ICHIM
PARIS 21-23 SEPT. 05
www.ichim.org

Digital Culture & Heritage
Patrimoine & Culture Numérique

Bibliothèque nationale de France, PARIS
Sept. 21st - 23rd, 2005
21 - 23 septembre 2005

DOCUMENTER ET PARTITIONNER UNE ARCHIVE DU WEB: VERS LE DEPOT LEGAL D'UN DOMAINE MEDIA

Bruno Bachimont, Thomas Drugeon, Geneviève Piéjut
http://www.ina.fr

Published with the sponsorship of the French Ministry of Culture and Communication

Actes publiés avec le soutien de la Mission de la Recherche et de la Technologie du Ministère de la Culture et de la Communication, France

Interprétation simultanée du colloque et traduction des actes réalisées avec le soutien de l'Agence Intergouvernementale de la Francophonie

Abstract (EN)

It is now widely acknowledged that the Web has given rise to radical new forms of contents and communication between people. A genuine culture is currently emerging and for which questions of memory and heritage should be answered. Since a few years, the archiving of the Web has been recognized to be a very crucial issue. Several initiatives, at national or international levels, are launched and illustrate the multiple ways of dealing with such an archiving, according to the context and constraints. After showing why the Web remains a complex object, this article deals with the different strategies that exist in order to archive the Web, presents the paramounts initiatives, and concludes on the french case by paying a special attention to the approach adopted by the french Institut National de l'Audiovisuel with respect to its targeted domain.

Keywords: Web archiving, legal deposit, French Web.

Résumé (FR)

Le Web est désormais compris et reconnu comme un lieu original où une culture propre se construit, que ce soit par les modes de communication entre les personnes ou par les types de contenus élaborés. Or, toute culture pose le problème de sa mémoire et de son patrimoine, ainsi la question de l'archivage du Web est devenue prégnante ces dernières années et s'est traduit par de nombreuses initiatives nationales ou internationales. Se heurtant aux difficultés propres au Web, elles adoptent des stratégies parfois fort différentes selon le contexte. Cet article fait le point sur l'enjeu de l'archivage du Web, précise la situation internationale et conclut sur le cas de la France, en s'intéressant particulièrement à l'approche adoptée par l'INA sur le domaine qui devrait lui être dévolu.

Mots clés : Archivage du Web, dépôt légal, Web français, Institut National de l'Audiovisuel.

Introduction

Le Web a souvent été perçu comme un support de diffusion et de transmission supplémentaire apportant de nouveaux moyens de communication, mais sans pour autant déterminer la forme ni le contenu des objets transmis. Ainsi, un article scientifique formaté selon les règles de la discipline et selon les prescriptions des revues du domaine trouvera-t-il une diffusion facilitée, mais sa forme reste déterminée par des contraintes extérieures au Web. Cependant, il est rapidement apparu que le Web est également un moyen d'expression, renouvelant à la fois les modes de communication entre les individus et les types de contenus échangés. Le Web devient alors un médium à part entière, suscitant sa propre forme éditoriale et communicationnelle. A ce titre, le Web constitue un univers culturel autonome dont il convient d'assurer l'organisation et la gestion, en particulier sur le plan de la mémoire et du patrimoine. A l'instar de supports plus anciens dont la maturité s'est accompagnée d'une politique patrimoniale de la mémoire, le Web devient progressivement un lieu où ces questions se posent. De fait, de nombreuses institutions, publiques pour la plupart mais pas exclusivement, envisagent un archivage de toute une partie du Web pour répondre à ces préoccupations. La France s'associe à ce mouvement en envisageant un dépôt légal du Web dont la Bibliothèque Nationale de France (BNF) et l'Institut National de l'Audiovisuel (INA) serait en charge. Cet article présente la problématique de l'archivage du Web, les initiatives en cours avec les différentes méthodes et approches possibles pour conclure sur le cas de la France, en s'intéressant particulièrement à l'approche adoptée par l'INA sur le domaine qui devrait lui être dévolu.

I. Pourquoi Archiver le Web

L'archivage du Web est de plus en plus perçu comme une nécessité, concernant la mémoire et la culture collectives. Cela tient à deux raisons essentielles :

- C'est une forme inédite de production, diffusion et consultation de contenus : on y trouve des contenus qui n'ont pas leur équivalent ailleurs ;
- Ces contenus possèdent un intérêt culturel et patrimonial.

Ne pas les conserver en en gardant une trace, même sélective, correspond à une négligence patrimoniale. Ce fait est d'ailleurs largement reconnu et les initiatives d'archivage fleurissent. Il est donc important d'avoir une politique de mémoire du Web pour se situer dans ce paysage déjà dense. Ce qui donne lieu à un troisième type de raison :

- Le contrôle de la mémoire est un moyen de domination culturelle, politique et économique. Ne pas avoir une politique de mémoire du Web, c'est laisser à d'autres le contrôle (la sélection des contenus et le contrôle des accès) de sa propre identité culturelle.

