

Conserver le support, conserver l'information : des défis pour les institutions patrimoniales

Stéphane Reecht

► **To cite this version:**

Stéphane Reecht. Conserver le support, conserver l'information : des défis pour les institutions patrimoniales : Communication donnée au congrès international du cinquantenaire de l'Institut des textes et manuscrits modernes, Paris, École normale supérieure, Bibliothèque nationale de France, 17-20 octobre 2018. 2018. hal-02162215

HAL Id: hal-02162215

<https://hal-bnf.archives-ouvertes.fr/hal-02162215>

Submitted on 21 Jun 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Congrès international du cinquantième de l'Institut des textes et manuscrits modernes

La critique génétique comme processus. (1968-2018)

Paris, École normale supérieure, Bibliothèque nationale de France, 17-20 octobre 2018

~~~~~

### **Conserver le support, conserver l'information : des défis pour les institutions patrimoniales.**

**Stéphane Reecht, conservateur, expert de préservation numérique - Bibliothèque nationale de France, département de la Conservation**

Le rôle des institutions patrimoniales est depuis longtemps de conserver du numérique, et de plus en plus non du numérisé mais du numérique natif. Dans cette communication sur les défis qui leur sont posés pour rendre intelligibles sur le long terme des documents numériques, nous nous intéresserons particulièrement à la question des traitements à appliquer au support, de la manière dont on peut aborder les problématiques d'accès, en particulier pour permettre un usage lié à la critique génétique, avant de finir sur les problèmes plus spécifiquement juridiques que pose cette double exigence de conservation et d'accès.

#### **1. S'abstraire du support**

##### A) La durabilité du support

C'est enfoncer une porte ouverte que rappeler la fragilité du support informatique, et la difficulté à le préserver sur le long terme. Du côté des institutions patrimoniales, on n'ose plus vraiment croire, en tout cas pour le moment, à l'existence d'un support qui, en plus d'être abordable financièrement et pratique d'utilisation, aurait des caractéristiques intrinsèques permettant d'envisager de le conserver plusieurs décennies voire plusieurs siècles avec un risque raisonnablement bas de perte de données. Quand bien même il existerait, il y a tout lieu de penser qu'il serait avant tout un outil pour les professionnels, hors de portée du grand public auquel appartiennent les producteurs de données que sont les écrivains, artistes et autres créateurs auxquels nous nous intéressons. Le besoin de gérer le support après réception existera toujours.

Il nous faut donc faire avec la fragilité du support, les risques d'effacement accidentel ou par excès d'utilisation, sa dégradation naturelle d'ordre physico-chimique, ses altérations dues à de mauvaises conditions de conservation, son obsolescence enfin (le matériel de lecture, la connectique et les logiciels associés finissent par disparaître du marché plus ou moins

brutalement). Une bibliothèque, un service d'archives, un musée, ont par vocation des moyens pour répondre à une partie de ces problèmes, mais il n'en reste pas moins nécessaire d'envisager toujours une migration à plus ou moins long terme. Oui mais quand ? L'envisager le plus tôt possible après réception du don ou du dépôt numérique permet de réduire le risque de perte de données, quelle que soit la cause. Mais repousser cette primo-migration, outre que c'est souvent une nécessité dictée par le manque de temps, de moyen, d'infrastructure, de connaissance ou d'organisation (ou des cinq à la fois), peut avoir un avantage : ne pas s'exposer à faire des erreurs lors de la migration qui pourraient conduire à de la perte d'information. Nous y reviendrons.

## B) Pérenniser l'information

Le mot est lâché : ce qui importe, au fond c'est l'information, c'est-à-dire, dans son acception la plus large, toute connaissance qui peut être échangée. La donnée est la forme que prend cette information, et cette donnée est encodée d'une certaine manière et organisée en un ou plusieurs objets numériques selon des règles particulières, qui sont celles du format. Sachant cela, on peut postuler que, si l'essentiel est de pérenniser de l'information, il est tout à fait admissible de transformer la forme qu'elle prend, pour les besoins de cette pérennisation. On se donne ainsi les moyens, par exemple, d'échapper à l'impasse des formats propriétaires. On peut aussi considérer sans état d'âme que le support n'est à prendre que comme un medium, qui n'a de toute façon pas vocation à accompagner indéfiniment l'information qu'il héberge. Un medium dont il faut s'abstraire. Position peut-être radicale. C'est en tout cas ainsi que nous invite à raisonner le modèle conceptuel de base utilisé dans le monde de la préservation numérique, l'OAIS (Open Archival Information System, norme ISO-14721, issue des travaux du CCSDS, l'organisme normatif des agences spatiales du monde entier<sup>1</sup>). Bien sûr, ce modèle ne nous dit pas que le support de stockage n'est pas important, mais qu'il faut pouvoir regarder au-delà de cet aspect du problème pour pouvoir envisager la conservation de l'information sur le long terme. En somme, il nous dit ce qu'il faut faire si l'on veut s'abstraire du support.

