Les collections de la Radio, Télévision, et du Web audiovisuel : continuités et ruptures

Claude Mussou
Institut national de l'Audiovisuel (INA)
Bry-sur-Marne, France

Session:

148 — *Copyright law and legal deposit for audiovisual materials* — Audiovisual and multimedia with Law Libraries

Résumé:

Depuis sa création en 1974, l'Ina a pour mission de collecter, préserver, et rendre accessible les collections de la radio et de la télévision françaises. D'abord organisé pour répondre à des besoins professionnels d'archivage des productions des télévisions publiques, l'institut est rapidement devenu l'organisme de référence pour la conservation du patrimoine audiovisuel français. Aujourd'hui, l'Ina est considéré comme l'une des archives audiovisuelles et numériques les plus importantes au monde, avec un fonds de plus de 4 millions d'heures d'enregistrements télévisés et radiodiffusés, qui court depuis les premières diffusions de chacun des medias et s'enrichit chaque année de plus de 800.000 heures de programmes collectées au titre du Dépôt Légal.

A l'aube du 21e siècle, alors que le Web offrait des opportunités inédites aux secteurs de l'édition et de la diffusion, que de nouveaux acteurs profitaient des avancées de la révolution numérique pour diffuser des contenus audiovisuels en ligne, en France le dépôt légal était étendu au Web. Notons que le législateur a jugé nécessaire d'en partager la responsabilité entre deux institutions dépositaires, la BnF et l'Ina ; l'institut ayant la charge de l'archivage des sites en relation avec le secteur de l'audiovisuel ainsi que des Services de Medias Audiovisuels à la demande, dits SMADs.

L'Ina a commencé à collecter les sites Web liés au secteur audiovisuel en février 2009. Cette activité vient compléter et assurer la continuité de ses collections de programmes radio et télédiffusés, elle s'appuie sur le développement et la mise en œuvre d'outils et dispositifs innovants pour la collecte, le stockage, et l'accès aux contenus issus des « nouveaux medias »

L'Ina : l'institution patrimoniale française pour le son, l'image et les contenus Web audiovisuels

Depuis sa création par la loi de 1974, l'Institut National de l'Audiovisuel, Ina, est chargé de la collecte, préservation et valorisation des collections audiovisuelles françaises. D'abord organisé pour répondre à des besoins professionnels d'archivage des productions des télévisions publiques, l'institut est rapidement devenu l'organisme de référence pour la conservation du patrimoine audiovisuel français, une position qui s'est confirmée au moment de la disparition du monopole national de la télédiffusion.

En effet, au milieu des années 1980, tandis que des autorisations d'émettre étaient accordées au jeune secteur privé de l'audiovisuel, aucun cadre légal ou réglementaire ne garantissait l'archivage et la sauvegarde, à des fins patrimoniales, des programmes diffusés par les nouvelles chaînes. La communauté universitaire et académique n'eut alors de cesse de défendre l'idée que les émissions de télévision et de radio constituaient, au même titre que les autres supports de connaissance et de mémoire, des témoignages et sources pour les futures générations de chercheurs. En vertu de son expérience et de sa légitimité dans le domaine de la collecte et de la préservation de fonds audiovisuels, l'Ina fut désigné comme responsable du Dépôt Légal de la radio et la télévision par la loi du 20 juin 1992 [1]. Deux décennies plus tard, l'institut est considéré comme le fonds d'archives audiovisuelles le plus important au monde, avec plus de 4 millions d'heures d'enregistrements télévisés et radiophoniques. Les collections courent depuis les premiers enregistrements de chacun des medias et connaissent une croissance annuelle de 800.000 heures de programmes captés numériquement, 24 heures par jour, 7 jours par semaine pour 100 chaînes de télévision et 20 stations de radio.