Le Web est une forme originale d'élaboration et diffusion de contenus. En effet le Web n'est pas un simple support de diffusion et de distribution. C'est un support *d'élaboration*, donnant lieu à de nouveaux contenus qu'on ne trouve pas ailleurs. Le Web ne prolonge donc pas seulement des collections existantes, mais inaugure de nouveaux contenus dans une forme éditoriale également nouvelle (inédite !). Ces nouveaux contenus possèdent ainsi une forme spécifique au Web qui ne se retrouve sur aucun autre support ; en particulier, ils sont à la fois interactifs et interconnectés. Leur logique d'élaboration et de consultation repose sur ces deux propriétés :

- Interactivité :

La plupart des contenus reposent sur une activité de l'internaute : leur nature éditoriale ne peut alors être conservée que sur un support permettant l'interactivité.

- Connectivité :

Les contenus sont reliés par des liens ; ces liens font partie du contenu et le constituent, il en est indissociable. Ces liens induisent un mode de consultation et d'appropriation particulier : navigation bien sûr, mais aussi reprise et réutilisation (blog), collaboration (wiki), syndication (RSS). Autrement dit, un contenu du Web ne peut se consulter que dans son contexte de connectivité. L'extraire de ce contexte pourrait s'apparenter au fait de regarder un film sans sa bande son, de lire un livre sans ses notes de bas de page, un article d'encyclopédie sans ses renvois : il en résulte bien un contenu, mais pas le contenu dans son intégrité originale.

Par ailleurs, le Web possède un intérêt culturel intrinsèque. L'interactivité et la connectivité donnent lieu à des formes culturelles nouvelles : forums, wiki, blogs, où les contenus circulent, s'agrègent et se propagent. On parle aujourd'hui de « liquidité » des connaissances, dans la mesure où elles s'écoulent sur le Web, par opposition aux autres médias qui ont plutôt tendance à

fixer les contenus. Le Web permet comme média et forme de communication de constituer ou d'animer des formes diverses de sociabilité. En outre, de nombreux sites possèdent une interactivité aboutie, en particulier les sites de création artistique et littéraire. Autrement dit, le Web n'a pas seulement un intérêt culturel au sens sociologique du terme (témoignage de la pensée et du mode de vie d'une époque), mais au sens des beaux-arts et de la culture savante.

Mais cette vie sociale et ces contenus s'inscrivent dans le temps court et constituent par nature un patrimoine en danger : média de circulation, le Web permet d'élaborer des contenus innovants pour l'interactivité et la communication où le propre d'un contenu est d'être échangé et transformé, non d'être conservé. La logique du Web n'inclut pas la persistance ni la pérennité, mais l'échange et la reprise. Si bien que le Web est très volatile : la durée de vie moyenne d'une page est estimée être inférieure à 2 mois (Lawrence 2001). Colin Webb (Webb 2001), des archives nationales australiennes, souligne ainsi que les sites Web associés aux jeux Olympiques de Sydney en 2000 ont plus vite disparu que les athlètes eux-mêmes. Média de la circulation, le Web n'a pas de mémoire.

Ce constat est désormais banal : face à un intérêt sociétal et éditorial grandissant et à une volatilité sans cesse constatée, il est nécessaire d'avoir des initiatives volontaristes de préservation de la mémoire et de constitution d'un patrimoine. Aussi les initiatives patrimoniales sont–elles multiples et de plus en plus relayés à différents niveaux. Ainsi l'UNESCO a adopté en 2003 une charte sur la conservation du patrimoine numérique. Les contenus numériques sont reconnus désormais comme possédant une *valeur*, valeur patrimoniale et culturelle. On relève également des initiatives publiques à des échelles nationales, notamment dans le cadre de dispositif de dépôt légal sur lesquels nous revenons plus bas. Enfin, des initiatives privées exemplaires et spectaculaires, dont la plus marquante est sans conteste InternetArchive, fondation américaine privée, qui archive le Web mondial. Ce projet est d'autant plus remarquable qu'à ce jour on ne compte aucun projet international émanant d'instances publiques.

II. Les difficultés propres au Web…

S'il paraît nécessaire d'envisager un archivage du Web du fait de son importance culturelle et sociétale d'une part, et de la volatilité de ses contenus d'autre part, la mise en œuvre effective

d'une telle entreprise paraît d'une complexité redoutable, au point de remettre en cause le principe même d'un archivage. Aussi les politiques adoptées en la matière renvoient à des choix politiques et des conceptions théoriques permettant d'aborder tant la masse que la complexité des contenus du Web. Tout d'abord, avant d'envisager la manière dont les différents projets se sont constitués, il est bon de rappeler en quoi le Web est un objet complexe, et en quoi il est particulièrement rebelle à une logique d'archivage et de documentation.
Le Web est un média de convergence : on retrouve toutes les formes de média dans les contenus du Web. De fait, la plupart des difficultés rencontrées avec les différents médias se présentent par conséquent dans le contexte du Web.

