Quentin Lobbé

Archives et Fragments Web

Désagréger les archives Web pour mener une exploration temporelle de traces numériques des migrations

Université Paris-Saclay, École doctorale des sciences et technologies de l'information et de la communication.
Thèse pour l'obtention du doctorat de Télécom ParisTech et de l'Université Paris-Saclay.

Thèse présentée par **Quentin Lobbé**
LTCI, Télécom ParisTech, Université Paris Saclay & Inria. Paris, France.
quentin.lobbe@telecom-paristech.fr

Sous la direction de :
Pierre Senellart, professeur à l'École Normale Supérieure
Dana Diminescu, professeure à Télécom ParisTech

Soutenue publiquement à Paris le 9 novembre 2018, devant un jury composé de :
Bruno Bachimont (Rapporteur), enseignant-chercheur à l'Université Technologique de Compiègne
Marc Spaniol (Rapporteur), professeur à l'Université de Caen Basse-Normandie
Anat Ben-David, professeure à l'Open University of Israel
Dominique Cardon, professeur associé à Sciences Po Paris
Bruno Defude, directeur adjoint de la recherche et des formations doctorales à Télécom SudParis

last modified May 2018

Il me demanda de chercher la première page.
Je posais ma main gauche sur la couverture et ouvris le volume de mon pouce serré contre l'index. Je m'efforçais en vain : il restait toujours des feuilles entre la couverture et mon pouce. Elles semblaient sourdre du livre.
- Maintenant cherchez la dernière.
Mes tentatives échouèrent de même; à peine pus-je balbutier d'une voix qui n'était plus ma voix :
- Cela n'est pas possible.
Toujours à voix basse le vendeur me dit :
- Cela n'est pas possible et pourtant cela *est*. Le nombre de pages de ce livre est exactement infini. Aucune n'est la première, aucune n'est la dernière.

Jorge Luis Borges - Le livre de sable

Remerciements

Ici je remercie plein de gens
Beaucoup de gens
Mais vraiment

Table des matières

List of Figures

List of Tables

Chapitre 1

Introduction

1.1 Introduction générale

Ici l'intro de la thèse.

1.2 Mise en garde

Penser le passé depuis le présent

Ici on fait un rapide détour par l'historiographie et les difficultés à parler du passé depuis le présent.

Conservation différentielle et nature des archives Web

Ici on parle de la raréfaction de la matière Web à mesure que l'on remonte le temps et également à mesure que le web fournit du contenu.

Chapitre 2

| Du Web aux Représentations en Ligne des Diasporas

2.1 Retour aux origines du Web

2.2 Le migrant connecté

2.3 Le Web, espace de communication et d'organisation

The Web is the main publishing application of the Internet. As such, it consists mainly of the combination of three standards, the URI (Berners-Lee 1994) defining a naming space for object on the Internet, 6 HTTP (Fielding et al. 1999) defining a client–server interaction protocol using hyperlinks at its core, and HTML (Berners-Lee and Connolly 1995) an SGML DTD that defines the layout rendering of pages in browsers. The implementation of these three standards enables any computer connected to the Internet to be- come a publishing system.

But the fact that it is actionable on the Web changes the way references are used by fragmenting content to smaller addressable pieces and overall favoring transversal navigation and access to content which in return, deeply changes the nature of writing as well as reading (Aarseth 1997; Landow 1997; Bolter 2001).

Géopolitique de l'hypertexte

Le web est un digital cultural artifact (Lyman and Kahle 1998)

the Web does, to a large extend, re-use previous forms of publishing 12 (Crowston and Williams 1997; Eriksen and Ihlström 2000; Shepherd and Polanyi 2000), it also invents new ones.

This characterization of the Web as a distributed hypermedia openly and permanently authored at a global scale entails that Web archiving can only achieve preservation of limited aspects of a larger and living cultural artifact.

the interconnectedness of content is a major quality of the Web that raises issues when it comes to archiving.

2.4 L'Atlas e-Diasporas

Chapitre 3

| Archiver le Web

Face à la disparition totale ou partielle des sites Web recensés par l'atlas e-Diasporas (Chapitre 2), il a été décidé de lancer une campagne d'archivage afin de préserver cet héritage numérique et d'anticiper, par là même, la tenue de futures recherches. Sans cette initiative, mes travaux de thèse n'auraient pas pu exploiter et questionner les traces d'un Web aujourd'hui passé.

De part la nature même du médium, le Web demande de penser et de mettre en place un archivage particulier, basé sur des techniques de collectage dédiées. Dans ce chapitre, nous évoquerons la genèse de l'archivage du Web qui, au tournant des années 2000, a connu un essor mondial, mobilisant nombre d'acteurs et d'institutions. Nous introduirons ensuite, d'un point de vue technique, les principales méthodes de sélection, collecte et stockage des corpus à archiver. Nous présenterons enfin les contours des archives e-Diasporas à proprement parler. Ses particularités et ses caractéristiques. Sa durée et son étendue.

Bien que cette thèse se concentre sur l'exploration d'archives Web déjà existantes, il nous semble important d'évoquer la façon dont ces dernières sont constituées en amont afin de mieux saisir les biais analytiques (Chapitre 4) qui motiveront la présentation de notre principale contribution (Chapitre 5). Ce faisant, les éléments que nous nous apprêtons à présenter s'appuieront principalement sur l'ouvrage de J. Masanes : "*Web Archiving*" (Masanès, 2006) qui reste encore aujourd'hui une référence.

3.1 Vingt ans d'archivage du Web

En Octobre 2016, se tenait à la Bibliothèque Nationale de France (BNF) une grande conférence anniversaire réunissant, pour les 20 ans de l'archivage du Web[1], les acteurs français de la pratique. Alors qu'était évoqués les conditions du partage du dépôt légal du Web national entre la BNF et l'INA, il a été rappelé qu'à l'origine chacune des

[1] http://www.bnf.fr/fr/professionnels/anx_journees_pro_2016/a.jp_161122_23_archivage_web.html

deux institutions souhaitait se voir attribuer la pleine gestion de ce dépôt. L'INA mettait en avant ses compétences techniques acquises en archivant les flux audiovisuels nouvellement introduits dans le paysage culturel. La BNF, pour sa part, s'appuyait sur son expérience pluricentenaire de préservation du patrimoine[2].

[2] http://multimedia.bnf.fr/video/prof/161123_10_dl_web.mp4

Cette querelle initiale et son dénouement (la cotutelle du dépôt légale) sont à l'image de l'histoire même de l'archivage du Web: la conjugaison d'une tradition longue de sauvegarde des savoirs et d'un ensemble de techniques de collecte nouvellement pensées pour cet objet complexe qu'est le Web, le tout porté par une poignée de pionniers.

Préserver la mémoire collective

L'archivage du Web s'inscrit dans la tradition longue des techniques d'élaboration et de conservation de la mémoire collective. Tradition qui remonte aux origines même de l'humanité où technique et mémoire se trouvaient étroitement liées.

A. Leroi-Gourhan fait émerger, de l'étude de séries d'objets (silex taillés, percuteurs, harpons, etc) et de figures préhistoriques (gravures et peintures des grottes ornées), une ligne de rencontre entre technique et mémoire (Leroi-Gourhan, 1964). Le préhistorien décrit la technique comme un système évolutif, soumis aux lois générales de la technologie et apparaissant comme transversal à des cultures parfois diverses et éloignées[3]. La technique est chargée, en elle même, de l'histoire passée des continuités, ruptures et transformations technologiques dont elle est l'aboutissement à un instant t.

[3] Leroi-Gourhan associe les formes animales des grottes ornées à des signes, réalisant des couplages basés sur l'observation (comptages et statistiques) de dizaines de cavités. Il cherche à établir une échelle évolutive des styles pariétaux, transversale aux premiers âges de l'Europe de l'Ouest (Leroi-Gourham, 1984)

Avec Leroi-Gourhan, la technique devient mémoire. Elle peut en être chargée et/ou être conçue à dessein de la conserver. Involontairement, le silex taillé porte en lui la trace de l'homme qui l'a élaboré. Lorsque le tailleur finit par mourir, son geste continue à s'extérioriser à travers l'outil qui demeure. Précieux indice pour celui qui vient à sa suite ou pour l'archéologue qui, des millénaires après, saura grâce à cet objet assembler les traces fragmentées d'une pratique passée. Mais l'homme aurait aussi très bien pu choisir, en conscience, d'inscrire son expérience individuelle sur des supports de mémoire dédiés. L'écriture est ainsi l'une des premières techniques de la mémoire, utilisée par l'humanité depuis le néolithique. L'écriture est en cela une *mnémotechnologie* (Stiegler, 1998).