L'ère numérique a évidemment ouvert de nouveaux horizons pour les collections vieillissantes de l'Ina et un vaste plan de numérisation a été engagé dès 1999 pour sauvegarder et transférer aux formats numériques environ 830.000 heures de ses collections analogiques menacées de vétusté ou de disparition. D'ici 2015, ce plan de numérisation s'achèvera et les contenus auront déjà « migré » plusieurs fois afin d'assurer leurs longévité et accès pérenne. Les verrous liés au droit de la propriété intellectuelle ont été levés pour une partie des collections (30.000 heures) afin de les rendre accessibles en ligne pour le grand public. Un accès en ligne restreint est offert pour l'usage professionnel des archives dont l'Ina détient les droits producteurs (plus d'un million d'heures). Les étudiants et chercheurs peuvent, quant à eux, rechercher, visionner et analyser, dans les emprises de l'institut, l'intégralité des collections, soit plus de 4 millions d'heures.

Parallèlement à la transition de l'analogique vers le numérique pour la transmission et la préservation du son et des images animées, le Web est devenu un outil incontournable pour la publication et consultation d'une grande variété de contenus. Les diffuseurs traditionnels, comme les nouveaux entrants issus du secteur des télécommunications ont bien sûr saisi les occasions offertes par les technologies du numérique (compression de fichiers, accès haut-débit) pour distribuer des contenus audiovisuels en ligne et engager des stratégies de diffusion multi plateformes, accompagnant ou suscitant une évolution – voire révolution - des usages pour la

[1] Loi du 20 juin 1992 sur le Dépôt Légal étendu à la Radio et à la Télévision http://legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000723108&categorieLien=id (Consultée le 6 mai 2012)

consommation de vidéos depuis des terminaux variés.[2]Suivant l'évolution rapide des technologies de l'édition, la réglementation du Dépôt Légal en France a été étendue au Web. Il est intéressant de constater que le législateur a jugé nécessaire de partager cette responsabilité entre la BnF, Bibliothèque nationale de France, et l'Ina, assurant ainsi la cohérence et la continuité de leurs fonds respectifs.[3]

L'Ina fut donc désigné comme dépositaire national des sites Web relatifs au secteur de l'audiovisuel - au sens le plus large possible - ainsi que des services de medias audiovisuels à la demande (SMADs). Un décret récent[4] définit les missions et le périmètre de chacune des institutions.
Si la cohérence et la continuité des collections ont en effet prévalu à la rédaction du cadre légal qui a élargi le domaine de compétences de l'Ina, sa mise en œuvre technique et concrète s'éloigne des méthodes, outils et pratiques utilisés pour l'archivage des collections de la radio et de la télévision. Cet article tentera de fournir une vision générale des problématiques liées à la sélection, l'acquisition, l'organisation, l'accès, le stockage et la préservation d'archives Web en relation avec le domaine de la radio et télédiffusion, qui y renvoient ou le documentent.

Sélection

En France comme dans la plupart des pays, c'est un organisme national, le CSA, qui est en charge de la régulation de la communication audiovisuelle, essentiellement pour la radio et la télévision et depuis peu pour les contenus audiovisuels diffusés via le web. Ses responsabilités couvrent notamment l'attribution et la réglementation des fréquences aux opérateurs. Celles-ci sont soumises à autorisation et limitées en nombre. Avant qu'une nouvelle chaîne ne soit autorisée à émettre - ce qui n'est pas si fréquent, à l'inverse des sites Web qui naissent et meurent chaque jour - elle est rendue publique à l'avance et si son archivage incombe à l'Ina, le processus est en général anticipé et se déroule sans heurt. Il en va autrement des sites Web, qui naissent et disparaissent sans préavis. L'agence française d'attribution de domaines (AFNIC) est en mesure de fournir régulièrement une liste des noms de domaine en .fr, mais ceux-ci ne représentent que 30% du Web français et la liste n'est classée, ni par activité de l'éditeur, ni par thème.

La première étape consiste donc à sélectionner les sites pertinents pour l'archivage, à évaluer leur fréquence de mise à jour et leur profondeur de collecte en fonction de leur taille. Parce que le Web est sans limite, volatil et éphémère, définir le périmètre des collections et les frontières des domaines à capturer est fondamental. C'est cependant une activité très chronophage qui repose essentiellement sur l'évaluation et le jugement humains. Les documentalistes de l'Ina s'appuient