À sacraliser le support, on risque d'aller au devant de désagréments certains. A minima, on peut en venir à s'intéresser à des fichiers de peu voire d'aucun intérêt, comme des fichiers-système qui sont produits automatiquement sur les disques durs ou les clés USB, et qui n'ont d'usage que pour le fonctionnement de l'appareil. Or, on ne veut justement plus utiliser le support de stockage mais le figer.

On risque également d'accorder une importance démesurée à une information qui est celle de la date de dernière modification d'un fichier. Bien souvent, cette date n'est que celle de l'arrivée du fichier dans un système de fichier, en l'occurrence celui du support de stockage. La dernière modification du contenu lui-même (« information de contenu » dit-on en OAIS)

---

<sup>1</sup> <https://public.ccsds.org/pubs/650x0m2.pdf> (version anglaise originale) ou <https://public.ccsds.org/Pubs/650x0m2%28F%29.pdf> (traduction française).

peut être bien antérieure et remonter à une époque où le fichier était sur un autre support. Inversement, il suffit parfois d'ouvrir un fichier sans le modifier pour que cette date de dernière modification soit changée pour celle du jour (encore une fois, ce comportement peut varier d'un système de fichier à un autre). Et modifier des fichiers, les déplacer, les réorganiser, peut être une étape du travail du bibliothécaire et de l'archiviste qui a en charge le classement des documents numériques qui lui sont confiés. On considère en effet que le bibliothécaire ou l'archiviste doit être un intermédiaire entre le producteur des données et l'utilisateur et que son travail consiste notamment à rendre intelligibles les données pour l'utilisateur. Dans ce contexte, il apparaît alors plus pertinent d'enregistrer à côté, sous forme de métadonnées, des informations telles que date de création ou de dernière modification (tout en ayant conscience de leurs limites), ou encore l'arborescence originale (i.e. l'emplacement logique du fichier dans le support de stockage).

Au-delà du support, il faut aussi veiller à ne pas trop sacraliser les objets numériques que le producteur des données y a laissés. En effet, si ceux-ci sont dans un format propriétaire qui nécessite un logiciel de lecture à clef de licence (c'est-à-dire payant), il y a de grandes chances que dans un délai de quelques années on ne puisse plus les lire, ou avec un logiciel qui ne les représentera pas fidèlement (malgré les promesses de son fabricant sur la compatibilité descendante de son produit...). On a tout intérêt alors à en faire une copie dans un format qui offre plus de garanties de pérennité. De même pour des formats propriétaires qui sont largement diffusés, mais en voie d'obsolescence comme RTF ou WORD, XLS, PPT etc. La stratégie sera alors de produire une copie des fichiers concernés dans un format meilleur pour la conservation à long terme. Dans certains cas, cette stratégie rendra probablement les meilleurs services aux chercheurs intéressés par la génétique des œuvres. Je prends un exemple dans le monde de la création vidéo. La plupart des professionnels du secteur utilisent des outils logiciels propriétaires, payants et à obsolescence programmée, fonctionnant sur système d'exploitation Mac-OS, notamment le logiciel de montage Final Cut Pro. Cet outil très perfectionné permet notamment d'enregistrer des états successifs d'un film en cours de montage, à des fins de simple sauvegarde, mais aussi pour pouvoir restaurer une version précédente de manière volontaire. On va ainsi pouvoir connaître de près le processus de création, pour peu que le monteur ait bien enregistré des états successifs. Mais pour que cela soit possible sur le long terme, rester dans cet univers propriétaire est beaucoup trop risqué. À la BnF, la stratégie envisagée est donc de créer des exports dans un format standard, qui correspondront aux états successifs du film en cours de montage. On se libèrera ainsi de la nécessité d'avoir recours au logiciel Final Cut Pro.