Poursuivant son évolution, l'humanité développe plus avant les techniques de transmission des savoirs pour sélectionner et agréger ses expériences individuelles en une mémoire collective. Des espaces et des structures voient le jour, appuyés par divers pouvoirs politiques ou religieux, avec le double objectif de préserver et d'administrer l'héritage collectif. J. Derrida décrit ainsi le geste d'archiver comme un "*geste de*

pouvoir" (Derrida, 2014). Choisir ce que l'on garde ou non dans les archives ne peut être que le fruit d'une hégémonie, d'une hiérarchie et "*d'un certain nombre d'opérations de pouvoir*" rendues légitimes par une institution. L'État est ainsi caractérisé par *sa capacité d'accumuler, contrôler et exploiter la mémoire collective* (Stiegler, 1991), capacité dont on retrouve divers incarnations au court de l'histoire :

– Au IVe millénaire av. J.C, les tablettes d'argiles étaient accumulées par les mésopotamiens pour constituer les premières bibliothèques

– Entre 535 et 555, Cassiodore pense le Monastère de Vivarium comme un lieu de transmission où, pour la première fois, seraient associés culture savante et christianisme

– François Ier crée le dépôt légal[4] en France, par l'ordonnance royale du 28 décembre 1537, à des fins de préservation culturelle mais également de contrôle politique

[4] "*Nous avons délibéré de faire retirer, mettre et assembler en notre librairie toutes les livres dignes d'être vues qui ont été ou qui seront faites, compilées, amplifiées, corrigées et amendées de notre tems*", extrait de l'ordonnance royale (Dougnac and Guilbaud, 1960)

Les siècle passent et les archives s'adaptent à la transformation des supports de mémoire et à l'émergence de formes nouvelles d'enregistrement. Avec l'arrivée des technologies analogiques[5], il faut désormais capter et archiver des flux d'images et de sons, ce qui conduira en France à la création de l'Institut National de l'Audiovisuel (INA) en 1974. L'apparition du numérique[6] marque la dernière étape de ce cheminement en ouvrant la voie à un renouveau des formes de lecture et d'étude des archives. L'accès à distance de documents numérisés facilite leur consultation, mais il devient également possible de les qualifier, de les annoter ou de les mettre en relation, et ce, de manière large voire exhaustive (Borgman, 2000) :

[5] Cinématographie, photographie, radiodiffusion, etc

[6] Bases de données, logiciels, interfaces, etc

– En 1971, le projet Gutenberg commence à collecter des copies numériques (recopiées et tapées *à la main*) d'ouvrages du domaine public

– Le Thesaurus Linguae Graecae cherche, depuis 1972, à numériser la plupart des textes littéraires rédigés en grecs ancien et toujours subsistants

Mais si le numérique permet aujourd'hui de revisiter des ressources anciennement archivées, il est aussi créateur d'objets nativement numériques tout autant porteurs d'un héritage à préserver. Le web en est la parfaite illustration.

Un héritage numérique

L'archivage du Web débute à la fin des années 90, lorsqu'en 1996 se développent les premières initiatives de préservation du Web, soit 4

années à peine après la publication de la première page sur la toile (Section 2.1). La National Library of Australia est ainsi à l'initiative du projet Pandora[7] qui vise à archiver les publications en ligne australiennes sur la base d'une collecte sélective et continue de sites Web. La Swedish Royal Library, quant à elle, lance le projet Kulturarw3[8] qui s'essaye à une collecte "*intégrale*" mais dispersée dans le temps des sites du Web suédois (Arvidson et al., 2000).

[7] http://pandora.nla.gov.au/

[8] https://web.archive.org/web/20040206225053/https://www.kb.se/kw3

Mais c'est avec la création d'Internet Archive par B. Kahle la même année (Kahle, 1997), que s'écrit véritablement la première page de l'histoire des archives du Web. Ingénieur et activiste, Kahle s'inspire de la Bibliothèque d'Alexandrie pour motiver la création de son organisation à but non lucratif et rendre accessible au plus grand nombre le passé du Web[9]. Utilisant un crawler développé pour le compte de son autre société Alexa Internet, Kahle revendiquait, dans les premières années de la collecte, être capable d'archiver au moins une fois tous les deux mois chacun des sites de l'ensemble du Web (Mohr et al., 2004). La revente d'Alexa au groupe Amazon en 1999 va lui permettre de pérenniser financièrement Internet Archive, qui depuis ce temps n'a eu de cesse d'archiver le Web.

[9] https://archive.org/

Ces pionniers de l'archivage sont rapidement suivis par la Finlande en 1997, le Danemark en 1998 et d'autres pays nordiques rassemblés autour du projet NWA[10] (Hallgrinsson and Bang, 2003). En 2003, la publication par l'UNESCO de la *Charter on the Preservation of the Digital Heritage* (UNESCO, 2003) marque un nouveau tournant pour l'archivage du Web en reconnaissant la valeur universelle d'une telle démarche. La charte fait l'effet d'un l'accélérateur (Figure) pour de nombreuses bibliothèques nationales qui se mettent alors à archiver leur Web national respectif (Gomes et al., 2011).

[10] Nordic Web Archive

Ces différents acteurs peuvent être classés suivant la terminologie introduite par J. Masanes (Masanès, 2006), entre initiatives publiques ou privées, poursuivant un but lucratif ou non. L'accès aux corpus archivés peut être entièrement public ou restreint, en ligne ou physique. Par exemple, Internet Archive est une initiative à but non commerciale, avec un accès public à l'ensemble de ses corpus, en ligne depuis 2001 et physique depuis 2002[11].

[11] La Wayback Machine est officiellement lancée en 2001. Avant cette date les corpus d'Internet Archives n'étaient pas accessibles. En 2002, une copie des archives est disponible à la Bibliotheca Alexandrina, en Égypte

Une autre manière de catégoriser ces initiatives serait de regarder la nature des corpus archivés. Sont-ils basés sur une thématique particulière (tel qu'Archipol, un corpus centré sur les sites politiques néerlandais (Voerman et al., 2002))? Ou sur une extension d'URL particulière pour délimiter les contours d'un Web territorial (d'un état entier en suède à une ville comme Anvers (Boudrez and Van den Eynde, 2002))? Il est aussi possible d'associer chaque corpus à une utilisation caractéristique :

– Pour de la consultation libre ou des

Bon là il faut une carte

Il y des archives régionales ou communales ou basés sur des sites précis The national archives of Australia (National Archives of Australia 2001), UK national library (Brown 2006), Canada, USA (Carlin 2004) have started systematic web archiving. See also the city of Antwerp DAVID's project (Boudrez and Eynde 2002).

la Library of Alexandria, Egypt donne un accès online

Internet Archive or the WA Pacific Islands also preserve information related to foreign countries

Bon il y a of course google et son cache

en france il y a l'ina et la bnf

il y a ceux qui sont centrés pour la recherche et trustés par des université : les jap, Digital Archive for Chinese Studies (DACHS) at Heidelberg University in Germany (see Chap. 10, Lecher 2006), or Archipol for analysis of Dutch political sites at Groningen University in the Netherlands, Voerman et al. 2002)

t'as aussi harchomem

t'as des truc qui se basent sur une problématique, par exemple les éléctions archiving electoral Web sites, such as the Minerva project from the Library of Congress (Schneider et al. 2003) or the French elections Web archive made by the Bibliothèque nationale de France (Masanès 2005)

Mais t'as 4 type : les sites centrés (UK), les wild harvest (Internet Archives), les national harvest (INA/BNF), les recherches (Jap)

L'unesco et puis là c'est la foire à la saucisse quoi However, since 2003 there was a significant and constant growth with the creation of 31 initiatives

The International Internet Preservation Consortium (IIPC) was founded in 2003 qui fait des survey réguliers http://internetmemory.org/ images/uploads/Web_Archiving_Survey.pdf

il y a une communauté d'archivsites Web Archiving Workshop mais peut de personnes bossent vraiment dedans cf le survey

mais bon internet memory stop ...

Si l'on part sur un principe plus ouvertement milantant et libriste de l'ouverture du savoir à tous.

- voire même le travail de Aaron Schwartz qui s'appuit même sur la notion de communs

Les archives sont des biens appartenant à tous pourquoi les privatiser ... les nationaliser ...

