L'archivage du Web

Marie Puren

L'ARCHIVAGE DU WEB

Marie Puren

13 novembre 2020 : CM4

UVSQ | M1 ACPCI, ECMAH, EMAS, HCS, RCL

PLAN

Un petit détour

L’archivage du Web

Initiatives et projets d’archivage du Web

D’autres initiatives

UN PETIT DÉTOUR

ARCHÉOLOGIE DES MÉDIAS

- Archéologie des médias = une forme d'exploration historique des médias (Jussi Parikka)
- Depuis le début des années 90, des travaux ont "cherché [...] à comprendre les cultures numériques et du logiciel par le prisme du passé".
- Archéologie des médias repose sur les archives nouvelles d'origine numérique.

ARCHÉOLOGIE DES MÉDIAS

- Sources histoire des médias peuvent être sous forme numérique uniquement. Cf. Frédéric Clavert : données numériques = les futures sources pour l'histoire
- Importance du Web dans notre société = utiliser le Web comme une source historique majeure.
- Exemple des travaux du chercheur danois Niels Brügger sur le groupe de radiotélévision public danois DR : utilisation des versions archivées du sites Web

ARCHÉOLOGIE DES MÉDIAS

- Reconstruire l'histoire du Web, en se basant sur les médias issus d'Internet, totalement dépendant de ce qui est archivé ou encore accessible.
- Les historiens du Web = archéologues du Web. Importance de conserver la mémoire du Web
- Exemple de l'art virtuel : plus anciens exemples d'art en ligne seulement connus par des descriptions dans des livres.

L'ARCHIVAGE DU WEB

POURQUOI ARCHIVER LE WEB ?

- Apparition du Web => jamais autant d'informations n'avaient été publiées et disponibles rapidement et pour le plus grand nombre ; et jamais n'y a-t-il eu autant de pertes.
    - 500 millions de tweets par jour => impossible de les archiver tous.
- Brewster Kahle, créateur de l'Internet Archive : "Le web est constamment en train d'évoluer et de disparaître : on peut dire qui si on s'en tient aux statistiques, le meilleur du web est déjà perdu..."
- Environ 1,78 milliards de sites Web dans le monde

POURQUOI ARCHIVER LE WEB ?

- Durée de vie d'un nom de domaine : environ 3,8 ans
- Le patrimoine numérique est donc un patrimoine particulièrement fragile.

POURQUOI ARCHIVER LE WEB ?

“Les leçons du passé ne doivent pas être oubliées à l’ère de la publication numérique, alors que les incertitudes sur la possibilité de la maîtrise de la conservation des nouvelles publications et l’accroissement considérable de leur quantité peut nous pousser à vouloir réserver des moyens insuffisants et que nous avons encore du mal à évaluer à une sélection draconienne initiale des objets qui seront conservés créant ainsi dès le départ des trous énormes dans notre patrimoine futur”

La conservation des publications électroniques et du dépôt légal, Catherine Lupovici (2007)

POURQUOI ARCHIVER LE WEB ?

- Dépôt légal : recenser tous les documents imprimés, graphiques et photographiques.
- Bibliothèque nationale de France en charge du dépôt légal.

POURQUOI ARCHIVER LE WEB ?

- **Loi sur les Droit d'auteurs et droits voisins dans la société de l'information (**DADVSI**)** du 1er août 2006 = intégration du dépôt légal des sites Web (Code du patrimoine, art. L.131-2) :

"Les logiciels et les bases de données sont soumis à l'obligation de dépôt légal dès lors qu'ils sont mis à disposition d'un public par la diffusion d'un support matériel, quelle que soit la nature de ce support. **Sont également soumis au dépôt légal les signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique**".

- Bibliothèque nationale de France et l'Ina (l'Institut national de l'audiovisuel) en charge du dépôt légal de l'Internet.

L'archivage du web d'un point de vue technique

- Infrastructure
- Préservation de l'environnement technologique de ces pages Web
- Difficulté à accéder au Web profond ou "deep Web".
    - Ne pas confondre "deep Web" et "dark Web" : Web profond (deep Web)= partie du Web qui n'est pas indexée par les moteurs de recherche (Toutes les raisons de la non-indexation sur Wikipedia).
- Respect du droit d'auteur.

L'ARCHIVAGE DU WEB D'UN POINT DE VUE TECHNIQUE

- Stratégie simple : compter sur le dépôt volontaire.
- Plus efficace = collecte automatisée par des robots
- Archivage du Web = collecter les sites en tant qu'unités, mais aussi conserver les liens qui matérialisent les relations entre les sites.
- Technique de collecte automatique utilisée dès 1996 aux Etats-Unis par l'Internet Archive, et dès 1997 par la Bibliothèque nationale de Suède.
- Approche semi-automatisée qui combine l'utilisation d'un robot de moissonnage et l'application de certains critères plus sélectifs.