Bien sûr, on ne peut exclure que le futur ne soit pas aussi sombre que nous le craignons, que l'obsolescence soit freinée, que les industriels mettent à disposition librement les anciennes versions de leurs logiciels (comme l'incite l'Unesco à travers le projet Persist), et que par des techniques de virtualisation et d'émulation, l'on parvienne à utiliser durablement ces objets numériques que l'on ne pensait pas pouvoir conserver sur le long terme initialement. C'est

pourquoi, par mesure de prudence, on choisit en général de garder le fichier original ainsi que son dérivé. Cela n'est pas sans poser de problèmes de place, toutefois, pour les fichiers volumineux, et c'est un élément qui doit être mis dans la balance par l'institution patrimoniale avant d'accepter un don ou de conclure une acquisition onéreuse.

C) Un exemple : la stratégie de la BnF en la matière

À la Bibliothèque nationale de France, il y a longtemps que nous recevons en don ou en dépôt, ou que nous achetons, des documents numériques uniques sur support (j'exclus les parutions, soumises au dépôt légal, qui sont hors sujet aujourd'hui). Des traitements leur ont pour la plupart été appliqués, mais nous sommes en train de formaliser un circuit et des procédures pour résorber le rétrospectif et être en mesure d'accueillir sereinement les entrées futures. Ce circuit et ces procédures sont inspirés par ce que je viens de vous exposer. Il s'agira ainsi de s'abstraire le plus vite possible du support, quel que soit son type et quelle que soit sa technologie, en copiant les données qu'il contient sur un espace de stockage sécurisé à tous points de vue. En même temps, une analyse doit être effectuée, qui produit pour chaque support dématérialisé un fichier de métadonnées dit « manifeste » qui fait la liste des fichiers, avec leur emplacement (chemin logique), leur date de modification, leur empreinte numérique etc. Immédiatement, les fichiers systèmes sont nettoyés. Le bibliothécaire qui devra ensuite traiter le fonds (le « gestionnaire de collection » dans notre jargon) pourra ensuite demander une copie de ces fichiers dans un espace de travail où il pourra les réorganiser, les modifier etc. Comme le support lui-même peut être porteur d'information (étiquette renseignant sur son contenu par exemple), et par souci de traçabilité, il est également prévu une prise de vue photographique qui sera jointe aux fichiers déchargés. On s'abstrait ainsi du support mais sans le nier tout à fait. Il est d'ailleurs bien entendu que celui-ci pourra être conservé à titre de témoignage. On peut aussi imaginer qu'un support qui n'est plus lisible aujourd'hui le sera demain si les progrès dans les techniques de récupération se poursuivent (cela s'est vu avec des disquettes par exemple).

## **2. Permettre l'accès**

Envisageons maintenant les choses sous le prisme de l'accès.

Aucune institution patrimoniale de nos jours ne travaille sur des documents sans penser à l'accès qu'auront le grand public, les chercheurs ou les professionnels, à ces documents. Conserver seulement pour conserver n'est pas une activité acceptable de nos jours (si tant est qu'elle l'ait jamais été). Mais si nous travaillons pour l'accès, il faut bien reconnaître que nous avons tendance à travailler pour un type d'accès particulier. Celui-ci est appuyé sur une description la plus fidèle et la plus normée possible dans un catalogue ou un inventaire, qui rend compte de la manière dont les professionnels du patrimoine ont compris les documents et leur organisation en tant qu'ensemble, et ont pu modifier cette organisation pour faciliter la compréhension, l'accès donc, à ces documents. C'est aussi un moyen de rendre compte de la complémentarité entre papier et numérique qui caractérise souvent un fonds de créateur.

Derrière ce type d'accès, il peut y avoir des traitements qui aboutissent à supprimer des doublons ou éventuellement même des documents jugés sans intérêt. Archiver, c'est choisir !

Ce type d'accès est également appuyé sur un principe de protection des documents, voire de leur contenu, qui conduit bibliothèques et archives à mettre des freins à la consultation, voire à l'empêcher pour certaines pièces. Dans le cas du numérique, cela revient concrètement à donner accès à des copies de consultation, qui peuvent être dans un format différent de celui de l'original, et qui peuvent être partielles. Et le mode de découverte reste le même que celui des documents physiques (recherche dans un catalogue, navigation dans un inventaire). L'utilisateur consulte également dans un environnement dédié sur lequel il n'a pour ainsi dire aucune prise, soit celui d'une application en ligne (bibliothèque numérique), soit un poste dédié (qui, grâce aux techniques actuelles, peut tout de même être déporté sur son ordinateur personnel<sup>2</sup>).