---

[2] Des études sur le comportement du public ont montré que la vidéo en ligne connaît une croissance sans précédent. D'après l'IAB, Internet Advertising Bureau, il a fallu 38 ans à la radio pour atteindre 50 millions d'usagers, 13 ans à la télévision, moins de 5 pour internet, et moins de 2 pour la vidéo sur internet. Au Royaume-Uni, 27,3 millions d'usagers sur 38,5 millions qui se sont connectés via leur PC en février 2012 ont visionné de la vidéo. (Source : UKOM/Nielsen, fév. 2012)

[3] Loi DADVSI du 1er Août 2006, Titre IV, extension du Dépôt Légal au web http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000266350 (Consulté le 6 mai 2012)

[4] Décret relatif au Dépôt Légal publié le 19 décembre 2011 http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000025002022 (consulté le 21 juin 2012)

sur les critères énoncés dans le décret pour identifier les sites pertinents et nourrir la liste des sites candidats pour la collecte.

Selon les termes de ce décret, la sélection se fonde sur des critères relatifs à l'activité de l'éditeur du site (activité de radio ou télé diffusion), sur le fait que sa thématique principale est étroitement corrélée à l'un ou l'autre de ces médias (beaucoup de blog et de sites de fans entrent dans cette catégorie), mais aussi qu'il offre majoritairement de la vidéo à la demande ou un accès non linéaire à des programmes (*Replay*, télévision de rattrapage, ou contenu vidéo original pour diffusion web).

La chaîne d'archivage est en œuvre depuis plus de deux ans, en février 2009, 3600 sites Web entraient dans le périmètre de l'Ina, il en compte aujourd'hui plus de 10.000. Il faut noter que l'ambition du Dépôt Légal qui fut de tendre à l'exhaustivité doit évidemment être reconsidérée pour ce qui concerne le Web. Il est difficile de garantir que tout les sites Web pertinents seront identifiés et sélectionnés, et l'approche est plutôt celle du « meilleur effort », grâce à une surveillance active du « Web vivant » par des professionnels experts. Des méthodes fondées sur une analyse sémantique semi-automatique de gros corpus issus du web ont été testées pour accompagner l'intervention humaine mais elles ne se sont pas avérées pertinentes dans le périmètre qui revient à l'Ina.

Nous montrerons plus tard comment cette approche du «meilleur effort» concerne également les phases d'acquisition et capture.

Collecte

A l'époque de l'analogique, l'Ina a collecté des documents audiovisuels fixés sur support physique, suivant les méthodes traditionnelles d'acquisition en bibliothèque. Avec le développement des technologies numériques, dès 2001 un dispositif d'enregistrement en continu depuis les régies de chaînes a permis de collecter à grande échelle les flux de plus de 100 chaînes et d'en organiser l'archivage et stockage à l'Ina.En un sens, les techniques et le dispositif mis en œuvre pour moissonner le contenu du Web s'inscrivent dans cette approche « invasive » de collecte, même si contrairement à la radio et la télévision, le Web n'est pas un flux temporel linéaire et requiert en conséquence le développement de techniques de collecte spécifiques.

Collecter les données et contenus directement depuis les serveurs Web est de fait l'approche générale, l'objectif étant de simuler autant que possible, au sein de chaque page, les interactions humaines, afin de générer un maximum de réponses et d'en télécharger les contenus. Ces techniques communément désignés par les termes anglais *harvesting* (moissonnage) ou *crawling* (rampant) sont utilisées par tous les moteurs de recherche pour collecter et traiter les données issues du Web. Les outils utilisés sont souvent appelés *robots* ou même *spiders* (araignées).

Le mot *spider* peut paraître étrange et désuet mais c'est probablement le terme le plus approprié pour définir cette méthode, puisqu'il évoque les chemins multiples et les carrefours que l'outil automatique va découvrir, suivre, ou ignorer dans sa collecte d'une partie de la toile mondiale (*World Wide Web).*Plus qu'un simple moissonnage il s'agit plutôt d'une technique particulièrement précise et systématique qui, en fonction de paramètres de capture donnés et d'une liste de sites de départs (*seed list*), suit des liens pour découvrir et télécharger des contenus. Tout ceci serait assez simple si la toile n'était pas en mouvement perpétuel, avec des liens qui

apparaissent, ou d'autres qui disparaissent de manière imprévisible, et des éléments, nouveaux ou mis à jour, découverts à chaque nœud. Il est de fait possible, dans une certaine mesure, de prévoir ou d'être informé des changements, par des flux RSS qui servent d'alarme, comme les vibrations sur la toile préviennent l'araignée et signalent que de nouvelles ressources sont apparues sur la toile. Mais cela est plus souvent l'exception que la règle, et l'araignée doit plutôt explorer continuellement la toile pour y découvrir des nourritures attrayantes.