Détour par la France

on reprend la construction juridique du bouzin et là c'est la bnf qui fait du .fr tous les ans et l'ina qui fait du média quotidiennement avec un format spécial et aussi des corpus Tweeter ou vidéo

La constitution juridique des corpus en france

3.2 Constituer des corpus d'archives

Archiver le Web a rapidement été considéré comme une nécessité

Si l'archivage du Web a rapidement été considéré comme une nécessité, les premières années n'ont pas été simples. 3 arguments contre revenaient souvent : 1) la qualité n'est pas au rdv & les éditeurs veulent garder le droit sur ce qui doit être fait, 2) le Web se préserve de lui même on peut donc le laisser tranquille, 3) c'est trop compliqué

Le Web est éphémère

Le Web est il self préserving (oui théoriquement mais en pratique uniquement au début) Le web est "relativement" self préserving mais il faut en fait bien comprendre la notion de ressource Web : Car le Web a une cardinalité particulière, double même : Jusqu'à l'imprimerie on faisait des copies de l'original mais il y avait toujours un original (Canfora 1989). En revanche depuis l'imprimerie il n'y a plus d'original. It stabilized content while permitting its wider distribution (Febvre and Martin 1976; Eisenstein 1979) un site a deux cardinalité : l'original et unique posté sur un serveur et l'infinité d'accès possible à cette ressourec de part le web. A resource has a unique source (the Web server) and a unique identifier, but can be generated virtually infinitely and undergoes some degree of variation for each of its instantiations. Donc en fait, archiver le web revient à archiver des ressources Web et tout leurs états successifs

On peut calculer la durée de vie d'un site par rapport à sa Half life Pour Cho and Garcia-Molina (2000) la half life est d'en moyenne de 50 jours Mais c'est en plus en fonction du contexte et des sujets (Fetterly et al. 2003) Et aujourd'hui ça doit être encore plus court (mais peut être paradoxalement peut on trouver des choses qui auront tendance à se perdre) 80% of the pages are updated or disappear after 1 year (What's new on the web?)

Biblio sur les changements du Web et sur comment les crawler peuvent s'adapter

En revanche on peut conserver cette idée de self archiving pour plus tard (car c'est une propriété vraie qui nous emmerdera au moment des

duplicate)

Critère de sélection

Qualité de quoi archiver ** IA utilise une stratégie par pattern - Kimpton et al. (2006) plutôt qu'une stratégie par requêtes à un moteur de recherche - Pandey and Olston 2005 Il peut aussi s'agir de définir à priori la valeur d'un site par son degré (Masanès 2002) (Page et al. 1998; Abiteboul et al. 2002) quelle profondeur

En 2006 IA lance le Archive-it-service qui est une sorte de crowd-sourcing des archives ouvert à tous pour aller contre cette idée de quoi sélectionner On peut aussi parler de l'externalisation des archives Web à diverses fin et des corpus d'archive ou de navigation privée Il y a le cas des tunisiens et de la collecte des vidéos, mais aussi (Rekimoto 1999; Dumais et al. 2003; Ringel et al. 2003). The Internet Archive, Internet Memory and California Digital Library provide web archiving services that can be independently operated by third-party archivists. The services are named Archive-it 3 , ArchiveTheNet 4 and Web Archiving Service 5 http://www.archive-it.org http://archivethe.net http://webarchives.cdlib.org

Et il existe les rogue archivists (Cf conf à Jeru)

Collecte

The term "acquisition" designates the various technical means used to get the content into the archive.

Tout d'abord d'un point de vue de l'architecture, les archives web, s'incrivent naturellement dans l'environnement Web et ne demande pas forcément de récréer des choses en cela qu'elle ne font que freezer un état dynamique. Le Web peut inclure ses propres archives

Premier problème, c'est qu'avec le protocol http il faut procéder fichier par fichier et non en bulk

Il y a 3 méthodes (techniques d'acquistion) qui dépendent en fait du point de vue de l'archiviste

- les crawler dérivés des systèmes de search (Roche 2006 chap 4 du bouqin de mazanes), c'est du client-side-archiving, le crawler étant un client comme un autre (devant sependant faire gaf aux robot;txt) qui peut capturer tout ou une partie d'un site. Ce que le crawler ne capture pas est appelé "Web profond" (Cf le chap 6)

- "transaction-archiving" qui se base sur la participation des visiteurs d'un site donnée Fitch (2003) où l'on ne capture que des tuples request/response et ainsi les diférentes version dès que qq1 la voit

- "server-side-archiving" où le producteur de la ressource fait tout

lui même. Mais ça pose un pb de tout archiver depuis le sserveur, déjà on perd le côté dynamique de ce que l'on voit depuis le client et du coup il faudra côté recréer l'ensemble de l'environnement (bdd, biblio, lybrari) pour tout refaire (tiens là on peut parler de l'exemple de ce site qui a été recré par les amsterdam)

La client-side est la plus simple car en plus elle reflète la manière actuelle de se connecter au web et aussi les intéractions avec les sites Et on explore lien par lien

en fait le crawler imite vraiment l'utilisateur, il doit être actif et on ne peux pas archiver de manière passive ou sans s'impliquer dans ce que l'on fait (comme avec les bouquin ou un flux vidéo que l'on archiver silentieusement)

Les crawler viennent des techno d'indéxation du web Sonnenreich (1997)

Les crawler d'archives capturent tous les fichiers possibles (pas uniquement ce qui est indexable) et comme il ne faut pas se faire banned et respecter les rêgles de politesse d'un site et du coup il peut y avoir un delay certain pour archiver l'entièreté d'un site

Du coup on va plutôt crawler les best page first Cho et al. 1998; Najork and Heydon 2001; Najork and Wiener 2001; Castillo et al. 2004; Baeza-Yates and Castillo 2005)

On peut avoir des stratégy qui cherche à minimiser cela et à garder la cohérence d'un contenu site (Masanès, 2004) (Castillo et al. 2004; Baeza-Yates and Castillo 2005) et où l'on va attaquer la front line d'un crawl Mohr et al. (2004) pour l'Heritrix

Also to be considered is the delay needed to find new sites. It can take lots of time for holistic crawl to discover sites. When it comes to ephemeral sites, related to an event for instance, the delay can be too long to locate and archive them. Si on veux faire un crawl thématique

là je peux parler de l'archivage des diverses sources autre depuis que le bouquin de maznes a été écrit

Chez IA le crawl est effectué par Alexa

à l'INA "A technical approach for the French Web Legal Deposit" Là on parle des sites média fr (mais 50% sont des .com et seulement 30% sont des .fr) ce qui montre aussi la limitation de l'archivage par estension environ 10000 sites ils ont un site crawler qui collecte (un par site) toutes les pages d'un site ils ont un scheduler qui définit les sites à archiver et estime les mises à jour le crawler ne sort pas du scop du site (il conserve une trace des liens externe) et prend un breath first stratégie plusieurs crawler peuvent être envoyés si besoin

parser le web c'est l'enfer, ils ont un ensemble de connecteur maison et en utilise des déjà existants, on va extraire les liens, les vidéos ... n-gram si besoin ou des keywords spécifics mais il n'y a pas qu'une seule et unique stratégie d'indexation "N - Gram - Based Text Catego-

rization"

puis c'est du daff

ha et !!! les vidéos ne sont pas jouées dans le player d'origine mais dans le player de l'ina

Stockage et accès

Bon idéalement l'archive devrait être iso au site, mais ce n'est plsu le web cf brugger Mais en fait on transforme le site pour l'archiver : re-creation of the Web information system

1) Local Files System Served Archives

Là on va carément recréer une copy locale du site archivé où l'on se premène de fichier en fichier en transformant les uri absolu en uri relatif ça marche si tu ne revient pas 36 sur le site (non incrémentale)

sinon y'a le pdf ou un truc non hypertext tout et réorganisé

sinon y'a le

Warc (et autre) vs Daff

et il faut construire un serveur web par dessus

the Internet Archive on the contrary, pays only for storage using compression (as crawl is donated by Alexa)

There are 21 initiatives (50%) that provide full online accessto search mechanisms Some initiatives hold the copyright of the archived contents (e.g. German Bundestag, UK WA, Canada WA) The Internet Archive and the Portuguese WA proactively archive and provide access to contents but remove access on-demand. On the other hand, for 16 initiatives (38%) the access to the collec- tions is somehow restricted. The Library of Congress, WebArchiv and Australia's WA provide public online access to part of their collections. BnF, Web@rchive and Preservation .ES grant access exclusively through special rooms on their facilities.

The Archive-Access tools are dominant (62%), including the Heritrix, NutchWAX and Wayback projects,

Nonetheless, NutchWAX supports full-text search for the Finnish WA (148 million), Canada WA (170 million), Digital Heritage of Catalonia (200 million), California Digital Library (216 million) and BnF (estimated 2 100 million). Australia's WA supports full-text search over 3 100 million contents indexed using an in-house developed system named Trove. I

Even with the high computational resources required for this purpose, 67search for at least a part of their collections [14]. In another survey about European web archives this percentage is 70[13].

The Living Web Archives (LiWA) aimed to provide contributions to make archived in- formation accessible [19]. It addressed problems shared with other information retrieval (IR) areas, such as web spam

de- tection, terminology evolution, capture of stream video, and assuring temporal coherence of archived content. LiWA was followed by the Longitudinal Analytics of Web Archive data (LAWA), which aims to build an experimental testbed for large-scale data analytics [28].

!!! aussi de dire que Internet Archives bouge ces data storage pour les garder à l'abri

ha et là je parle de comment on fait des analyses on top of web archives

et on se dit que tout de même ça ressemble un peu à une capsule temporelle que l'on regardera plus tard là placer la ref aux capsules tempo

3.3 Les archives Web de l'Atlas e-Diasporas

Présentation rapide de l'ensemble des corpus et focus sur les Marocains (expliquation ...)