Il faut bien reconnaître que ce mode d'accès ne permet pas tous les usages possibles, notamment ceux qui visent à « fouiller » un support de stockage ayant appartenu à un créateur, comme nous venons de le voir avec l'exemple de Jacques Derrida. On ne peut cependant pas amener le support informatique sur la table du chercheur, il faut donc donner accès à une image de ce support, la fameuse image disque au format ISO, IMG ou autre. La consultation doit toujours s'envisager dans un environnement sécurisé, mais il y a là un progrès indéniable. Techniquement, les difficultés sont minces en théorie, tant ces technologies sont connues aujourd'hui. Il faut seulement se les approprier et mettre en place l'infrastructure technique adéquate. Une limite toutefois : le bibliothécaire ou l'archiviste est ainsi privé de la possibilité de sélectionner du contenu à ne pas divulguer, ce qui peut être un frein important à l'adoption de cette technique. Heureusement, le projet BitCurator, mené par l'université de Chapel Hill (USA, Caroline du Nord), fournit enfin des outils aux institutions patrimoniales qui souhaitent mettre en place la dématérialisation de supports en vue d'une analyse poussée<sup>3</sup>. C'est l'application des *digital forensics* rendue possible aux bibliothèques et archives. Cette suite d'outils open source, relativement faciles d'utilisation, permet la création d'images de supports, un certain nombre de traitements dont l'anonymisation, l'exploration du contenu et la mise à disposition des images de supports aux chercheurs. Déjà en vogue dans le monde académique, BitCurator est regardé de près par les bibliothèques nationales notamment, qui sont de plus en plus conscientes que les services aux chercheurs ne doivent pas se limiter à la constitution de bibliothèques numériques et à l'accueil dans des salles de lecture confortables et à l'ambiance studieuse.

L'utilisation d'images de support doit toutefois être faite à bon escient. Il ne faut pas perdre de vue en effet que cette technique revient à capturer l'ensemble du support, y compris ses emplacements inutilisés. Si un disque dur de 1 téraoctet est rempli à moitié, le déchargement par simple copie des fichiers qu'il contient nécessitera un volume de stockage de 500 Go ;

---

<sup>2</sup> Voir par exemple le service AVEC de la BnF : <http://avec.bnf.fr/>

<sup>3</sup> Voir le site officiel du projet et du produit : <https://bitcurator.net/>

alors que la création d'une image disque demandera un volume de 1 téraoctet. Dans ce cas, si l'intérêt scientifique de la deuxième technique n'est pas avéré, on aura tendance à se poser sérieusement la question de sa pertinence.

Au-delà, cette approche, qui vise à capturer un contenu pour recréer virtuellement la « boîte » qui le contenait, peut être adoptée pour des documents qui n'ont pas à proprement parler de support, dont le support en fait, si l'on peut dire, est déjà lui-même virtuel. C'est le cas des boîtes-mail : on peut considérer chaque courriel écrit ou reçu comme une lettre écrite ou reçue, indépendamment de son contexte matériel de production et de gestion. On peut aussi s'intéresser aux à-côtés (brouillons, corbeille, contacts, organisation des dossiers) pour en tirer un sens supplémentaire et potentiellement fécond. On peut aussi vouloir analyser le contenu des courriels dans leur globalité et non individuellement, pour faire de l'analyse sémantique par exemple, mais même sans aller jusque-là, pour identifier rapidement des interlocuteurs ou des noms cités sans avoir à faire un dépouillement fastidieux. Dans cette optique, l'Université de Stanford a développé un outil très prometteur appelé ePADD<sup>4</sup>, qui permet de capturer une boîte courriel à distance (avec l'assentiment du propriétaire, cela va sans dire), l'empaqueter dans un format archivable, et la charger dans un espace de consultation qui permet sa visualisation classique, mais aussi la découverte par d'autres moyens notamment par des mots-clés extraits automatiquement. Le mode de découverte reste encore relativement sommaire, mais au moins est-il presque immédiatement compatible avec les exigences d'une institution patrimoniale. Une autre limite d'ePADD est que l'indexation par mots-clés est conçue pour le moment seulement pour la langue anglaise ; il n'en reste pas moins que c'est un outil très prometteur. On entrevoit ainsi que l'étape suivante pourrait être un outil analogue qui permette de capturer un Google drive ou une Dropbox.