Initiatives et projets d'archivage du web

UNE IDÉE ANCIENNE

Nécessaire de conserver sur le long terme les publications électroniques, “faute de quoi le passé ne laisserait plus de traces, et les recherches antérieures ne pourraient plus être retrouvées, comprises ou reproduites comme il se doit, pour entretenir le cycle continu de l’expérimentation qui fait progresser la connaissance”.

Conférence des directeurs de bibliothèques nationales (CDNL) en 1996

L'APPROCHE INTÉGRALE : INTERNET ARCHIVE

- **Approche intégrale** : collecter l'ensemble du Web, sans faire de distinction ni de sélection : approche retenue par Internet Archive.
- Avril 1996, Brewster Kahle, présente le site Archive.org qui a pour but de sauvegarder toutes les pages du Web, celles qui fonctionnent encore, mais surtout celles qui n'existent plus, qui ont été abandonnées ou remplacées.
- But : "Construire la bibliothèque d'Alexandrie, 2ème version".

L'APPROCHE INTÉGRALE : INTERNET ARCHIVE

- Développement de deux robots “Crawle” et “Spider” avec pour mission de prendre tous les deux mois des instantanés (ou snapshots) des sites publics.
- Pour accéder aux contenus archivés par Internet Archive :
  - naviguer dans les versions archivées d'un site Internet grâce à la Wayback Machine
  - faire des recherches sur des contenus grâce au moteur de recherche d'Archive.org.

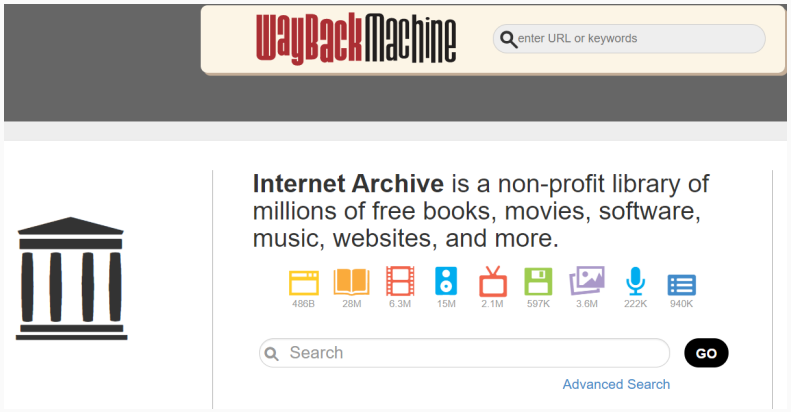

FIGURE : Internet Archive

L'APPROCHE INTÉGRALE : INTERNET ARCHIVE

- Brewster Kahle, qui a vendu sa société Alexa - qui a produit les deux robots - à Amazon pour 250 millions de dollars en 1999, utilise sa fortune pour financer l'Internet Archive (association à but non lucratif).
- Financée par les fonds provenant des bibliothèques nationales qui payent l'Internet Archive pour archiver le Web national et leur livrer.
- Difficultés rencontrées par l'Internet Archive :
  - Robots incapables d'archiver les sites payants ou à abonnement
  - Possible d'interdire aux robots d'accéder auxsites, avec une commande nommée "robot.txt" (protocole d'exclusion des robots)
- 450 milliards de pages archivées par l'Internet Archive. En moyenne, 100 Téraoctets de données de plus chaque mois.

LES AUTRES APPROCHES

- **Approche exhaustive** basée sur le nom de domaine = récolte de site internet ou de ressources en ligne correspondant à un espace national défini. Ex. : Kulturarw3 en Suède, Australian Web Archive.
- **Approche sélective** : archiver certaines portions définies du Web ou des ressources particulières. Exemple projet PANDORA (explorable sur Trove) de la bibliothèque nationale australienne.
- **Approche thématique** : collection de sites Web à l'occasion d'un événement particulier.
    - Collecte du Web électoral par la BnF
    - Collections with Archived Web Sites de la Bibliothèque du Congrès

LES APPROCHES COMBINÉES

- France et Danemark : une collecte exhaustive sur l'ensemble des sites du domaine national + une collecte sélective en fonction de critères définis et campagnes de collectes thématiques.
- En France :
  - Ina chargé de la collecte et de la conservation des sites concernant la télévision et la radio,
  - BnF collecte les autres.

LA STRATÉGIE FRANÇAISE

- Cinquantaine de robots Heritrix, un logiciel libre développé par Internet Archive, qui moissonnent les sites.
- Sites liés à des événements particuliers - et par nature très éphémères pas toujours repérés + profondeur de sites au volume conséquent.
    - Sites signalés aux robots + précise jusqu'à quelle profondeur il faut les capturer.