Chapitre 4

| Traces Discrétisées et Temporalité Figée

4.1 Détruire pour mieux archiver

on parle en fait de pages depuis alta vista The launch of the AltaVista service in December, 1995 proved that all of the pages on the Web could be treated as a single collection, and indexed and made searchable for all users on the Net.

De Derrida aux traces discrétisées, de la sélection effectuée par le crawler et l'archiviste, les archives sont des traces discrètes du Web, comme Funes on ne peut tout garder

4.2 Un temps sans extension

Ici on part de Saint Augustin et de sa définition d'un présent sans extension qui a influencer le rapport des occidentaux au temps. Ce rapport au temps se retrouve lorsque l'on étudie en détail les modèles d'exploration des archives web qui s'appuient sur la date de capture d'un contenu. S'en suit plusieurs remarques qu'il faut conserver en tete avant de se plonger dans toute exploration

proposer ici une échelle de datation (date de téléchargement et date de last modified)

Niveau	Nature de la date
page	téléchargement
page	dernière modification

↓ Validité historique

Table 4.1: Échelle de datation d'une page Web archivée

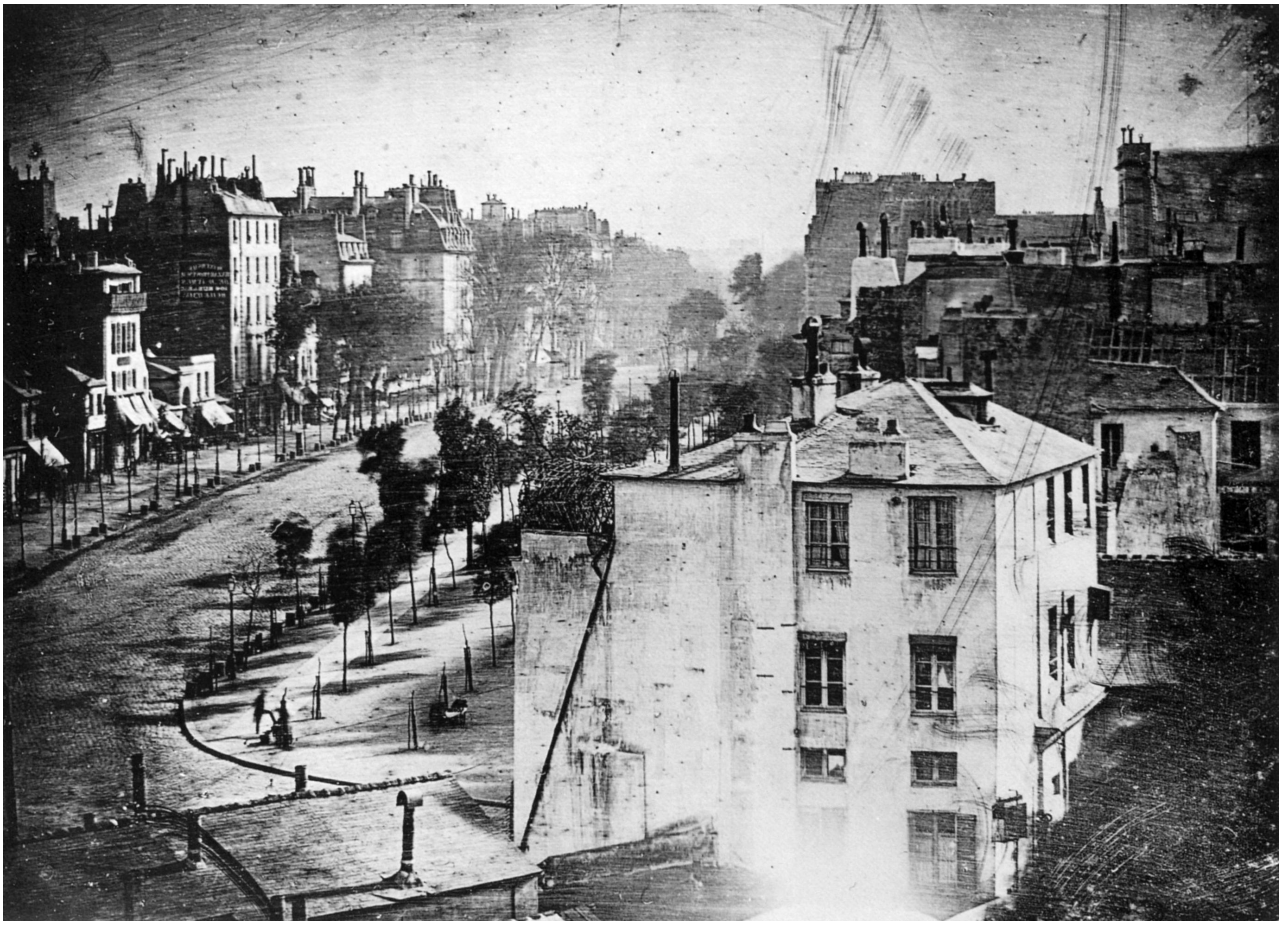

Figure 4.1: "Boulevard du Temple", Louis Daguerre, 1838

Figure 4.2: Warc vs Daff

Crawl blindness

Cohérence

Duplicata

4.3 Construire un moteur d'exploration d'archive

Extraction et enrichissement

Définition du schéma d'indexation

bon y'a déjà des travaux

Web archives receive a significant number of queries with a specific time interval of interest, denoted time-travel queries "Characterizing search behavior in web archives". Thus, partitioning the index by time enables searching only the sub-index corresponding to that interval.

In this work, we refer to the time when the web documents were crawled, but we could use other time attributes associated to documents, such as their creation or publication dates. Some works use instead the time mentioned on the document text "On the value of temporal information in information retrieval".

We can have at least four types of partitions, where the index is split per time and then per document or term, or the opposite.

When partitioning first by document and then by time, subsets of documents are allocated to computer clusters ac- cording to their identifiers "A time machine for text search"

Les gars préfèrent rester à la page The document-based partition of the index offers superior query throughput and does not require rebuilding of the in- dexes each time new documents are indexed, contrary to the term-based partition.

moi je vais en dessous

Alternatives for rebuilding the indexes when using the term-based partition exist, but are also more complex and less efficient than using the document-based partition "Index maintenance for time-travel text search".

The open source Wayback Machine (WM) is a set of loosely coupled modules that can be exchanged and configured ac- cording to the web archive needs "Wayback' for Accessing Web Archives"

The Internet Archive's WM uses as index a flat file sorted alphabetically by URL and divided in similar size buckets. The buckets are distributed across web servers and map URLs to the ARC files storing them [10]. Thus, each web server responds over a subset of documents (i.e. URLs), meaning that this architecture uses a document-based index partition. When the WM receives a URL query, a broker routes

it to the web server handling that range of URLs, which replies with the list of all URL versions. To access one of these versions, the web server replies with the respec- tive ARC name. Then, the ARC file is located by sending a UDP broadcast to all storage servers. There are more than 2500 storage servers in the primary data center, each with up to four commodity disks. The storage server containing the ARC file replies to the broadcast with the local file path and its identification. Finally, the client browser is redirected to the respective storage server, which runs a lightweight web server to deliver the ARC file. The Portuguese Web Archive (PWA) http://archive.pt is based on the WM, but uses NutchWAX as its full-text and URL search engine "Introducing the Portuguese web archive initiative."

NutchWax est un search basé sur lucene mais adapté pour bouffer du WARC "Full text searching of web archive collections." qui travaille sur une indéxation à la page puis sur un bucket of time

Il y a d'autre search qui font du solr pur au lien de Nutch "A survey on web archiving initiatives" Everlast est un systeme de search en peer to peer ça partitionne par terme puis par time "EverLast: a distributed architecture for preserving the web."

à la bpi c'est du elasticsearch

Détection d'événements

4.4 Les archives sont des traces indirectes du Web

Les archives sont les traces directes du crawler et non du web (Cf mises en gardes précédentes) + exemple sur yabiladi.com donc il faut descendre au niveau de la page et y extraire d'autres temporalités, d'autres forme d'exploration qui ne dépendent pas non plus de la linéarité proposé par les moteurs d'exploration classique. La désagragation se fait dans le modèle de données mais également dans la façon de conduire sont exploration.

Chapitre 5

| Fragmenter les Archives Web

Les effets de crawl legacy (Section 4.4) sont indissociables des archives Web telles que nous les connaissons. Liés organiquement à la structure même des fichiers archivés (Figure 4.2), ils en sont les artéfacts directs. Pour qui souhaite conduire l'analyse d'un ensemble de sites Web archivés, ces effets induisent nombres d'obstacles : collectage non régulier, sur-représentation de certaines parties d'un site, incohérences entre contenus archivés, etc. Lors de la consultation des archives, l'explorateur n'a que très rarement la main sur les commandes du crawler et doit se contenter de l'état du corpus qui lui est proposé.