1. Des contenus hétérogènes…

L'archivage de la télévision amenait son lot de contraintes techniques, de supports et de formats, mais il s'agissait d'un medium mature, relativement stable et aux normes bien établies (PAL/SECAM, Beta SP, MPEG-2, *etc*.). La situation du Web est toute autre. Les technologies mises en œuvres évoluent sans cesse, tant il est vital pour beaucoup de sites de présenter des contenus toujours plus attractifs et adaptés aux technologies du moment. L'immatérialité de ces technologies facilite leur développement : là ou un changement de norme en diffusion télévisée nécessite la mise en place de chaînes de traitements spécifiques et un investissement matériel pour l'utilisateur[1], l'apparition d'une nouvelle norme sur le Web ne requiert souvent que l'installation ou la mise à jour d'un composant logiciel spécifique (*plugin* ou *codec*). La structure même du Web rend ces mises à jours aisées, voire parfois transparentes pour l'utilisateur, exacerbant ainsi la prolifération de nouveaux formats éphémères et non normés, pouvant émaner de n'importe quel acteur du domaine. Constituer l'archive pérenne d'un Web en perpétuelle évolution technologique est l'un des enjeux majeurs de ce Dépôt Légal : comment collecter, indexer, conserver et préparer la future consultation de ces contenus ?

[1] La mise en place de la télévision numérique terrestre (TNT) nécessitera par exemple pour l'utilisateur l'achat d'un décodeur adapté ou d'un nouveau téléviseur.

2. … dans une structure interactive

Les contraintes techniques ne se situent pas uniquement dans les contenus, mais aussi dans la manière dont ils sont structurés au sein d'un site. Le Web est construit sur une logique d'interaction, l'utilisateur sollicitant lui-même l'affichage d'une page en cliquant sur un hyperlien ou en remplissant les champs d'un formulaire. La consultation d'un site Web suit ainsi en permanence une logique de question/réponse dans laquelle chaque contenu envoyé par le site (page, film, image) fait écho à une requête de l'utilisateur. Le principe de collecte des sites Web diffère ainsi de celui d'une collecte radio ou télévision où l'archivage peut être assimilé à la captation passive d'un flux prédéfini, puisqu'il faut ici reproduire les interactions d'un utilisateur pour accéder aux contenus. La collecte complète d'un site Web à un moment donné prend ainsi forme par la copie de l'ensemble des réponses reçues à chacune des interactions possibles. La reproduction systématique et contrôlée de ces interactions par un système automatique constitue l'enjeu principal de la phase de collecte : une interaction manquante ou tronquée peut rendre inaccessible à la copie une partie importante du site, alors qu'un manque de contrôle dans les interactions peut engendrer la copie en masse d'informations redondantes ou inutiles.

3. Une unité documentaire à définir

La définition du concept de document et sa délimitation est le préalable obligatoire au travail d'indexation de l'archive, puisqu'il faudra conserver les contenus archivés, les structurer et les organiser pour permettre de les retrouver et de les re-exploiter.

Dans le contexte du Web, ce préalable recèle des difficultés importantes. Traditionnellement, un document se caractérise comme étant la fixation d'un contenu sur un support matériel (livre, feuillet, cassette) ; la permanence du support assure au contenu sa conservation de même que les délimitations du support déterminent l'unité du contenu.

Dans le contexte numérique du Web, les contenus ne sont pas véritablement fixés ou enregistrés sur un support, mais inscrits dans un codage informationnel qui n'a pas de lien avec un support particulier. Cela se traduit par conséquent par une volatilité forte des contenus : révisions innombrables des pages, mises à jour indépendantes de différentes parties d'un site, etc. Il devient

alors difficile non seulement d'accéder à l'information avant sa révision, mais surtout de savoir à quel état, quelle identité du contenu on s'intéresse.

Par ailleurs, les contenus sont difficiles à délimiter et à cerner, compromettant ainsi la définition documentaire d'une archive. On peut cependant distinguer deux logiques documentaires.

Selon la première, des éditeurs et producteurs de contenu élaborent des documents suivant les critères traditionnels du monde de la presse, des éditeurs ou des diffuseurs de média. Ces contenus sont alors mis en ligne et l'archivage du Web peut s'appuyer sur la cohérence documentaire des contenus, antérieure à leur mise en ligne, pour les collecter et organiser leur conservation.