Cette analogie atteint ici ses limites, puisque l'araignée tisse et contrôle sa propre toile, tandis que les robots utilisés pour l'archivage du Web sont invasifs, plutôt semblables à des parasites parcourant une toile gigantesque que des millions d'araignées auraient tissée. De plus, de manière tout à fait inattendue, à l'image des insectes prisonniers des pièges du fourmilion, les robots se trouvent aspirés par des pièges (posés intentionnellement ou pas) entraînant leur chute. Cette métaphore animale trouve ici sa conclusion et nous lui préférerons désormais le terme de « robot ».

Le périmètre du dépôt légal de l'Ina étant précisément délimité, il ne répond pas au mêmes besoins que les collectes larges de domaines, un système de collecte modulable a donc été conçu sur mesure qui s'adapte à la diversité des sites de ce périmètre (fréquence de mise à jour, profondeur, fonctions interactives).

Le système est basé sur une architecture à deux niveaux, avec un ordonnanceur principal qui donne des ordres à une multitude de robots, chacun en charge d'un seul site à la fois. En général, 500 à 1000 robots tournent sur une machine.

L'ordonnanceur (partie supérieure de cette architecture à deux niveaux) s'occupe des aspects de programmation et de configuration de chaque site Web. Grâce à une stratégie d'échantillonnage, il contrôle la fréquence et les rythmes de mise à jour : les sites sont catégorisés (de manière semi-automatique) selon des critères de fréquence (mises à jour constantes, quotidiennes, hebdomadaires, etc.) sur lesquels se fonde la fréquence de collecte.

Le Web de surface (défini par un nombre limité d'interactions ou de « clics » pour y accéder depuis la page d'accueil) étant généralement mis à jour plus souvent que les pages plus « profondes », les pages de surface de chaque site sont collectées plus souvent que les plus profondes.

Certains sites Web ont des flux RSS qui renvoient vers les nouveaux contenus. Ces flux sont utilisés pour lancer des robots spécifiques sur une page, un article ou un contenu donnés dès qu'il sont publiés. De nouvelles visites automatiques sont prévues pour les mises à jour suivantes ainsi que les commentaires.

Cette approche, qui permet l'utilisation simultanée de plusieurs robots, considère séparément, d'un côté les stratégies de programmation (paramètres de fréquence et de profondeur selon les conditions de chaque site) et de l'autre, les problèmes intrinsèquement liés à la collecte (pièges, règles de capture et de netiquette, stockage).

Parce que le Web est une jungle envahie d'une multitude de systèmes et formats hétérogènes, les robots sont toutefois susceptibles de se tromper ou d'échouer au moment où ils s'efforcent de déclencher une interaction.

Cette approche spécifique de robots multiples vise à collecter divers types de contenus suivant la stratégie, précédemment citée, du « meilleur effort » (*best effort* [5],, Elle participe de l'amélioration de la qualité de l'archive. Les robots ont été développés en interne et au format open source, et à l'heure actuelle trois types de robots différents sont utilisés avec le même ordonnanceur, chacun affecté à une tâche précise :

PhagoSite est un robot généraliste capable de traiter des sites Web de taille importante et qui nécessite peu de ressources matérielles.

Fantomas est un robot plus spécialisé basé sur Phantom-JS, qui utilise les mêmes fonctions de base WebKit que les navigateurs Google Chrome ou Apple Safari. Ce robot est capable de capturer la plupart des sites Web « 2.0 » avec des interactions javascript avancées, sans être trop gourmand en ressources matérielles.

Crocket est une extension du navigateur Firefox qui peut capturer des sites Web riches et complexes (riches en média et en interactions). Il nécessite cependant d'importantes ressources matérielles.

