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

Chapitre 1 Introduction 13

1.1 Introduction générale 13

1.2 Mise en garde 13

Chapitre 2 Du Web aux Représentations en Ligne des Diasporas 15

2.1 Retour aux origines du Web 15

2.2 Le migrant connecté 15

2.3 Le Web, espace de communication et d'organisation 15

2.4 L'Atlas e-Diasporas 15

Chapitre 3 20 ans d'archivage du Web 17

3.1 Les pionniers 17

3.2 Préserver notre héritage numérique 17

3.3 Constituer des corpus d'archives 17

3.4 Les archives Web de l'Atlas e-Diasporas 17

Chapitre 4 Traces Discrétisées et Temporalité Figée 19

4.1 Détruire pour mieux archiver 19

4.2 Un temps sans extension 19

4.3 Construire un moteur d'exploration d'archive 21

4.4 Les archives sont des traces indirectes du Web 21

Chapitre 5 Fragmenter les Archives Web 23

5.1 Vers une nouvelle unité d'exploration 24

5.2 Le fragment Web : définition
5.3 Scraping et méthodologie d'extraction 32
5.4 Penser une exploration désagrégée 32
5.5 Intégration à un moteur d'exploration 32

Chapitre 6 | Explorations de Collectifs Migrants Éteints 33
6.1 À la recherche de l'étonnement : l'analyse exploratoire de données 33
6.2 Les traces d'une mutation numérique 33
6.3 Un soulèvement en ligne éphémère 34
6.4 Les Moments Pivot du Web 35

Chapitre 7 | Au Delà Des Archives Web 37
7.1 Remettre l'humain au cœur des archives 37
7.2 Fouiller les archives du Web profond 37
7.3 Les traces nativement numérique 37
7.4 Vers une sociologie numérique des migrations 37

Chapitre 8 | Ressources 39
8.1 References 39
8.2 Figures and Tables 39

Chapitre 9 | Conclusion 41

Chapitre Bibliographie 43

List of Figures

4.1 "Boulevard du Temple", Louis Daguerre, 1838 20
4.2 Warc vs Daff 20

5.1 Extraits de "*Je Vous Salue, Sarajevo*", J.L. Godard (1993) à partir d'une photographie de R. Haviv (1992) 25
5.2 Les 5 strates analytiques du Web, d'après (Brügger, 2009) 27
5.3 Répartition des archives de *yabiladi.com* dans la WayBack Machine (`https://web.archive.org/web/*/www.yabiladi.com`) 28
5.4 Date de création (rouge) d'un post de forum sur *yabiladi.com* 28
5.5 Dimensions d'exploration des archives Web (ajout de l'acteur) 30

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/`). 39
8.2 This graph shows $y = \sin x$ from about $x = [-10, 10]$. *Notice that this figure takes up the full page width.* 40
8.3 Hilbert curves of various degrees n. 40

List of Tables

4.1 Échelle de datation d'une page Web archivée 19

5.1 Échelle (actualisée) de datation d'une page Web archivée 29

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

2.4 L'Atlas e-Diasporas

Chapitre 3

20 ans d'archivage du Web

3.1 Les pionniers

Internet Archive et le pre-Unesco

3.2 Préserver notre héritage numérique

L'unesco et faire des archives un commun Un tour du monde des initiatives La constitution juridique des corpus en france Et l'état de l'archivage aujourd'hui (fin de Internet mémory et les rogues archivistes)

3.3 Constituer des corpus d'archives

Méthodologie d'acquisition

là on parle des changements dans les pages

Où l'on fait le tour de l'état de l'art en matière de création d'archives Web, de crawl, etc ...

le web n'est plus le web (brugger) le web s'archive lui même et l'archive ne capture que le front end

Un format unique ?