Nos travaux portant sur l'exploration de corpus d'archives Web déjà existants et/ou constitués de longue date, nous ne proposerons pas ici d'alternative aux formats WARC et DAFF. Nous chercherons plutôt à définir une stratégie d'exploration capable de s'affranchir de l'héritage pesant des crawlers sur les archives Web ou tout au moins d'en atténuer les effets. Il faut également préciser que nous conditionnons notre réflexion à la réalisation d'une exploration large (en terme de pages à visiter) et profonde (en terme de durée à parcourir) des archives Web qui implique le développement d'une méthodologie dédiée à une échelle si large (Chapitre 6). Il va sans dire, que pour l'étude à la main d'une poignée de sites ou de pages (depuis la WayBack Machine par exemple), les effets de crawl legacy restent parfaitement surmontables. En revanche, cette tâche devient rapidement fastidieuse voire titanesque à mesure que grandit le périmètre d'exploration et que la validation humaine s'efface au profit d'un algorithme ou d'un ensemble de scripts.

Sur ce point, nous proposerons l'introduction d'une nouvelle unité d'exploration des archives Web : le **fragment Web**. L'essentiel de ce chapitre sera consacré à explicité l'intuition selon laquelle il peut être bénéfique de descendre au delà du niveau des pages Web archivées en proposant un changement d'échelle analytique. Le fragment Web devra d'une part offrir aux explorateurs du Web passé une plus grande souplesse et de nouveaux outils pour déconstruire, recomposer et questions les archives, et d'autre part, cette unité devra devenir un object

d'étude à part entière. Plutôt que d'être une trace directe du crawler, le fragment Web cherchera à témoigner des gestes de l'auteur ou du lecteur d'un site archivé dont le passages parmi ses pages aura laissé des indices qu'il nous faudra exploiter. Nous nous concentrerons particulièrement sur la question de la bonne datation d'une page Web archivée en nous basant sur les dates de création et d'édition de tel ou tel contenu. En associant le fragment Web à une date d'édition nous discuterons du changement de temporalité ainsi constaté : passant du crawler à la page Web et ses fragments. Nous en illustrerons ensuite le bénéfice potentiel en terme de précision historique. Enfin, nous reviendrons en miroir sur les modalités techniques et théoriques d'un moteur d'exploration d'archives Web (Chapitre 4) prenant dès à présent le fragment Web comme unité principale d'indexation. Une démonstration en sera donnée via un cas simple de détection d'événements parmi le contenu fragmenté des archives de *yabiladi.com*.

Tout au long de ce chapitre, nous appuierons nos réflexions sur les travaux de l'historien du Web N. Brügger et sur la notion de *strates analytiques du Web* qui servira de cadre à la définition de nos propres fragments Web. La question du changement de temporalité sera, quant à elle, abordée à l'aune des recherches de l'historien médiéviste J. Baschet sur les enjeux historiographiques d'un tel déplacement.

5.1 Vers une nouvelle unité d'exploration

Toute archive est une matière destinée à être désagrégée ou ré-arrangée en vue de la questionner et d'écrire l'histoire. Ainsi le professeur d'archivistique E. Ketelaar dit d'une archive qu'elle ne parle pas seule (Ketelaar, 2006), qu'elle n'est jamais fermée ou complète. Une archive se tient toujours prête à être réinterprétée par une nouvelle génération d'explorateur ou de chercheur. Mais s'il est clair que la direction prise avec l'introduction du fragment Web est celle d'une déconstruction des corpus d'archives Web existants, gardons à l'esprit comme le rappel Derrida[1], que le document d'origine ne doit en aucun cas être altéré ou modifié. Et ce, pour justement permettre à d'autres, après nous, d'à nouveau s'y référer, le faire parler.

[1] "*je peux interroger, contredire, attaquer ou simplement déconstruire une logique du texte venu avant moi, devant moi, mais je ne peux ni ne dois le changer*", (Derrida, 1995), p.374.

Précisons donc avant toute chose, que le fragment Web, ne sera pas une version modifiée d'une page Web archivée et de ses fichiers sources, mais bien une entité autre, issue du fractionnement de cette page et utilisable en parallèles des modèles d'exploration d'archives déjà existants.

Découper, déplacer, monter

Nous évoquions déjà dans la section 4.1 le personnage de Funes imaginé par Borges qui, dans la fable, à force de ne plus jamais rien oublier, voyait décroitre ses capacités à penser et à raisonner. Funes vit dans l'indexation d'un perpétuel présent. Il se redécouvre sans cesse et n'arrive plus à se recréer des souvenirs, à se raconter de mémoire sa propre histoire[2].

Pour mémoriser ou archiver il faut oublier. Ré-arranger et faire du montage. Nos souvenirs sont des sélections qui, mises bout à bout, collées, accélérées ou ralenties forment le fil de nos histoires. Le cinéaste C. Marker donne corps à cette idée dans son film d'archives "*Le Fond de l'Air est Rouge*"[3] où la posture de l'historien face à un document archivé se rapproche de celle du monteur de cinéma face à une matière filmée. Leurs outils sont semblables. Lorsqu'il invente l'histoire, l'historien découpe, isole et rapproche des sources archivées potentiellement très éloignées.

Dans son court métrage *"Je Vous Salue, Sarajevo"*, réalisé en 1993 pendant la Guerre de Bosnie-Herzégovine[4], J.L. Godard déconstruit une photographie du reporteur de guerre R. Haviv. Il fragmente et fait se confronter des inserts éclatés à la manière d'un collage-poème ou d'un cinétract[5]. Par le collage, les fondus et les découpes Godard rompt la continuité de l'archive qu'il utilise comme source première afin de rendre compte image après image de la cruauté qui frappe Sarajevo, une ville de son temps (Figure 5.1). Le film finit par dévoiler entière, l'image dans toute son horreur. Décomposer pour mieux recomposer.

Figure 5.1: Extraits de "*Je Vous Salue, Sarajevo*", J.L. Godard (1993) à partir d'une photographie de R. Haviv (1992)

[2] "*[Funes], ne l'oublions pas, était presque incapable d'idées générales, platoniques. Son propre visage dans la glace, ses propres mains, le surprenaient chaque fois.*", (Borges, 1974), p.??.

[3] Film réalisé en 1977. C. Marker superpose par exemple aux souvenirs de S. Signoret des extraits remontés et recolorisés du "*Cuirassé Potemkine*" d'Eisenstein (`https://youtu.be/d01E4GYjF1s`)

[4] Voir `https://youtu.be/WKbfu8rRrho`

[5] Mini-films non signés à caractère militant, réalisés en mai et juin 1968 (`https://fr.wikipedia.org/wiki/Cin%C3%A9tract`)

Il y a dans les travaux de Godard et de Marker une souplesse d'action vis à des archives à même de se révéler profitable si nous l'appliquions aux archives Web. Chercher à avoir en main des éléments fragmentés de pages Web éloignées, que nous pourrions associer, à souhait, afin de traiter plus largement d'un moment particulier de l'histoire du Web. Comment se donner la possibilité de rapprocher automatiquement deux contenus archivés hors du carcans de leurs pages respectives ? Peut on par exemple demander à un moteur d'exploration d'archives de nous retourner l'ensemble des messages postés sur un forum, par une seule et même personne il y a dix ans de cela ?

Les strates analytiques du Web

Le glissement d'un niveau d'analyse à un autre, vers un en-dessous de la page archivée, est formulé pour la première fois par l'historien du Web N. Brügger lorsque, cherchant à définir le site Web comme objet potentiel de recherches historiques (Brügger, 2009), ce dernier en vient à introduire la notion de **strates analytiques du Web**[6].

[6] En anglais : *analytical Web strata*.

Brügger suggère de construire un système d'analyse dynamique pour réajuster, à souhait, le périmètre d'une recherche portant sur le Web. L'observateur doit ainsi pouvoir passer d'un ensemble de sites, à une page unique, voire descendre jusqu'aux éléments constitutifs de cette dernière (un texte, une image, etc)[7]. Cette approche, notons le, n'est pas confinée au Web archivé, elle peut très bien s'adapter au Web vivant. Brügger définit 5 niveaux d'analyses (présentés ci-dessous du plus englobant au plus précis):

[7] "*One can distinguish the following five analytical strata: the web as a whole; the web sphere; the individual website; the individual webpage; and an individual textual web element on a webpage, such as an image*", (Brügger, 2009), p.19

1. le Web dans son entièreté
2. la sphère Web
3. le site Web
4. la page Web
5. l'élément Web

La Figure 5.2 donne à voir une illustration de l'agencement de ces strates. Le premier niveau englobe l'entièreté des sites du Web vivant. Il inclut également les éléments de back-end (base de données, code côté serveur, etc) et plus généralement l'ensemble de l'infrastructure physique du Web (serveurs, câbles réseax, supports numériques, etc).

La sphère Web, inspirée des travaux de Foot et al. sur le volet numérique des campagnes électorales états-uniennes du début des années 2000 (Foot and Schneider, 2006), désigne un ensemble de sites Web sélectionnés par un chercheur. C'est une construction ad hoc motivée par une question de recherche donnée, une thématique précise.