Par ailleurs, les contenus vidéos représentant une partie importante de l'archive (tant en nombre qu'en taille), des outils de capture dédiés ont été développés qui téléchargent des vidéos depuis YouTube ou Dailymotion, ou collectent les contenus diffusés en *streaming*.

Ce système de collecte « sur mesure », est en production depuis février 2009, il suit un rythme annuel de 6 milliards de requêtes. Les outils sont sans cesse mis à jour et améliorés afin de s'adapter au monde encore peu mature et en perpétuelle évolution de l'Internet.

Description, métadonnées, accès

A l'inverse de ce qui prévalait à l'ère de l'imprimé, l'accès aux contenus du cinéma, de la radio télévision, ainsi que du Web, repose sur une médiation technologique. Le contenu d'un livre peut être immédiatement accessible, à condition de savoir lire, tandis que les informations consignées sur des supports technologiques tels que film, cassette, disque, ou fichier numérique, nécessitent d'être reconfigurées, calculées, interprétées, pour être rendues intelligibles.

A l'Ina, au vu de la masse des données considérées, tant pour les flux de radio-télédiffusion que pour les contenus issus du Web, les pratiques documentaires ont progressivement évolué et l'extraction et la gestion de métadonnées est devenu la règle.

Cependant, conformément aux approches traditionnelles d'organisation des collections, les sites Web sont décrits et catalogués suivant des taxonomies spécifiques de manière à rapprocher les collections web et celles de la radio et de la télévision.

Le contenu des sites Web archivés est automatiquement indexé pour offrir un accès direct en combinant une URL et une date. C'est de fait le stockage sur disque des fichiers et de leurs métadonnées qui constitue l'archive. De par les caractéristiques de l'archive numérique, qui stocke des unités discrètes, et la nature du Web comme outil de publication, l'accès aux pages

[5] Terme anglais généralement utilisé pour désigner cet état de fait assez transversal dans Web (NDLT)

Web archivées implique une reconstruction du processus de publication qui consiste à accéder à tous les fichiers capturés à un instant T,afin de permettre leur affichage dans une page reconstruite (une estimation récente évalue à 50 le nombre moyen de fichiers qui constituent une page Web[6]).