Selon la deuxième approche, les contenus sont produits et mis en ligne en suivant la logique propre du Web de circulation d'information dans un réseau. Il devient alors difficile de repérer les limites d'un document, le support devenant immatériel et se confondant avec les sites en réseau. L'unité d'un document sur le Web résulte en effet le plus souvent du parcours de l'utilisateur, ce parcours pouvant aussi bien être linéaire (chaque page représentant par exemple le chapitre d'un récit que l'on parcourt), que circulaire (*Web Rings*), hiérarchique (structure arborescente), ou même chaotique (flânerie au hasard des pages). Certes, la logique du parcours est plus ou moins établie par le créateur du site qui, suivant sa logique éditoriale, constitue un document composite. Mais, outre le fait que cette structure peut énormément différer d'un site à l'autre, cette structure préétablie éclate le plus souvent quand l'utilisateur parvient à une page particulière du site en suivant un hyperlien depuis un autre site ou un moteur de recherche, ou lorsqu'il sauvegarde l'adresse de la page pour y revenir sans devoir à nouveau parcourir le site (*bookmark*). Il échappe ainsi aux logiques éditoriales gouvernant les mises en ligne pour créer son propre document, fruit de son parcours. On constate d'ailleurs que cette logique éditoriale disparaît comme telle, bon nombre de sites comme les blogs, les RSS (syndication), les wikis reposant sur la logique d'échange et de circulation sans qu'il y ait nécessairement une régulation explicite.

Tout semble indiquer que l'unité documentaire ne peut être fixée à l'avance. La structure de l'archive doit donc permettre de documenter et de consulter ces documents selon différentes échelles d'analyses (page, partie de site, site, groupement de sites, *etc.*). Autrement dit, l'archive du Web doit collecter des contenus pour en faire des documents, selon une logique de *documentarisation* : de la même manière que l'INA fait de la diffusion radio et télévision une

archive documentaire qu'il structure en collection, programmes, séquences, l'archivage du Web devra déterminer et expliciter les points de vue documentaires permettant de constituer l'archive et de l'exploiter. L'archivage n'est pas une conservation, mais une constitution.

III. Les différentes initiatives d'archivage du Web

Les initiatives autour de l'archivage du Web sont nombreuses (Day 2003). On peut les catégoriser selon trois axes :

- Type de l'acteur assurant l'archivage ;
- Type d'archivage assuré par l'acteur.
- Méthode d'archivage.

1. Les acteurs

Les acteurs peuvent être regroupés en trois catégories essentielles :

- Les *organisations particulières* assurant la publication et mise en ligne de contenus ; acteurs du Web elles assurent elles-mêmes l'archivage des données qui les concernent (leur domaine d'intervention) ou qui leur incombent (ce qu'elles publient). Par exemple, la BBC assurera l'archivage de son site d'information.
- Les *institutions publiques de niveau national* qui assurent l'archivage et la conservation des publications Web considérées dans le domaine national (par exemple, les sites .fr, en français, ou dont le serveur est domicilié en France etc.).
- Les *organisations privées agissant sur un plan international* et qui assurent un archivage du Web. C'est par exemple le cas de la *fondation* américaine (Etats-Unis) privée, Internet Archive, qui assure un archivage du Web mondial. D'autres initiatives émergent, notamment autour de Google, *société commerciale*, qui entreprend de numériser des bibliothèques pour offrir leur consultation accessible en ligne. Cette démarche, ne relevant pas directement de l'archivage du Web, montre cependant l'intervention d'initiative

privée dans des questions relevant traditionnellement de la sphère publique, voire des privilèges régaliens des États (politique culturelle de la mémoire).

2. Le type d'archivage

L'archivage s'aborde à partir d'une politique générale déterminant le périmètre concerné. Deux types de périmètre sont à distinguer :

- Le Web global ;
- Un périmètre donné, restreignant le Web global à un segment ou une portion strictement inférieure. Ce périmètre peut se définir selon ces critères :
    - Linguistiques : les sites de langue française, de langue suédoise, etc. ;
    - Territoriaux : les sites en .fr ;
    - Thématiques : les sites se rapportant à la médecine (par exemple www.cismef.org, projet UKWAC cf. *infra*), à la migration, etc. ;
    - Événementiels : les sites se rapportant à un événement particulier et ne survivant pas à cet événement : les jeux olympiques, un mariage royal, etc.

Quand un périmètre est fixé, on peut retenir différentes approches :

- *L'approche exhaustive* : l'archivage porte sur une aspiration/récupération aussi complète et systématique que possible de tous les sites du périmètre retenu. L'intérêt de cette approche est ne pas à avoir à sélectionner en amont. L'inconvénient est d'avoir une qualité très hétérogène de l'archive résultante, la correction manuelle de l'aspiration étant impossible vu la masse. L'exhaustivité est difficile à atteindre du fait du Web caché et du rythme des mises à jours. Cette approche est coûteuse en termes techniques (stockage, crawler, indexation) et économique en termes de personnels (processus très automatisés).
- *L'approche sélective* : l'archivage porte sur des sites sélectionnés *a priori* dans le périmètre retenu selon des critères définis par l'acteur assurant l'archivage. L'objectif est de ne retenir que ce qui paraît pertinent, et de bien le faire. Un premier intérêt est de ne pas s'encombrer des innombrables données sans intérêt (en termes de qualité – c'est très mauvais, faux, etc., et en termes de notoriété – ça n'intéresse personne). Un second intérêt est de pouvoir affiner manuellement les outils automatiques d'aspiration des sites pour

