qui a eu lieu le 19 octobre 2018 à Paris, pour le 50e anniversaire de l'Institut des textes et manuscrits modernes (ITEM) dans la demi-journée consacrée au volet de la critique génétique appliquée aux supports numériques.

Bac à sable hexadécimal

Jeremy Pedrazzi

1 Rencontre

Quand l'aventure du projet Derrida Hexadécimal a commencé, nous étions face à un fonds hybride sur le plan numérique. Nous avons donc cherché à caractériser ce qu'était un fonds numérique. Selon les interlocuteurs et les contextes, on parlera tantôt des supports, tantôt des fichiers. Il arrivera même que la notion couvre l'information, hors fichiers, stockée sur les supports.

Sur le plan matériel, nous avions accès à des supports numériques de différentes natures (disquettes, disques durs, CD-Rom). Ces derniers étant obsolètes, nous avions besoin d'en extraire les contenus. Le travail était déjà commencé pour deux disques durs et quelques supports externes de stockage. Il s'agissait alors principalement de dossiers compressés. L'IMEC nous donnait aussi accès à plus de 300 disquettes qui n'étaient pas encore sauvegardées. Nous avions bien conscience qu'il était important de transférer rapidement les informations présentes sur les différents supports magnétiques qui s'abîment dans le temps[1].

Pour procéder à la sauvegarde de l'information présente sur les supports physiques, il existe deux méthodes. Soit par l'extraction du contenu, où l'on copie les zones du support où des fichiers sont officiellement présents, soit par copie complète, ou « image-disque », incluant les zones de mémoire vierges.

[1] La compréhension que nous avons du numérique nous fait souvent oublier qu'il est dépendant d'un support physique.

Dans le premier cas, on peut par exemple récupérer 20 Mo de données sur un support pouvant en contenir 100. Les supports issus d'extractions pouvaient être sous la forme d'une arborescence de fichiers ou d'un fichier d'archive[2].

L'image-disque (ou "dump") est la méthode la plus intéressante pour le généticien. Elle permet une copie complète des informations en gardant tous les 0 et 1 d'un support, y compris des zones considérées comme libres, sans les interpréter.

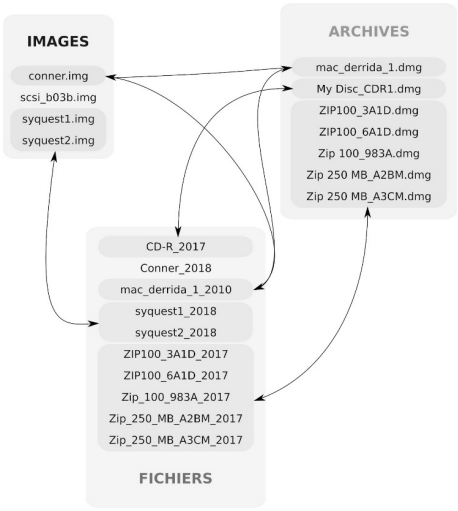

IMAGE 1

Cela a pour conséquence de permettre, entre autres, l'extraction des fichiers effacés, ou plus exactement de lire les traces de fichiers laissées sur le support. En effet, les zones libres sont en réalité des zones que l'on ne référence pas, ou plus, dans le catalogue du support, mais dont l'information, si elle a existé, est toujours présente. Elle ne disparaît en réalité que lorsqu'on utilise à nouveau cette même zone pour un nouveau fichier. L'action de "supprimer" un fichier ne fait en réalité qu'effacer la référence à ce dernier dans le catalogue du support[3]. La conséquence est que plus un support est rempli, moins il est possible de trouver des fichiers effacés ou plus généralement des traces sur sa surface.

2 Volume

Pour donner un ordre d'idée du volume numérique du corpus Derrida de l'IMEC, nous disposons actuellement d'environ 700 Mo de données, avec de la redondance dans les fichiers, sous forme d'images-disques, d'extractions de disques durs et de cartouches de mémoire[4].

2 Dans le fichier archives, les dates du système d'exploitation d'origine pouvaient être conservées.

3 Il s'agit là d'un mensonge du numérique, qui, pour des questions de performance (rapidité) et d'usure des supports, n'efface pas toute l'information, par défaut, du support.