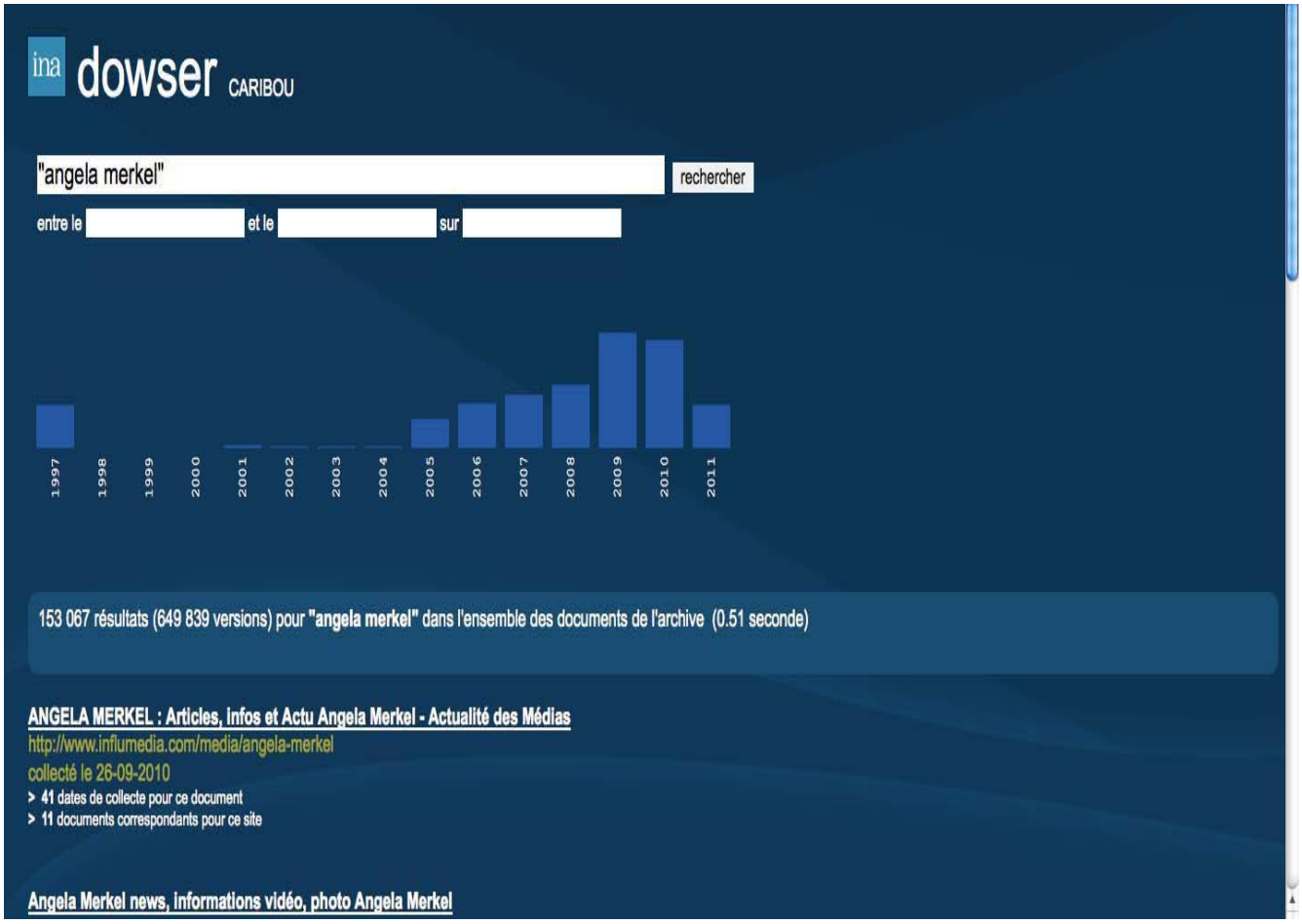
L'outil de navigation dans l'archive est une extension de Firefox. A partir d'une page donnée, il est possible de naviguer en avant ou en arrière dans le temps ou d'afficher les différentes versions existant dans l'archive. La plupart des interactions sont encore actives (la navigation par liens et interactions sommaires représente aujourd'hui 95% de l'expérience d'un utilisateur), certaines ne le sont plus pour des raisons techniques (contenu Flash, interactions complexes avec javascript intégrant un lecteur vidéo).

L'objectif est bien sûr de capturer et d'afficher l'aspect et le fonctionnement des contenus et pages tels qu'ils étaient dans leur contexte de publication d'origine. Ça n'est pas toujours possible en raison d'obstacles ou verrous techniques.Il faut parfois recréer certaines interactions perdues pour que l'expérience de navigation soit complète. Par exemple, une vidéo qui ne peut être visionnée à partir d'un lecteur intégré pourra s'afficher dans un lecteur externe ; de nouvelles fonctionnalités pourront être ajoutées, comme la recherche au sein de la vidéo.

Il est évidemment impossible d'archiver des moteurs de recherche comme Google ou Bing (simuler la totalité des interactions des utilisateurs est impossible), c'est pourquoi un moteur de recherche spécifique a été développé afin de permettre l'interrogation plein texte dans les archives Web de l'Ina. Ce moteur considère la dimension temporelle de l'archive et regroupe les doublons ou quasi-doublons en lots résultats.

Dowser est donc notre moteur de recherche « magique » qui permet des requêtes textuelles « à la Google » au sein de l'archive. Des histogrammes aident l'usager à naviguer entre les dates et les résultats pour chaque requête.

[6] La plupart des sites audiovisuels dépassent allègrement ce chiffre



C’est l’authenticité qui fonde la légitimité d’une archive, mais cette notion est fortement remise en question dans l’environnement numérique. La simple notion de document «original» aura bientôt disparu et les données numériques sont vouées à être copiées, migrées et manipulées. Il est possible d’accéder à des versions successives d’un même site archivé, mais la version canonique d’un site n’existe pas. Le choix de l’Ina est d’informer l’usager de toute altération éventuelle d’un contenu ou d’un contexte original (discontinuité temporelle entre un lien et un site cible, vidéo visionnée dans un lecteur externe, contenus manquants liés aux frontières de la collecte, etc.), en insistant sur le fait que la page Web archivée n’est pas un artefact mais une reconstruction le plus souvent incomplète d’un média original. Cette fois encore, l’approche du « meilleur effort » est privilégiée !