Les acteurs Web regroupés au sein de ces sélections n'ont pas forcément conscience d'appartenir à un tel groupe. Par exemple, les réseaux de sites e-Diasporas (Section 2.4) peuvent être considérés comme des sphères Web. Sites et pages Web sont ensuite définis de manière égale à ce que nous proposions au Chapitre 4.

L'élément Web est, quant à lui, défini comme l'élément textuel minimal d'une page Web[8]. Ce peut être un ensemble de caractères écrits sur une page, des images fixes ou mobiles, ainsi que des sons. Brügger en revanche écarte de cette liste les menus, barres d'informations et autres éléments de navigations.

L'historien propose ensuite trois pistes d'analyse pour chaque éléments Web : 1) considérer le médium (l'écran, l'ordinateur, le support de lecture, etc) 2) considérer l'élément textuel seul 3) considérer un état hybride médium/texte où le code informatique derrière la page Web serait à la fois object et support du message[9]. Cet état, propre au numérique (Finnemann, 1997), est explicité dans un texte antérieur (Brügger, 2002) comme la rencontre d'une matérialité matérielle (les ordinateurs constitués de pièces de plastique, de métal ou de verre) et d'une matérialité immatérielle (le code informatique[10]).

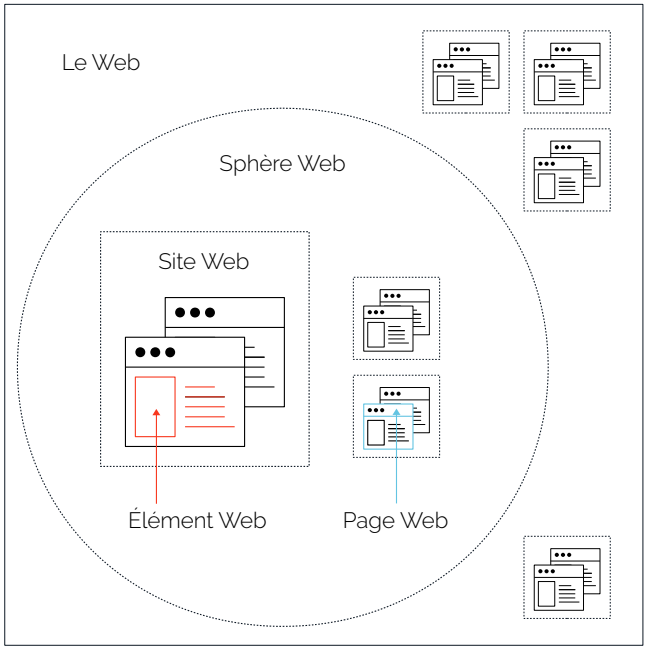

Figure 5.2: Les 5 strates analytiques du Web, d'après (Brügger, 2009)

Le fragment Web pourra assez naturellement s'inscrire dans la continuité des strates analytiques du Web, se situant quelque part entre l'élément Web et la page Web. Mais s'il ne nous semble pas nécessaire de descendre à un tel niveau de différentiation entre matérialité et immatérialité du Web[11], nous tenons tout de même à conserver la dif-

[8] "*The Web element is the minimal textual element on a webpage*", (Brügger, 2009), p.20.

[9] "*the intermediate textual level that the codes of 0/1 and their syntax constitute, a level that is in itself a text, insofar as it is composed of letters and a syntax*", (Brügger, 2009), p.9.

[10] "*the energy-based binary alphabet*", (Brügger, 2002), p.21.

[11] Rappelons que le Web archivé n'est déjà plus le Web, qu'il n'est que l'enregistrement sur fichier de ses éléments de front-end (fichiers HTML et CSS). Back-end et supports physiques ne faisant pas l'objet d'un archivage.

férence entre un élément tel qu'affiché à l'écran depuis le navigateur et la portion de code informatique dont il est l'expression. Nous voulons par exemple pouvoir demander à un moteur d'exploration d'archive de retourner l'ensemble des pages contenant une balise HTML donnée tout autant qu'un mot clé dans un paragraphe. Le fragment Web traduira ainsi une portion de code informatique et son interprétation à l'écran.

Dater une page archivée

Notre réflexion sur les différents niveaux d'analyse ne doit pas être menée que d'un point de vue purement structurel. Il serait également intéressant de discuter de la datation d'une page Web archivée pour tendre vers une plus grande précision historique. Comment bien dater une page Web et le contenu dont elle est le support ? Comment approcher le plus précisément possible de la date de création réelle d'un message posté sur un forum ?

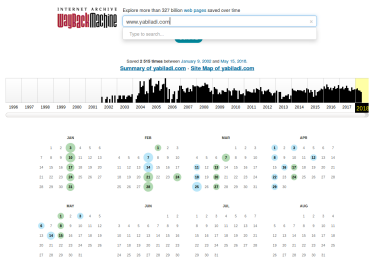

Figure 5.3: Répartition des archives de *yabiladi.com* dans la WayBack Machine (`https://web.archive.org/web/*/www.yabiladi.com`)

Les archives Web sont les traces directes des crawlers en charge des collectages (Section 4.1). En base de données, rappelons-le, une page archivée est indexée par sa seule date de téléchargement (aussi appelée date d'archivage) sur laquelle s'appuient les moteurs d'exploration existants pour retourner les résultats d'une recherche (par exemple la WayBack Machine, Figure 5.3). En comparant deux versions d'une même page archivée, il possible de déterminer une date de dernière modification (Section 4.2) et de l'intégrer à une échelle de datation (Table 4.1).

En introduisant le fragment Web, nous souhaitons améliorer la précision historique des contenus archivés. Ne pas se satisfaire des seules date de téléchargement et de dernière modification. Pour se faire, nous nous donnons comme objectif de chercher à identifier la date d'édition de chaque fragment Web et de répercuter cette donnée sur la datation des pages Web associées.

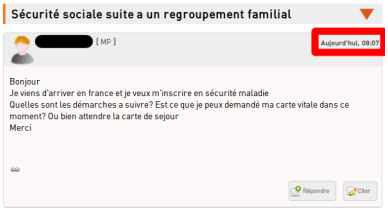

Figure 5.4: Date de création (rouge) d'un post de forum sur *yabiladi.com*

Une page Web évolue (Section 3.2) dès que son contenu est édité par un tiers : humain ou robot. Par *édition*, nous entendons la création, la modification ou la suppression d'un élément d'une page. Comme les actes de modification et de suppression demandent, pour être datés (même approximativement), de comparer deux versions archivées d'une même page (Rocco et al., 2003; Nunes et al., 2007), leur détection semble de prime abord compliquée à intégrer à notre moteur d'exploration d'archives, ce dernier travaillant sur des pages uniques et non des séquences de versions successives.

La création d'un message ou d'un commentaire peut en revanche être plus facilement datée. Des indices sont souvent dispersés à même

la page (Figure 5.4), reste alors à les interpréter et à les formater avant indexation (De Jong et al., 2005; Kanhabua and Nørv\a ag, 2009). Si l'en-tête HTTP d'une page Web a été archivé, celui-ci peut nous renseigner sur une date de dernière modification qui ne dépende pas directement du crawler (Amitay et al., 2004). À défaut, la création d'un contenu donné sera rapportée à sa première apparition sur l'ensemble des versions archivées d'une même page (Jatowt et al., 2007), cette comparaison peut être affinée si des URIs ont été par ailleurs collectées[12] (Aturban et al., 2017). Notons enfin qu'il existe des stratégies de datation adaptées à la nature interdépendante de certains contenus archivés, comme un réseau de citation d'articles de blogs par exemple (Toyoda and Kitsuregawa, 2006; Spitz et al., 2018).

[12] Le système Memento propose de voir une page archivée comme la concaténation de toutes les URIs qu'elle agrège. Cette vue est appelée TimeMaps (Van de Sompel et al., 2013) et peut être exploitée pour comparer les dates de certaines URIs d'images par exemple.

À partir de ce point, nous assumerons, pour simplifier, que la date d'édition d'un fragment Web équivaut à sa date de création. Par extension, il est alors possible de dire que la date de création de toute page Web est, au mieux, la plus ancienne date d'édition de l'ensemble de ses fragments. De là, nous pouvons proposer une mise à jour de l'échelle de datation introduite par la Table 4.1, telle que :

Niveau	Nature de la date	
page	téléchargement	
page	dernière modification	Validité historique ↓
page	création	
fragment	édition	

Table 5.1: Échelle (actualisée) de datation d'une page Web archivée

Tout fragment Web devra donc, dans le mesure du possible, être associé à une date d'édition. Dans le cas contraire, sa date sera rapportée à celle de la page à laquelle il est associé. Et si bien dater une page archivée participe de son émancipation vis à vis du crawler, cela donne, par la même occasion, corps aux acteurs qui l'ont fait vivre.