4 Les disques durs de l'époque étaient de petites capacités. 40Mo pour ceux déjà numérisés (pour donner un ordre d'idée, les disques durs internes des ordinateurs actuels sont de l'ordre de 1To, soit 25.000 fois plus).

Il est difficile d'évaluer le volume d'un corpus numérique. Pour le physique, on peut rapidement avoir une idée de volume, mais dans un espace sans topographie normée, la mesure est compliquée.

Si l'on devait évaluer la taille physique que prendrait ce corpus numérique une fois imprimé, il représenterait en hauteur 5 m de papier[5].

3 « Computer forensics »

Les copies systématiques des supports impliquées par les méthodes de la « computer forensics » offrent une immersion dans l'environnement de travail numérique de l'auteur. Elles ont pour objectif de transférer la totalité de l'information existante au moment de la capture, sans tenir compte du sens ou de la structuration de cette information. Il s'agit d'une photo instantanée d'un support, laissant alors au chercheur, dans le futur, la liberté de l'exploiter. À titre de comparaison dans le monde physique, il faudrait ajouter, aux manuscrits papiers, le bureau entier de l'auteur, corbeille à papier et agendas compris, pour que l'on apprécie mieux l'analogie.

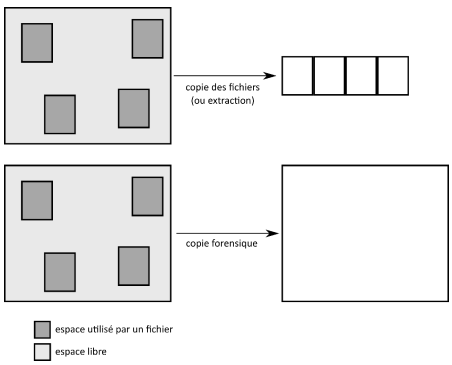

Image 2

4 Nativement numérique

Si nous avons aujourd'hui l'habitude de travailler sur un ordinateur, cela reste en majeure partie sur des documents physiques (manuscrits, dessins) qui ont été numérisés.

Il est ainsi possible de travailler sur des reproductions numériques de manuscrits de grands auteurs sans avoir à accéder aux originaux, avec les précautions que cela impose. Si le rendu visuel est excellent pour des tâches de simple lecture, il faut passer par des étapes de transcription, ou de reconnaissance automatique de caractères, pour obtenir des documents dans lesquels il est possible de faire des recherches textuelles.

5 Calcul fait avec 1page = 15ko, 750Mo environ 5m de papier imprimé.

Les contraintes du manuscrit sont alors transposées dans le monde numérique. Il est en effet difficile de dater un document à partir de l'image. Ce type d'information doit donc être encodé sous forme de métadonnées par le chercheur après une étude codicologique et grâce à la connaissance qu'il a d'un plus grand corpus.

Le travail sur des documents numérisés est donc hérité du travail sur les manuscrits. Pour les informations liées aux manuscrits physiques qui ne seraient pas prises en compte par les méthodes de numérisation, il nous faut alors enrichir le fichier numérique obtenu par des données complémentaires « métadonnées ». Loin du manuscrit original, le chercheur ne peut en effet pas connaître l'épaisseur du papier utilisé pour un document avec la simple image issue d'une numérisation. Ces métadonnées doivent être ajoutées après la numérisation pour être liées au document numérique.

Il n'y a donc pas d'information supplémentaire par rapport à celles portées par le manuscrit sur sa forme numérisée.

Avec les documents « nativement numériques », ou « born-digital », une quantité non négligeable d'informations, liées tantôt aux systèmes d'exploitation, tantôt aux logiciels de traitements de texte, se crée, se transforme et se répand au fil du temps, souvent sans que l'utilisateur en ait conscience.

Pour reprendre l'exemple des dates, nous savons aujourd'hui que tous les fichiers que nous créons sur un ordinateur[6] portent, au minimum, une date de création. Le travail de classement, comparé à celui sur papier, change radicalement. Mais il n'y a pas que les dates dont le traitement a changé, et l'auteur, sans le savoir, en fonction des couches logicielles utilisées, peut laisser dans ses documents ou simplement dans ses supports des informations riches que le généticien saura apprécier.