Dans le même ordre d'idée, on peut envisager de capturer des SMS. Il n'est en effet pas imaginable de conserver le téléphone portable d'une personne physique. L'expérience a déjà été tentée en bibliothèque, ainsi à la BnF avec Pierre Guyotat<sup>5</sup>. Le sujet est moins mûr, mais apparemment moins complexe également.

### **3. Les problèmes juridiques**

Ce sujet nous mène aux problèmes d'ordre juridique que nous rencontrons, notre activité étant assez encadrée de ce point de vue, en particulier dès lors que l'on donne accès à des documents au public.

La généralisation d'internet et la possibilité de diffuser en ligne du patrimoine numérique ou numérisé a amené les institutions patrimoniales à se poser assez tôt la question de ce qui est diffusable au-delà de la salle de lecture, de le définir (avec l'aide de leur tutelle et du législateur) et de s'organiser en conséquence en mettant en place des niveaux de diffusion

---

<sup>4</sup> <https://library.stanford.edu/projects/epadd>

<sup>5</sup> <http://bibliographie-historique.bnf.fr/Biblio/preuves-de-vie-les-sms-et-e-mails-de-pierre-guyotat-entrent-a-la-bnf>

différenciés. Dans cette optique, on pourrait penser que la consultation de supports et de documents numériques récupérés de créateurs divers ne devrait pas poser de problème particulier. Après tout, il s'agit d'œuvres de l'esprit, qui relèvent de la propriété intellectuelle et sont par conséquent soumises au droit d'auteur, lui-même justement pris en compte par les institutions patrimoniales.

Mais une difficulté naît du fait que l'on ne sait pas toujours à l'avance ce qu'on découvrira sur le support ; le créateur lui-même peut l'avoir oublié. Il peut aussi avoir effacé des fichiers et les croire, de bonne foi, impossibles à récupérer. Sauf que, tant qu'il n'y a pas eu de réécriture sur les emplacements du support concernés, les données ne sont pas physiquement effacées, et peuvent être récupérées pour peu que l'on dispose de l'outillage adéquat. De même pour des copies de sécurité faites automatiquement par les logiciels ou le système et qui peuvent subsister (cf. intervention de Thorsten Ries). Même si l'utilisateur qui consultera l'image du support n'aura probablement pas cet outillage à disposition, il est de notre responsabilité d'avertir le producteur ou son ayant droit de cette possibilité.

Autre élément qui peut se révéler source de complexité : un créateur peut avoir rassemblé à titre privé, pour son travail propre, de la documentation sous forme de texte, d'image, de vidéo etc. À partir du moment où une institution patrimoniale les met à disposition du public, il en fait ce qu'on appelle en droit une représentation. Elle peut pour cela s'appuyer sur l'exception au droit d'auteur prévue par le Code du patrimoine (si l'on reste dans le périmètre français), mais celle-ci ne l'autorise ni à rendre disponible la ressource sur internet, ni à permettre la reproduction par l'utilisateur en salle de lecture. Pour donner un exemple, c'est un cas que rencontre la BnF avec le fonds donné par Amos Gitai en avril de cette année. En plus se pose la question des droits voisins, ceux du monteur, du mixeur etc. Il faut que le donateur les ait négociés en amont du don avec eux.

De même, on se rend compte que la transmission de propriété du support n'est pas toujours évidente. On ne sait pas toujours si le don ou l'acquisition concerne des fichiers numériques ou les supports qui les contiennent, et cette indétermination peut exister jusque dans la convention et le pacte adjoint qui l'accompagne et donne le détail. On a déjà vu également le donateur revenir vers le conservateur pour récupérer des fichiers qu'il a perdus. En général on ne montre pas de mauvaise volonté dans ce cas. Mais que se passe-t-il si l'on a transformé les fichiers entre temps sans conserver l'original ? Et comment justifier qu'on ait supprimé d'éventuels doublons ?

La solution consiste donc à négocier les droits au cas par cas. Cela peut paraître étonnant de le rappeler, mais ce n'est pas une évidence pour tout le monde, et particulièrement pas pour le donateur. Dans le cas de dons mixtes (papier et numérique), il a souvent tendance à se focaliser sur le papier. Alors que c'est souvent pour le numérique que l'on a le plus besoin de clarifier les droits.



## **Conclusion**

Les institutions patrimoniales ont commencé à prendre la mesure des besoins de la recherche liés aux supports informatiques. Des outils existent déjà pour s'abstraire du support tout en gardant la trace et la logique. Un champ d'exploration demeure, bien que le défrichage ait déjà commencé dans des musées ou dans des laboratoires universitaires : celui de la même démarche appliquée à des ordinateurs entiers.