être sûr de les avoir récupérés de manière complète et adéquate. L'inconvénient est qu'il faut déterminer en amont l'intérêt et la pertinence des contenus. Pour un archivage de données médicales, les médecines alternatives peuvent être écartées, alors qu'elles intéresseront l'historien, l'anthropologue, le pharmacologue à la recherche de nouveaux principes actifs. Cette approche est coûteuse en termes de personnels, mais plus économique en termes techniques.

- *L'approche par échantillonnage :* l'archivage porte sur la récupération de quelques sites jugés particulièrement représentatifs d'un type de contenus. Ce sera par exemple l'archivage d'un site de création littéraire pour témoigner de la manière dont les techniques interactives du Web ont fécondé la création artistique. L'approche par échantillonnage est une espèce d'approche sélective. Mais, alors que l'approche sélective retient tous les sites qu'on a repérés comme étant importants, l'approche par échantillonnage n'en retient que quelques uns à titre d'exemple ou d'illustration.

Enfin la procédure d'archivage oscille entre ces deux modèles :

- *L'aspiration automatique*; effectuée par des robots, elle peut être continue (chaque site est aspiré en fonction de son propre rythme de mise à jour) ou discrète (on récupère une photographie du domaine à intervalle régulier, 4 fois par an par exemple).
- *Le dépôt manuel* par les éditeurs de site eux-mêmes des contenus, par voie électronique par exemple (transfert ftp des fichiers). Cette approche semble pertinente pour le « deep Web » ou « Web caché », correspondant aux bases de données ou de contenus utilisées pour publier dynamiquement des contenus. En effet, ces bases ne peuvent pas être récupérées par aspiration.

Ces différentes méthodes peuvent être librement combinées.

3. Les initiatives dans le monde

Muni de ces catégories, on peut tenter de repérer et situer les différentes initiatives. Deux constats préalables concernant les projets nationaux d'inspiration publique :

- Tout d'abord, il semble qu'aucune institution ne vise un archivage global d'Internet, à l'exception notable d'Internet Archive. La raison est que les différents acteurs potentiels ne veulent refaire ce que fait déjà Internet Archive, ce qui reviendrait à avoir une concurrence coûteuse et inutile. Autrement dit, chacun se concentre sur un périmètre donné.
- Les approches et méthodes choisies dépendent largement du cadre juridique. En l'absence d'un dépôt légal, on a plutôt une approche sélective reposant sur des autorisations préalables ; avec un dépôt légal, on adopte plutôt une approche exhaustive.

On distingue alors les initiatives nationales que l'on peut regrouper ainsi.

A) Les approches sélectives

- PANDORA (Australie)

Il n'y a pas de dépôt légal pour le Web en Australie. Aussi l'approche est-elle sélective, et repose sur une autorisation préalable des détenteurs de site. L'archivage a commencé dès 1997. En Août 2004, elle comptait environ 6000 sites, comprenant 13 000 pages, 21 millions de fichiers et 700 Goctets. L'approche sélective a permis de s'y prendre à l'avance, avant que tous les outils soient prêts, et est plus qualitative dans le sens où l'on peut vérifier si l'archivage a été fait correctement afin d'être utile. La National Library of Australia estime que près de 6% des contenus archivés ont désormais disparu du Web. PANDORA a mis au point un outil d'archivage collaboratif, PANDAS, qui doit être repris par le projet anglais, UKWAC.

- MINERVA (US)

C'est un projet de la Library of Congress américaine ayant porté sur l'archivage des sites en rapport avec les élections. Accessible uniquement en interne, cette archive ne porte que sur certains sites prédéfinis. Une collaboration s'est cependant faite avec Internet Archive pour avoir enrichir ce fonds.

- UKWAC (Grande-Bretagne, UKWAC UK Web Archiving Consortium http://www.webarchive.org.uk/index.html)

Projet récent, de 2 ans, lancé en juin 2004, et mené par la British Library, avec The National Archives, National Library of Wales, National Library of Scotland, JISC, et Wellcome Trust. La Joint Information Systems Committee of the Higher and Further Education Funding Councils (JISC) finance des projets donnant lieu à de nombreux sites. La JISC veut assurer leur préservation au delà de la durée des projets. La Wellcome Trust est une fondation pour la recherche médicale qui veut assurer une mémoire de la pensée médicale, académique mais aussi alternative. Ce projet utilise l'outil PANDAS du projet PANDORA. L'objectif est d'archiver 6000 sites.