5 Machines obsolètes

Pour resituer le contexte des archives et fichiers à notre disposition, nous avions des données, vraisemblablement nativement numériques, créées entre la fin des années 1980 et le début des années 2000, sur des machines de la marque Apple. Exploitant les fichiers sur des systèmes d'exploitation actuels, nous avons rencontré des difficultés concernant notamment l'encodage caractère ou encore des dates de fichiers qui n'étaient pas fiables (ex. : 2032 ou 1970)[7].

6 C'est vrai aussi sur les tablettes et smartphones.
7 Certains reconnaitront la date Unix 0 attribuée au fichier par le système d'exploitation.

6 Logiciels obsolètes

Si le matériel qui avait permis la création de ces documents était obsolète, le format des fichiers l'était tout autant. Impossible d'installer, sur des machines récentes, des logiciels comme Apple MacWrite ou même Microsoft Word 4. Après avoir écarté l'idée de travailler sur des machines de l'époque, nous avons « tenté » l'ouverture des fichiers avec LibreOffice, qui nous avait déjà aidés sur des formats aujourd'hui exotiques. Nous avions trouvé l'outil idéal pour lire la totalité des fichiers à notre disposition[8].

Il faut remarquer que la quasi-totalité des fichiers, tous supports confondus, n'avait pas d'extension comme nous avons aujourd'hui l'habitude d'en voir (.doc, .odt, ...). La brique logicielle libmwaw permet alors de détecter par les premiers octets d'un fichier quel est le logiciel de traitement de texte à l'origine de ce dernier.

7 Premiers tests

Pour tester nos outils et notre méthodologie, nous avons choisi de traiter le corpus du séminaire sur la peine de mort, qui s'est déroulé sur deux années à l'EHESS (1999-2000 et 2000-2001)[9].

Sur les supports numériques étudiés, nous avons des fichiers de toutes les séances, sauf la dernière (10e de la deuxième année, qui a été transcrite à partir d'une captation audio). Le corpus était intéressant pour l'exercice, car facile à isoler, présent sur plusieurs des supports à notre disposition et créé avec le même logiciel (MacWrite Pro).

Dans la prolifération des fichiers, nous avons trouvé, pour chaque séance, de multiples fichiers de même nom. Pour vérifier qu'ils n'avaient pas le même contenu, nous avons calculé pour chacun son "empreinte numérique"[10].

Les séances, pour lesquelles nous avions parfois douze fichiers de mêmes noms, ont révélé beaucoup de fichiers identiques. Pour la séance 6 par exemple, les douze copies trouvées n'étaient qu'un seul et même état. Nous avons aussi pu montrer qu'il existait, pour la séance 14, quatre états différents.

8 LibreOffice est une suite bureautique libre et gratuite qui intègre les bibliothèques du Document Liberation Project et plus spécifiquement depuis sa version 4.1, de juillet 2013, la librairie libmwaw portée par Laurent Alonso, qui permet d'ouvrir certains formats de fichiers anciens et obsolètes utilisés sous Mac OS Classic.

9 Et qui a été édité en deux volumes aux éditions Galilée en 2012 et 2015.

10 Sous la forme d'une clé MD5 de 32 caractères qui assure, pour deux fichiers identiques, une même clé.

Avec l'utilisation des clés MD5, il est donc possible, sans programmation, de repérer des états génétiques dans des corpus numériques. Nous avons ensuite exporté les quatre fichiers correspondants en mode texte pour les comparer avec l'outil d'alignement de texte MEDITE[11].

Il a donc été possible d'organiser les quatre états de cette 14e séance et de repérer les différences entre eux.

8 Conclusion

Ce simple exercice de comparaison de fichiers de mêmes noms nous a permis de faire le constat que l'archive numérique est un corpus qui ne va pas de soi. Les sauvegardes, par exemple, ont créé des clones qui n'existeraient pas en aussi grand nombre dans un corpus physique.

Pendant ce travail, le contenu des disques (pilotes, allusions dans les textes) nous a permis de retrouver des informations sur le matériel utilisé (imprimante laser, lecteur Syquest, ordinateurs dont nous n'avons plus de trace…). Il nous faut maintenant continuer l'exploration de la totalité des supports numériques accessibles de Derrida et enfin intégrer dans l'étude des sous-corpus les archives papier comme objets numérisables.

[11] Porté par Jean-Gabriel Ganascia https://obvil.sorbonne-universite.fr/developpements/medite.