Stockage et préservation

Beaucoup des contenus publiés sur le Web sont en général « nés numériques », et n’ont pas d’existence préalable sur support analogique. Ils constituent des traces et témoignages de notre époque et de nos sociétés et leur préservation à long terme est au coeur des préoccupations de l’archiviste , tout comme les solutions de stockage massif qui permettent la migration des données sur bandes magnétiques ou disques durs et dont les technologies et capacités sont en constante évolution.Si la communauté des archivistes du Web s’accorde sur la nécessité d’un

cadre fiable et robuste pour la préservation à long terme, les développements en sont le plus souvent ralentis au profit des dispositifs d'accès, et ce, afin d'amorcer les usages qui donnent vie aux collections. Nous nous pencherons ici sur les questions qui font débat dans la communauté et rejoignent d'ailleurs celles des archives numériques au sens large.

Supposant que l'intégrité des données encodées est maintenue dans les processus de conservation et préservation, ces données pourront-elles toutefois être réinterprétées correctement dans l'avenir ? De même que l'on doit pouvoir continuer à lire des films ou des cassettes vidéo, la question de la migration des formats de fichier et la maintenance des navigateurs ou des plugins est essentielle.

Intégrité à long terme des données (utilisation de migrations à court terme)

La préservation à long terme des données encodées (*bit preservation*) des fichiers d'archives est une préoccupation partagée par toutes les archives numériques qui s'engagent généralement dans des stratégies de migration à court et moyen terme (stratégies sur 20 ans environ).

A l'Ina, deux copies des archives web sont stockées sur deux générations distinctes de disques, tandis que deux copies de sauvegarde sont stockées surbande.Le dispositif prévoit une migration des fichiers vers de nouveaux disques de stockage tous les 3 à 5 ans, 4 à 5 ans pour les bandes.

Stockage sur disque

Les problèmes liés au stockage sont d'autant plus importants que progressivement, les Kilo-octets des pages de texte formaté (HTML) sont supplantés dans l'archive par un nombre croissant de fichiers audio et vidéo qui se mesurent en Mega-octets. Contrairement à celui du stockage sur bande, le coût du stockage sur disque est très élevé en raison de la consommation en énergie nécessaire pour l'alimentation et le refroidissement du dispositif. Il faut également prendre en compte la fiabilité et le pourcentage élevé de dégradation des disques, qui en limite en général l'usage entre 3 et 5 ans. Pour faciliter la migration et limiter les coûts, les capacités de stockage doivent être doublées lors de chaque migration (ainsi, les disques de 1,5 To seront remplacés par d'autres de 3 To)

LTO (Linear Tape-Open)

La technologie LTO de stockage de données sur des bandes magnétiques a été développée à la fin des années 90 sous la forme d'un standard ouvert, se distinguant du format propriétaire des bandes magnétiques disponibles à l'époque. A partir de 2000, le standard LTO définit une taille standard de cartouche avec une garantie d'augmentation de ses capacités sur 8 générations - soit une période de près de 20 ans (2000 à 2017) - cette capacité doublant pour chaque nouvelle génération, les stratégies de migration s'en trouvent simplifiées et permettent de normaliser le stockage physique.

Même si en théorie les bandes magnétiques peuvent se conserver jusqu'à 25 ans, rien ne garantit que les fournisseurs de LTO aient dans le futur des lecteurs capables de lire de vieilles versions des bandes. Les spécifications recommandent d'ailleurs, que chaque version des LTO soit compatible avec les 2 générations précédentes (par exemple, les lecteurs LTO-4 doivent pouvoir lire les bandes LTO-2). Ainsi, une migration fiable des données repose sur la disponibilité d'un

nouveau format économiquement viable. A titre d'exemple, même si le lecteur LTO-5 était en vente au 2ème trimestre 2010, il a fallu attendre le 2ème trimestre 2012 pour un prix de cartouche avantageux par rapport à la version LTO-4. De même, une optimisation des coûts et du ratio rendement/espace sera considérée au moment choisi pour la migration d'une génération à l'autre de LTO afin de multiplier par quatre les capacités de stockage, soit une migration de 4 cartouches LTO-4 vers 1 cartouche LTO-6 par exemple.