L'approche sélective a été choisie notamment parce qu'il n'y a pas de dépôt légal concernant le Web, si bien que pour échapper aux manquements aux lois sur le copyright, des autorisations préalables à l'archivage sont nécessaires.

B) Les approches exhaustives

Ces approches se caractérisent par le fait qu'elles s'abritent sous une loi de dépôt légal qui les autorise à capter les contenus sans autorisation préalable.

- KULTURARW (Suède)

Projet (Arvidson et al., 2001) lancé dès 1996 par la bibliothèque Royale de Suède. Elle effectue des photographies à intervalles plus ou moins réguliers. En octobre 2003, 10 passes avaient été faites, 185 millions de fichiers collectés pour 5,5 Téraoctets de données. Sur le plan juridique, un décret de mai 2002 autorise explicitement la bibliothèque royale à faire des captations, et des préservations de contenus du Web ; la consultation est ouverte au public dans le cadre de la bibliothèque. Avant que les travaux se soient effectués et que le cadre légal permette de réaliser des captations, la consultation restant interdite au public.

- AOLA (Autriche)

Ce projet (Rauber et al., 2002), *Austrian On-Line Archive* (AOLA), est une initiative de la bibliothèque nationale autrichienne (Österreichische Nationalbibliothek) et du département de

technologies logicielles de l'université de technologie de Vienne. L'approche repose sur une captation des sites .at, et des sites d'intérêt pour l'Autriche. Après des essais infructueux, une archive de 488 Go a été réunie au printemps 2002.

C) Les approches mixtes

- La **Bibliothèque Nationale de France** adopte une approche mixte combinant une aspiration systématique reposant d'une part sur un balayage exhaustif des sites et une captation des sites possédant la plus forte notoriété, et d'autre part sur un dépôt de sites de référence, appartenant au Web caché, car consistant en bases de données massives et volumineuses (Masanès 2002).
- **NetArchive.dk** (Danemark, http://netarchive.dk)

  Un projet pilote a été réalisé entre 2001 et 2002. Il repose sur une triple approche : photographies par captation automatique du Web danois (SnapShot Harvesting), captation sélective de sites de référence (Selective Haversting), captation sélective de sites rattachés à un événement (Event-driven harvesting). Ce projet profite d'une modification en 1997 de la loi sur le dépôt légal, qui inclut les contenus Internet dans le périmètre du DL. Cependant, seuls les contenus « statiques » semblent concernés. La dynamique restant hors périmètre.

  Le projet se poursuit, avec la Royal Library qui se concentre sur la collecte automatique, et la State and University Library qui se concentre sur la collecte sélective et la collecte événementielle. Parallèlement, un travail législatif est en cours.

  Le 16 décembre 2004 a été votée une nouvelle loi sur le dépôt légal prévoyant la captation du Web, pour une prise d'effet le 1 juillet 2005. Cette loi reprend les dispositions anciennes sur les films, les documents publiés sous une forme physique et impose un dépôt légal radio et télé ainsi qu'Internet (pris en sens large : communication électronique, filaire ou sans fil). La loi reprend les propositions issues du projet :

  - Une collecte exhaustive, avec trois collectes incrémentales, par an ;
  - Une collecte sélective sur 70-100 sites, collectées plus fréquemment ;
  - Une collecte événementielle 2-3 fois par an.

Les sites collectés de manière sélective doivent couvrir les sites représentatifs des médias, les sites les plus utilisés, les sites présentant un caractère expérimental. Un comité, sous la responsabilité du ministère de la culture, effectuera la sélection (collecte sélective et événementielle).

D) Les initiatives internationales

- Networked European Deposit Library (NEDLIB www.kb.nl/coop/**nedlib**/)
  Programme (Van der Werf-Davelaar 1999) qui a été soutenu par la Commission Européenne de 1998 à 2001. Il a regroupé huit bibliothèques nationales européennes (Pays-Bas, Norvège, Allemagne, Portugal, Suisse, France, la Bibliothèque de Florence, la Bibliothèque universitaire d'Helsinki). Sont également impliquées les sociétés d'édition Elsevier Science, Kluwer Academic et Srpinger Verlag. Le programme avait pour objectif la constitution d'un modèle de spécifications fonctionnelles et techniques relatif aux documents électroniques publiés sur support ou diffusés sur le web.

- Nordic Web Archive (NWA)
  Projet de collaboration internationale entre les bibliothèques nationales scandinaves, commencé en septembre 2000 et terminé en juin 2002, poursuivi par NWA II en mars 2003. Le projet a permis de développer ou adapter des outils de captation, indexation et conservation, que NWA II doit rendre utilisables par tous les acteurs du Web.

- International Internet Preservation Consortium (IIPC, http://netpreserve.org)
  Consortium regroupant des bibliothèques de référence, mené par la BNF :
  - Library of Congress,
  - National library of Italy (Florence),
  - Royal Library of Denmark,
  - The National Library of Norway ,
  - National and University Library of Iceland,
  - Library and Archives Canada, National Library of Australia
  - The British Library,

- The Royal Library, National Library of Sweden,
- Helsinki University Library, The National Library of Finland,
- Internet Archive.