Un article de blog ne s'écrit pas de lui même, il est le fruit du geste d'un auteur (unique ou collectif, humain ou robot) qui l'a mis en ligne. Derrière les dates d'édition des fragments Web, peuvent transparaitre les gestes de divers auteurs : blogueurs, commentateurs ou contributeurs qui deviennent dès lors objets ou dimensions possibles d'une exploration d'archives Web[13]. Serait-il alors possible, comme le suggère l'historien J. Morsel, d'écrire une histoire *symptomale*[14] (Morsel, 2016) à partir de nos corpus d'archives Web ? Cela reviendrait à considérer que certains fragments Web se trouve chargés de la présence l'attente d'un auteur, fossilisée sous la surface des pages archivées, prête à être questionnée. Cette nouvelle perspective d'exploration nous mènera

[13] Pour le philosophe V. Flusser les gestes sont des séries de mouvements significatifs dont le but est déchiffrable, ils "*montrent la façon dont nous sommes au monde*", (Flusser, 2014), p.319.

[14] Alors que la trace, telle que nous la décrivions jusqu'ici, suggère l'absence de l'agent qui l'a produite (elle s'en est détachée), le symptome, selon Morsel, suppose la présence latente de l'agent, coprésent à ce dont il est le signe (Morsel, 2016)

à considérer, depuis les archives Web, le devenir de communautés d'utilisateurs ou de collectifs d'auteurs tel que nous l'illustrerons dans le Chapitre 6. Une nouvelle dimension d'analyse des archives s'offre donc à nous : l'exploration par acteur (auteur, contributeur, commentateur, etc).

Figure 5.5: Dimensions d'exploration des archives Web (ajout de l'acteur)

Désagréger pour changer de temporalité

Il n'est pas simple de retrouver une date d'édition dans une page Web vivante ou archivée. Cela a déjà été proposé par plusieurs travaux. On peut partir sur des indices ou des comparaisons entre version. C'est compliqué oui mais les bénéfices en terme de précision historiques sont impressionnant, afin de l'illustrer nous allons procéder à une expérience dans laquelle nous allons rapidement extraire l'ensemble des dates de création de tous les postes de yabiladi. Tout au long de sa vie la structure du fichier HTML derrière le forum a peux évolué et les class de noeuds sont restée identiques. Nous allons donc faire une extraction spécifique à yabiladi forum et comparer pour chaque page sa première date d'édition vs sa date de download. Nous reprenons donc la répartition proposée au chapitre 4 à laquelle nous ajoutons en bleue la répartition des dates d'éditions.

La réaprtition est plus linéaire et ne relève pas de trous entre 2013 et 2014. PLus intéressant encore, en passant de la temporalité vue depuis le crawler à celle vue depuis le site ou le fragments on peut remonter jusqu'en 2003. Soit une année seulement après la création véritable de yabiladi.com. Un contenu archivé contient plus de mémoire que ce qu'il ne semble offrir de prime abord

Là est tout l'enjeu du chagement d'unité que nous proposons qui est en réalité un changement de temporalité.

De la même manière, dans ses derniers travaux historiographiques (Baschet, 2018), l'historien médiéviste J. Baschet à recourt à W. Benjamine pour réaffirmer la nécessité de rompre avec une vision unilinéaire de l'histoire. Il faut, selon lui, faire éclater la continuité de l'histoire pour en isoler des constellations afin de mieux saisir l'ensemble d'un mouvement historique.

Et il existe d'autres temporalités, comme le présente Husserl tmtc ou les indiens du chiapas. Bref le fragment web c'est un changement d'échelle spatial et temporel. Et voici maintenant le moment de vous le présenter.

5.2 Le fragment Web : définition

A partir de maintenant nous assumons la nécessité de trouver une nouvelle unité d'exploration des archives web baptisé fragment web, s'inscrivant dans la 5ème strate du web et émancé (autant que possible) de tout lien avec le crawleur. Cette unité devra autant que possible être relié à une éditiond ate pour éviter les crawl legacy et maximiser la présision historique.

Comme la forme de ces fragments sera toujours lié au contexte de l'exploration dans laquelle elle sera utilisée, et comme nous voulons que des sociologues oud es historiens puissent s'en saisir (car un historien demandera toujours à conaitre le contexte (cf le papier sur les archives là) la définition suivante sera générique à dessein. Une définition pratique et son extraction technique sera proposer dans la section suivante et targété pour la question des collectifs migrants etteinds.

Définition

Considérant la page Web comme unité de consultation de base du Web, bâtit sur des modalités d'écriture propre au support numérique et constatant que du point de vue de la perception humaine une page web est le résultat de l'agencement logique de fragments sémantiques distincts, alors :

le fragment est un sous ensemble cohérent ...

5.3 Scraping et méthodologie d'extraction

Extraire de l'information issue d'une page Web

Le scrapping c'est quoi ? Les méthode classique (visuel vs le reste) Néttoyer ou tout conserver ? Pourquoi scrapper ? (orienter business ?) l'on parle de readability et de fathom...

Implémentation technique

Là on parle de rivelaine et de la fonction distance ... L'algo les différents filtres on dit qu'il y a plusieures implémentations

Exemples et discussions

là on donne qq exemples et on dit que en tant qu'ingé on avait cherché un truc qui fonctionnait tout le temps mais c'est pas possible

et un historien a besoin de vouloir resizer à volonter Là on parle de l'automatique vs le fait à la main avec le truc firefox Du coup dans la suite, le treshold de la fonction distance sera toujours testées avec le truc firefox

5.4 Penser une exploration désagrégée

Atténuer les "crawl blindness"

Cohérence relative entre archives

Dédupliquer les corpus

5.5 Intégration à un moteur d'exploration

D'un schéma à un autre

Retour à la détection d'événements

S'éloigner des moteurs d'exploration

Chapitre 6

| Explorations de Collectifs Migrants Éteints

Où l'on parle d'exploration de blogs, de forum et de moments

6.1 À la recherche de l'étonnement : l'analyse exploratoire de données

De Tuckey à Fry

Où l'on explique l'EDA de où ça vient

Abduction, déduction, induction

Où l'on introduit la philosophie générale de l'EDA et on peut faire un lien avec Ginsburg

Méthodologie technique d'exploration

Où l'on explique comment techniquement nous allons procéder en suivant plutôt Fry

6.2 Les traces d'une mutation numérique

D'une communauté vibrante de blogs ...

Là on raconte l'état des blogs en 2008

En revanche, une blogosphère dont les acteurs sont parfois à l'origine même de sa construction et de sa promotion[1] n'est pas strictement équivalente à une sphère Web. (Keren, 2006)

[1] De 2007 à 2011 la blogosphère marocaine a organisé ses propres *blog awards* pour récompenser, promouvoir et connecter ses acteurs. Voir `https://fr.wikipedia.org/wiki/Maroc_Web_Awards`

... à un collectif éteint

Là on raconte l'état des blogs en 2018

Définir l'espace d'exploration

Là on explique la forme des fragments que l'on va chercher à retrouver

Migration d'un territoire Web à un autre

Là comprend que les blogs se sont déplacé vers Fb et Twitter

Conserver son identité numérique

Là on parle de la communauté des blogs

Le Printemps Arabe vu comme un moment-clé

Là on introduit le Printemps arabe marocain

6.3 Un soulèvement en ligne éphémère

Yabiladi.com : porte d'entrée sur la diaspora

Là on explique ce qu'est Yabiladi

La manifestation du 20 Février 2011

Là on rappelle ce qu'est cet événement

Définir l'espace d'exploration

Là on explique la forme des fragments que l'on va chercher à étudier

Voir un site évoluer

Là on explique comment on va visualiser ces fragments

Agréger les contributeurs

Là on s'intéresse au graph des contributeurs

De l'embrasement à l'évasion

Là on regarde les clusters de Threads

6.4 Les Moments Pivot du Web

Les limites de l'archivage du Web

Les archives ne capturent pas le Web comme un environnement

Les moments pivots du Web

Un moment pivot c'est quoi ? Les geste et compagnie ainsi que la micro-histoire

Temporalités d'analyse

Là on se dit que l'exploration désagrégée c'est quand meme pas mal et que l'on peut étudier les archives autour de moments singuliers

Repenser nos archives vis à vis des moments pivots

Là on commence à parler de la suite, du web que l'on souhaite, de la neutralité et des défis à venir de l'archivage

Chapitre 7

Au Delà Des Archives Web

7.1 Remettre l'humain au cœur des archives

7.2 Fouiller les archives du Web profond

voir el bouquin de mazanès p63

7.3 Les traces nativement numérique

7.4 Vers une sociologie numérique des migrations

Chapitre 8

Ressources

8.1 References

References are placed alongside their citations as sidenotes, as well. This can be accomplished using the normal `\cite` command.[1]

The complete list of references may also be printed automatically by using the `\bibliography` command. (See the end of this document for an example.) If you do not want to print a bibliography at the end of your document, use the `\nobibliography` command in its place.