Où l'on parle du WARC (et de ces prédécesseurs) vs le DAFF

3.4 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

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

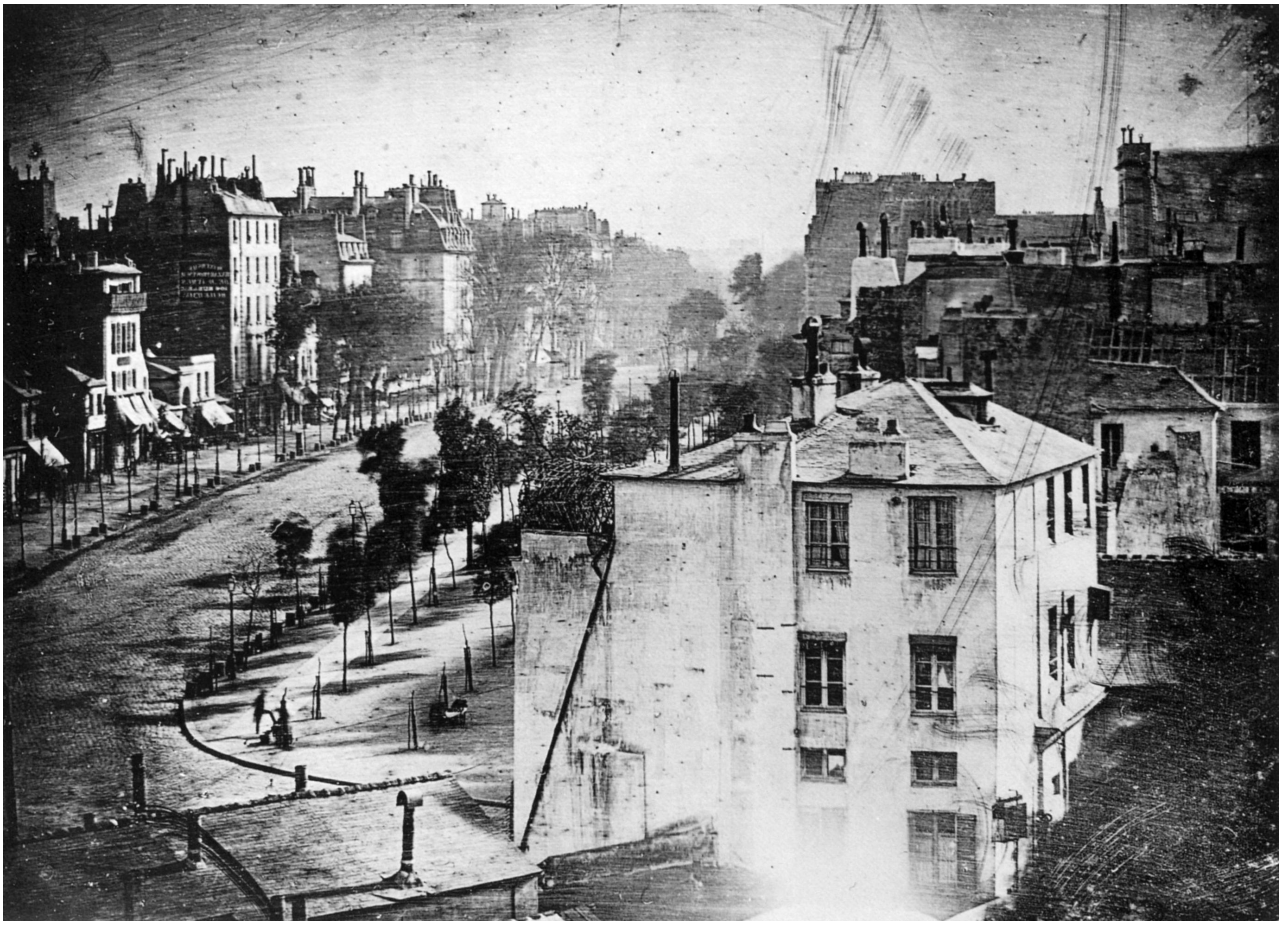

Figure 4.1: "Boulevard du Temple", Louis Daguerre, 1838

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

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]).

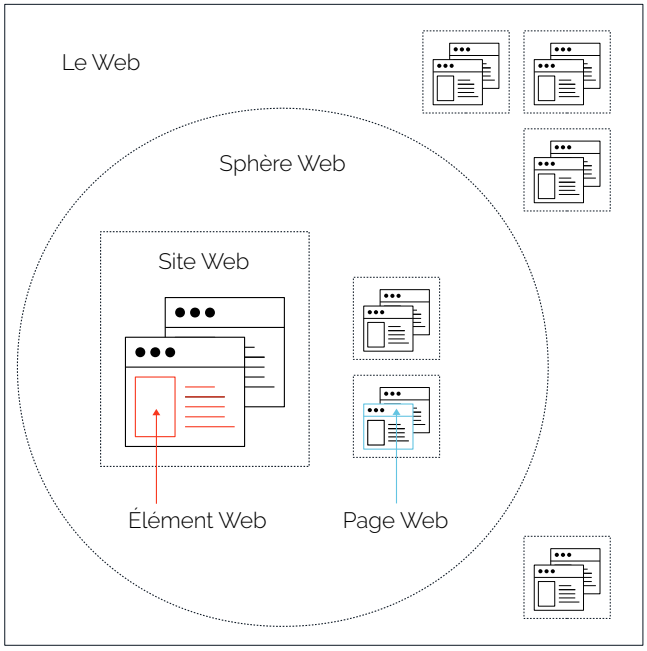

Figure 5.2: Les 5 strates analytiques du Web, d'après (Brügger, 2009)

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 ?