Checksums

Après la création d'un fichier d'archive, son contenu est vérifié à l'aide de divers outils. Si le contenu du fichier n'est pas corrompu, une clé *checksum* (SHA) du fichier est stockée dans une base de données séparée. Sur bande magnétique, les fichiers sont aussi stockés avec les *checksums* qui leur correspondent. Le contenu et la validité des fichiers peuvent ainsi être régulièrement vérifiés, et si besoin, une copie de sauvegarde corrompue pourra être remplacée.

Préservation à long terme : émulation ou migration ?

L'archive du Web doit être considérée selon de deux points de vue :

- Ce que l'on appelle le *look and feel* des pages Web – la dimension «esthétique » qui exprime le style, l'imagination, et le talent du créateur-

- Les différentes composantes de la page ; texte, images, sons et vidéos

La première dimension se prête plutôt à des stratégies de type émulation. Un grand nombre d'institutions s'engagent dans des projets de classification et émulation de vieux navigateurs et plugins. Une manière de vérifier que l'émulation fonctionne correctement peut aussi être envisagée en comparant avec une vidéo ou une capture d'écran des pages web d'origine.

Le deuxième aspect se prête davantage à une stratégie de type migration. Au moment de la capture (ou dans les années qui suivent) un fichier de format standard peut faire l'objet d'une migration à moyen terme (disons 20 ans) des données.

Selon l'exemple ci-dessus, les formats PDF ou ASCII pourraient être choisis pour le texte, JPEG2000 pour les images, AIFF pour le son, et MPEG4 H264 pour tous les formats vidéo ;lLes derniers étant actuellement les formats standards utilisés à l'Ina pour la migration à long terme des fonds audio et vidéo. Cette approche peut également s'appliquer à l'extraction de métadonnées, un certain nombre de logiciels pouvant être utilisés pour identifier fichiers et formats (DROID, JHOVE, ou Apache Tika par exemple) afin d'organiser le stockage des résultats dans des formats standardisés.

Au bout du compte, l'objectif est de consigner le plus d'informations possible relatives à une page Web au moment de sa capture. Même avec les meilleures stratégies de migration et/ou d'émulation, il est impossible de garantir qu'une page s'affichera exactement telle qu'elle a été initialement archivée. Il sera probablement possible de restaurer les aspects visuels de la page, mais dans la plupart des cas, il ne sera pas possible d'obtenir exactement les mêmes modalités d'interaction.

Cependant, une combinaison des deux approches pourra permettre à l'usager futur de recourir à divers outils. Celui-ci pourra ainsi consulter une page créée par un émulateur sans pouvoir lire un document associé dans la page reconstruite. Il sera toutefois possible de comparer cette émulation à une capture d'écran de la page originale afin d'avoir un aperçu de sa forme visuelle, puis de prendre ensuite les métadonnées extraites (migrées) pour s'efforcer de récupérer le contenu du document original.

Conclusion

Il est vraiment difficile d'évaluer l'impact à long terme du stockage numérique de tous les savoirs et productions, mais archives, institutions patrimoniales et universitaires s'efforcent aujourd'hui d'inscrire leurs stratégies de préservation dans le temps long. Dans cette perspective, les archivistes du Web ont un train d'avance et le mérite d'archiver des contenus qui ont espérance de vie très courte sur le web vivant. Ils suivent et s'adaptent à l'évolution de formats nombreux et éphémères, s'attachent au maintien de l'interactivité au sein des pages archivées, et s'efforcent de reproduire « l'expérience utilisateur » originelle. C'est désormais une activité partagée, et même si les approches diffèrent d'une institution à l'autre, leur coopération au sein de l'IIPC encourage le dialogue, les échanges de pratiques, en impliquant des usagers -- chercheurs ou professionnels -- de nombreux pays afin que les générations à venir.puissent écrire l'histoire du World Wide Web et de son contenu.