LA STRATÉGIE FRANÇAISE

- Entre 2004 et 2007 : mise en place du modèle d'archivage des sites
- Signature entre BnF et Internet Archive en 2004 une convention de recherche pour campagne d'archivage du domaine national en .fr.
- Internet Archive a réalisé des instantanés du domaine nationale français de 2004 à 2008.
- Collecte d' archives du Web français qui avait été collectées par l'Internet Archive depuis 1996.
- Chaque année : environ 4,5 millions de sites identifiés avec l'aide de Afnic : des sites du domaineen .fr et en .re (pour La Réunion), produits par des Français ou des personnes domiciliés en France avec des noms de domaine en .com, .org, etc.
    - Pas possible de les conserver en totalité => instanés une fois par an.
    - Collectes plus régulières sur 20000 sites

LA STRATÉGIE FRANÇAISE

- Ina : collectes sur environ 11000 sites.
- Sites émanant des services des médias audiovisuels : Web TVs et Web radios, sites des programmes radio et télé, et sites des organismes du secteur de la communication audiovisuelle.
- Archive également des tweets liés au monde de l'audiovisuel
- Recherches dans ces archives via catalogue de l'Inathèque.

LA STRATÉGIE FRANÇAISE

- Collectes pas exhaustives, mais représentatives.
- Collectes cibées de la BnF
  - Exemples : “Les journaux personnels ou littéraires” ou “La révolution tunisienne à travers le Web”.
  - Depuis 2010 : dispositif de collectes d’urgences pour les événements extraordinaires et des sites en voie de disparition. Exemple : Skyblogs
  - Collectes thématiques extraordinaires. Exemple : collecte sur les réseaux sociaux après les attentats de janvier et de novembre 2015.

2016 : 668 Téraoctets d'archives stockées, soit 26 milliards de fichiers.

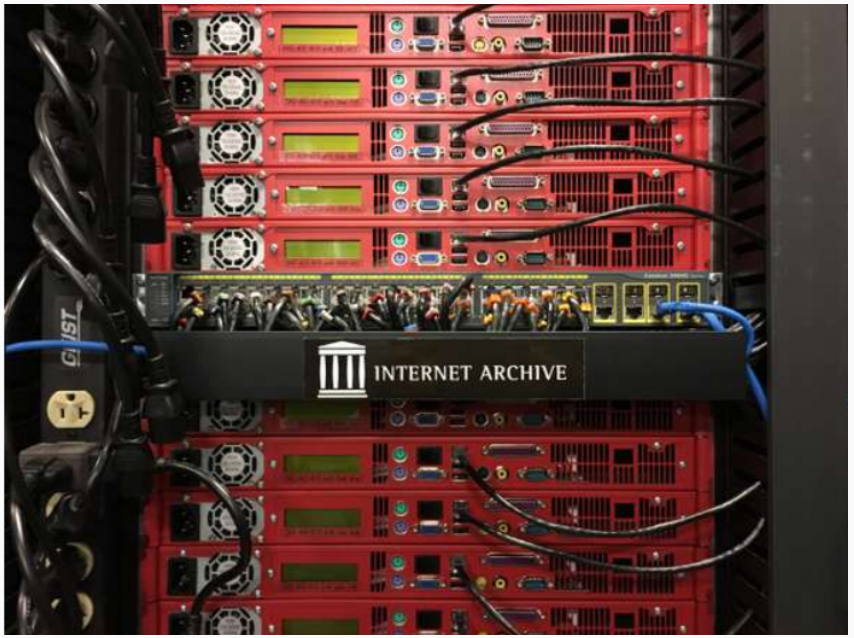

Figure : Stockage à la BnF

LA STRATÉGIE FRANÇAISE

- Consultation des sites archivés uniquement sur place, à la BnF, pour respecter droit d'auteur.
- Le dépôt légal "suspend" l'application du droit d'auteur car permet de rendre disponibles des publications protégées. Pour assurer sa protection : lecteurs ont seulement le droit de consulter ces contenus sur place.

D’AUTRES INITIATIVES

LA BIBLIOTHÈQUE DU CONGRÈS

- Bibliothèque du Congrès : Library of Congress Web Archive.
- Acquisition en avril 2010 l'ensemble des 21 milliards tweets publics publiés entre 2006 et 2010. Archivage systématique des tweets publics abandonné à la fin de l'année 2017. Collection sous embargo.

AU ROYAUME-UNI

- Collecte du Web britannique dans la UK Web Archive
  - De 2004 à 2013 : collectes thématiques.
  - A partir de 2013 : ensemble du domaine en .uk archivé.

CONSORTIUM INTERNATIONAL POUR LA CONSERVATION DE L'INTERNET

- Consortium international pour la conservation de l'Internet (International Internet Preservation Consortium - IIPC)créé en juillet 2003.
- Initiative de Brewster Kahle et Julien Masanès, alors en charge de l'archivage du Web à la BnF :
  - Encourager la collecte et la conservation à long-terme d'une part importante des contenus de l'internet
  - Soutenir le développement d'outils et de normes pour la création d'archives internationales
  - Aider les établissements engagés dans une démarche du Web
  - Faciliter la mise en place d'initiatives et de législation en faveur de la collecte, de la conservation et d'un accès au contenu d'Internet

Explorer d'autres projets

Listes d'exemples :
https://fr.qaz.wiki/wiki/List_of_Web_archiving_initiatives