Son but est d'échanger sur les outils et méthodes, et de construire et proposer à la communauté les référentiels, guides de bonnes pratiques, et outils nécessaires à l'archivage du Web.

4. Conclusion sur le positionnement international

La plupart des initiatives adoptent un positionnement « classique » sur l'archivage : le Web contient des contenus qu'il faut aller chercher et récupérer. On aborde donc le Web comme un support parmi d'autres de fixation de contenu : on récupère les imprimés, les vidéogrammes, les films, etc. On ne prend pas assez en compte le fait que le Web ne soit pas uniquement une affaire de *contenus fixés sur un support*, mais de *contenus circulant entre des sites.*

Il est donc possible que ces projets soient amenés à sensiblement évoluer dans les prochaines années pour répondre aux demandes des communautés présentes sur le Web, constitutivement sensibilisés à la nature éditoriale spécifique du Web : circulation des contenus grâce à connectivité, élaboration statique ou dynamique des contenus grâce à l'interactivité.

IV. Le dépôt légal du Web en France

La France possède une longue tradition en terme de dépôt légal, qui remonte au décret de Montpellier (1537, sous François 1er) pour les imprimés. Ce dépôt légal s'est progressivement étendu pour prendre en compte les nouvelles formes de publication et les nouveaux supports. A présent, trois institutions sont responsables du dépôt légal en France qui se partagent ainsi la tâche (très grossièrement) : la BNF est en charge des imprimés et plus généralement des contenus fixés sur des supports, le CNC est en charge des films, et enfin l'INA est en charge des flux radio-télévisions. La France s'oriente vers la mise en place d'un dépôt légal pour le Web dont la charge serait assumée par la BNF et la l'INA, en distinguant deux périmètres d'intervention distincts. En effet, le principe retenu est de confier à l'INA l'archivage des sites relevant de la

communication audiovisuelle pour confier à la BNF l'archivage des autres sites. Dans ce contexte, il convient de déterminer à partir de ce principe les périmètres respectifs de chaque archivage. Ainsi, pour l'INA, il faut déterminer le domaine « média ».

L'INA a entrepris une démarche expérimentale pour déterminer ce périmètre. Cette démarche suit une approche de type « sous-réseau » : on recherche, en utilisant les liens qui les relient, les sites qui ont le plus à voir ensemble, en partant d'un ensemble *a priori* de sites reconnus appartenir au domaine. En effet :

- Le Web est structuré : tout site ne pointe pas sur tous les autres sites. Le Web se répartit en sous-réseaux ou communautés qui constituent chacune un tout cohérent. Pour conserver l'environnement pertinent d'un site, qui permet de retrouver sa logique éditoriale au sein du Web, il suffit d'archiver la ou les communautés auxquelles il appartient.
- Ces communautés s'observent par un examen documentaire du Web global. Un travail spécifique permet d'en délimiter des contours et d'en affiner les critères d'archivage.
- Ces communautés correspondent à la réalité du Web, dans ses contenus et sa connectivité, qu'un chercheur pourra consulter. La conservation de la topologie globale lui permettra en outre de comprendre la position de la communauté archivée dans le Web global.

Cette approche se définit donc non pas par rapport à un type de contenu, mais par rapport à des communautés inscrites dans la réalité du Web. En pratique, cette reconnaissance s'effectue en utilisant les outils issus des travaux de Kleinberg (1999) qui permettent de repérer les agrégats sur le Web et leurs principales structures. Ces principes ont été intégrés à l'outil expérimenté à l'INA (voir *infra*).

V. La chaîne expérimentale de l'INA

1. La chaîne technique de collecte

La collecte du domaine est articulée autour d'un système distribué ou chaque site Web est traité de manière indépendante des autres. L'ensemble du système est contrôlé par un ordonnanceur automatique décidant des fréquences de collecte de chaque site répertorié. La collecte de chaque

site repose sur le principe déjà abordé de la reproduction des interactions d'un utilisateur avec le site, à l'aide de logiciels appelés « robots de collectes » (*crawlers*). Le principe en est simple : le robot rapatrie une page d'un site, l'archive, répertorie l'ensemble des hyperliens qu'elle contient, puis rapatrie chacune des pages pointées par ces hyperliens, et procède ainsi pour chacune d'entre elles et cela de manière itérative, jusqu'à ce que l'ensemble des liens aient été suivis[2]. L'intégralité des interactions possibles en terme de suivi de liens d'un page à une autre est ainsi reproduite, et leur résultat archivé. La mise en œuvre pratique d'un tel système recèle de nombreuses difficultés liées aux problèmes de structures et de formats précédemment évoqués. Le robot doit ainsi être capable d'appréhender l'ensemble des formats de contenus existants sur le Web afin de suivre les liens qui s'y trouvent. Les outils de collecte développés à l'INA répondent aux spécificités du domaine visé (fortes contraintes techniques, flux en *streaming*, volumes importants) avec le souci de produire une copie du site la plus fidèle possible, correspondant aux critères d'un dépôt légal.