To enter multiple citations at one location,[2] you can provide a list of keys separated by commas and the same optional vertical offset argument: `\cite{Tufte2006,Tufte1990}`.

`\cite[`⟨*offset*⟩`]{`*bibkey1*`,`*bibkey2*`,...}`

[1] The first paragraph of this document includes a citation.

[2] ; and

8.2 Figures and Tables

Images and graphics play an integral role in Tufte's work. In addition to the standard `figure` and `tabular` environments, this style provides special figure and table environments for full-width floats.

Full page–width figures and tables may be placed in `figure*` or `table*` environments. To place figures or tables in the margin, use the `marginfigure` or `margintable` environments as follows (see figure 8.1):

```
\begin{marginfigure}
\includegraphics{helix}
\caption{This is a margin figure.}
\label{fig:marginfig}
\end{marginfigure}
```

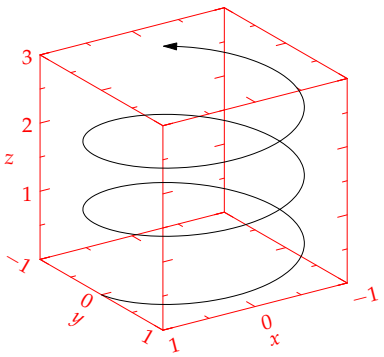

Figure 8.1: This is a margin figure. The helix is defined by $x = \cos(2\pi z)$, $y = \sin(2\pi z)$, and $z = [0, 2.7]$. The figure was drawn using Asymptote (`http://asymptote.sf.net/`).

The `marginfigure` and `margintable` environments accept an optional parameter ⟨*offset*⟩ that adjusts the vertical position of the figure or table. See the "**??**" section above for examples. The specifications are:

```
\begin{marginfigure}[⟨offset⟩]
...
\end{marginfigure}

\begin{margintable}[⟨offset⟩]
...
\end{margintable}
```

Figure 8.2 is an example of the `figure*` environment and figure 8.3 is an example of the normal `figure` environment.

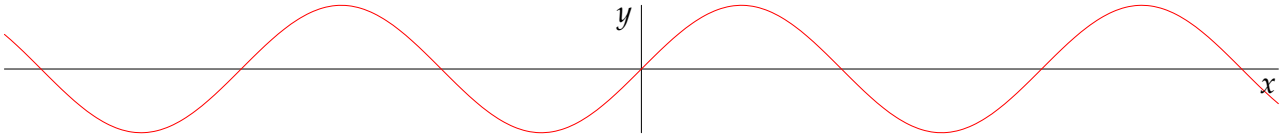

Figure 8.2: This graph shows $y = \sin x$ from about $x = [-10, 10]$. *Notice that this figure takes up the full page width.*

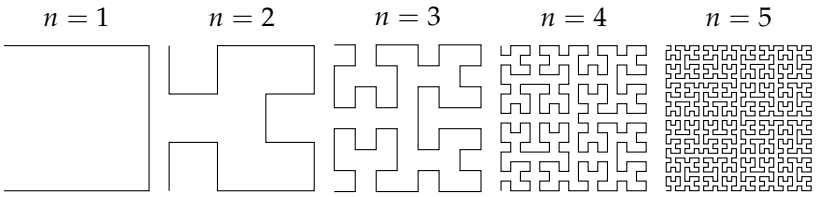

Figure 8.3: Hilbert curves of various degrees n. *Notice that this figure only takes up the main textblock width.*

Chapitre 9

Conclusion

| Bibliographie

Amitay, E., Carmel, D., Herscovici, M., Lempel, R., and Soffer, A. (2004). Trend detection through temporal link analysis. *Journal of the Association for Information Science and Technology*, 55(14):1270–1281.

Arvidson, A., Persson, K., and Mannerheim, J. (2000). The Kulturarw3 Project–The Royal Swedish Web Archiw3e–An Example of" Complete" Collection of Web Pages.

Aturban, M., Nelson, M. L., and Weigle, M. C. (2017). Difficulties of Timestamping Archived Web Pages. *arXiv preprint arXiv:1712.03140*.

Baschet, J. (2018). *Défaire la tyrannie du présent: Temporalités émergentes et futurs inédits*. L'horizon des possibles. Editions La Découverte.

Borges, J. (1974). *Fictions*. Collection Folio. Editions Gallimard.

Borgman, C. L. (2000). Digital libraries and the continuum of scholarly communication. *Journal of documentation*, 56(4):412–430.

Boudrez, F. and Van den Eynde, S. (2002). Archiving websites. *State Archives of Antwerp, Antwerp-Leuven*.

Brügger, N. (2002). Does the materiality of the Internet matter. *The Internet and society*, pages 13–22.

Brügger, N. (2009). Website history and the website as an object of study. *New Media & Society*, 11(1-2):115–132.

De Jong, F., Rode, H., and Hiemstra, D. (2005). Temporal language models for the disclosure of historical text. In *Humanities, computers and cultural heritage: Proceedings of the XVIth International Conference of the Association for History and Computing (AHC 2005)*, pages 161–168.

Derrida, J. (1995). Mal d'archive. *Paris, Galilee*, page 371.

Derrida, J. (2014). *Trace et archive, image et art*. Collection Collège iconique. INA.

Dougnac, M.-T. and Guilbaud, M. (1960). Le dépôt légal: son sens et son évolution.

Finnemann, N. O. (1997). Modernity modernised: The cultural impact of computerisation.

Flusser, V. (2014). *Les gestes*. Cahiers Du Midi. Al Dante Eds.

Foot, K. and Schneider, S. M. (2006). *Web campaigning (acting with technology)*. The MIT Press.

Gomes, D., Miranda, J., and Costa, M. (2011). A survey on web archiving initiatives. In *International Conference on Theory and Practice of Digital Libraries*, pages 408–420. Springer.

Hallgrinsson, B. and Bang, S. (2003). Nordic web archive. In *Proceedings of the 3rd Workshop on Web Archives in conjunction with the 7th European Conference on Research and Advanced Technologies for Digital Libraries (ECDL 2003)*, pages 37–48.

Jatowt, A., Kawai, Y., and Tanaka, K. (2007). Detecting age of page content. In *Proceedings of the 9th annual ACM international workshop on Web information and data management*, pages 137–144. ACM.

Kahle, B. (1997). Preserving the Internet. *Scientific American*, 276:82–83.

Kanhabua, N. and Nørv\a ag, K. (2009). Using temporal language models for document dating. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 738–741. Springer.

Keren, M. (2006). *Blogosphere: The new political arena*. Lexington Books.

Ketelaar, E. (2006). (Dé) Construire l'archive. *Matériaux pour l'histoire de notre temps*, (2):65–70.

Leroi-Gourham, A. (1984). *L'Art des cavernes: Atlas des grottes ornées paléolithiques françaises (Atlas archéologiques de la France) (French Edition)*. Impr. nationale.

Leroi-Gourhan, A. (1964). *Le geste et la parole*. Albin Michel.

Masanès, J. (2006). Web archiving: issues and methods. In *Web Archiving*, pages 1–53. Springer.

Mohr, G., Stack, M., Ranitovic, I., Avery, D., and Kimpton, M. (2004). An Introduction to Heritrix An open source archival quality web crawler. In *In IWAW'04, 4th International Web Archiving Workshop*. Citeseer.

Morsel, J. (2016). Traces? Quelles traces? Réflexions pour une histoire non passéiste. *Revue historique*, (4):813–868.

Nunes, S., Ribeiro, C., and David, G. (2007). Using neighbors to date web documents. In *Proceedings of the 9th annual ACM international workshop on Web information and data management*, pages 129–136. ACM.

Rocco, D., Buttler, D., and Liu, L. (2003). Page digest for large-scale web services. In *E-Commerce, 2003. CEC 2003. IEEE International Conference on*, pages 381–390. IEEE.

Spitz, A., Strötgen, J., and Gertz, M. (2018). Predicting Document Creation Times in News Citation Networks. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 1731–1736. International World Wide Web Conferences Steering Committee.

Stiegler, B. (1991). *Etat de la mémoire et mémoire de l'Etat*, volume 1.

Stiegler, B. (1998). Leroi-Gourhan: l'inorganique organisé. *Les Cahiers de médiologie*, (2):187–194.

Toyoda, M. and Kitsuregawa, M. (2006). What's really new on the web?: identifying new pages from a series of unstable web snapshots. In *Proceedings of the 15th international conference on World Wide Web*, pages 233–241. ACM.

UNESCO (2003). Charter on the Preservation of Digital Heritage.

Van de Sompel, H., Nelson, M., and Sanderson, R. (2013). HTTP framework for time-based access to resource states–Memento. Technical report.

Voerman, G., Keyzer, A., Den Hollander, F., and Druiven, H. (2002). Archiving the Web: Political party Web sites in the Netherlands. *European Political Science*, 2(1):68–75.