2. Gestion des mises à jour

Les contenus sont conservés dans leur entier et sous leur forme originale, sans aucune modification ou conversion de format, et sont certifiés à l'aide de signatures cryptographiques. Ce système de signature permet une gestion des versions et facilite le suivi des sites, en détectant au moment de la collecte les contenus ayant été mis à jours par rapport à la précédente collecte. L'ordonnanceur automatique utilise ces informations pour ajuster les fréquences de collecte de chaque site et tenter de les synchroniser au mieux avec les mises à jour réelles du site. L'analyse de l'historique des mises à jour permet ainsi de modéliser le mode de publication du site (publications régulières pendant la journée, mise à jour complète la nuit, sites événementiels dont l'évolution est lié à un festival ou une manifestation ponctuelle, etc.). Dans tous les cas, l'archive est exempte de toute redondance, puisque seuls les contenus nouveaux ou ayant subi des modifications (c'est à dire dont la signature est nouvelle) font l'objet d'un archivage.

[2] Les hyperliens pointant vers des pages déjà rapatriées sont ignorés, de même que ceux pointant vers des pages extérieures au site.

3. Prospection et extension du domaine

La collecte des sites répertoriés et l'analyse de leurs connexions avec d'autre sites permet également la détection de nouveaux sites du domaine pouvant à leur tour faire l'objet d'un archivage, constituant ainsi un outil de prospection de choix pour le maintien du domaine.
L'archive suivie dans le temps de chaque site conserve l'ensemble de ses connexions avec les autres sites archivés, permettant un travail d'analyse et d'enrichissement lors de la phase d'indexation. L'ensemble des informations de collecte de chaque site est également conservé, permettant la restitution lors de la consultation de toute sa dimension interactive. L'utilisateur accédant à l'archive produit ainsi des requêtes auxquelles il suffit de répondre par les contenus envoyés par le site au moment de sa collecte, en réponse aux mêmes requêtes faites par le robot de collecte à la date voulue.

References

Arms, W.Y., Adkins, R., Ammen, C., Hayes, A. (2001). Collecting and preserving the Web: the Minerva prototype. *RLG DigiNews*, 5 (April 2001) http://www.rlg.org/preserv/diginews/diginews5-2.html#feature1

Arvidson, A., Persson, K., Mannerheim, J. (2001). The Royal Swedish Web Archive: a "complete" collection of Web pages. *International Preservation News* 26 10-12 http://www.ifla.org/VI/4/news/ipnn26.pdf

Brygfjeld, S.A (2002). Access to Web archives: the Nordic Web Archive Access Project. *Zeitschrift für Bibliothekswesen und Bibliographie* 49, 227-231

Day M. (2003). Preserving the fabric of our lives: a survey of Web preservation initiatives. In: T. Koch & I. T. Sølvberg, (eds.), *Research and Advanced Technology for Digital Libraries: 7th European Conference, ECDL 2003, Trondheim, Norway, August 17-22, 2003, Proceedings*. (Lecture Notes in Computer Science, 2769). Heidelberg: Springer-Verlag, 2003.

Jon M. Kleinberg (1999). Authoritative Sources in a Hyperlinked Environment *Journal of the ACM* Volume 46, Number 5, 604-632.

Lawrence, S., Pennock, D.M., Flake, G.W., Krovetz, R., Coetzee, F.M., Glover, E., Nielsen, F.Å, Kruger, A., Giles, C.L. (2001). Persistence of Web references in scientific research. *Computer* 34 ; 26-31

Masanès, J.(2002). Towards continuous Web archiving: first results and an agenda for the future. *D-Lib Magazine* 8. http://www.dlib.org/dlib/december02/masanes/ 12masanes.html

Rauber, A., Aschenbrenner, A., Witvoet, O. (2002). Austrian Online Archive processing: analyzing archives of the World Wide Web. In: Agosti, M., Thanos, C. (eds.): *Research and advanced technology for digital libraries: 6th European conference, ECDL 2002*, Rome, Italy. Lecture Notes in Computer Science, Vol. 2458. SpringerVerlag, Berlin (2002) 16-31.

Van der Werf-Davelaar T. (1999). Long-term Preservation of Electronic Publications The NEDLIB project *D-Lib Magazine* Volume 5 Number 9 ISSN 1082-9873

Webb, C. (2001). Who will save the Olympics? *OCLC/Preservation Resources Symposium, Digital Past, Digital Future: an Introduction to Digital Preservation,* OCLC, Dublin, Ohio http://www.oclc.org/events/presentations/symposium/preisswebb.shtm