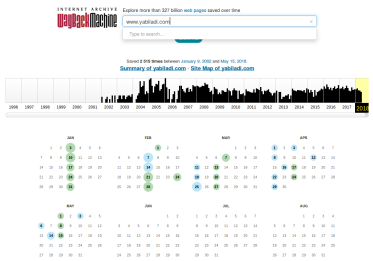

Figure 5.3: Répartition des archives de *yabiladi.com* dans la WayBack Machine (`https://web.archive.org/web/*/www.yabiladi.com`)

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.

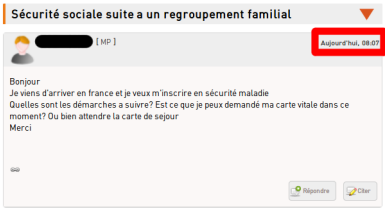

Figure 5.4: Date de création (rouge) d'un post de forum sur *yabiladi.com*

Une page Web évolue (Section 3.3) 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	
page	création	Validité historique ↓
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.

Du geste du crawler au geste de l'acteur

Bien dater une page archivée participe de son émancipation vis à vis du crawler et 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 publié en ligne. Derrière les dates d'édition des fragments Web, peuvent transparaitre les gestes de divers acteurs : auteurs, lecteurs ou visiteurs des pages et sites archivés. Pour le philosophe V. Flusser, un geste est une série de mouvements significatifs dont le but est déchiffrable[13] (Flusser, 2014). Comme pour la date d'édition, un auteur laisse des

[13] Le geste par le déchiffrement de sa signification devient ainsi objet d'études : "*les gestes montrent la façon dont nous sommes au monde*", (Flusser, 2014), p.319.

indices de sa présence sur une page Web, signant certains messages sous pseudonymes. Le geste de l'acteur devient ainsi objet d'études et une dimension à par entière de l'exploration des archives Web. Ou, comme le suggère J. Morsel, la possibilité d'écrire une histoire *symptomale* (Morsel, 2016). Alors que la trace implique l'absence de l'acteur qui l'a produite, le symptome suppose la présence l'attente de l'acteur (coprésent à ce dont il est le signe). L'archive renferme "*présence symptomal fossilisée*" de l'acteur.

Certains fragments Web auront pour l'historien ou l'explorateur d'archives une valeur particulière si tant est qu'il témoigne directement de l'acte d'un acteur donné. Cette nouvelle perspective pourra nous mener 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

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}
```

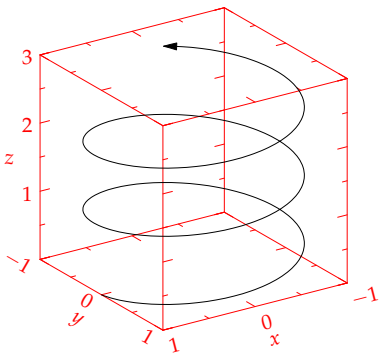

Figure 8.1: This is a margin figure. The helix is defined by $x = \cos(2\pi z)$, $y = \sin(2\pi z)$, and $z = [0, 2.7]$. The figure was drawn using Asymptote (`http://asymptote.sf.net/`).

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.

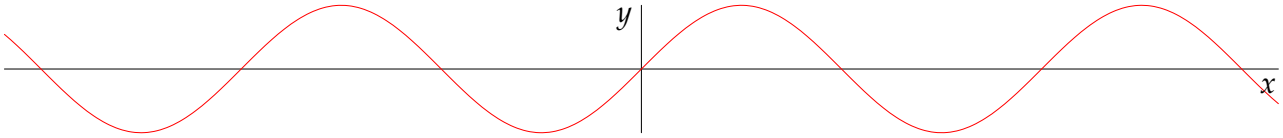

Figure 8.2: This graph shows $y = \sin x$ from about $x = [-10, 10]$. *Notice that this figure takes up the full page width.*

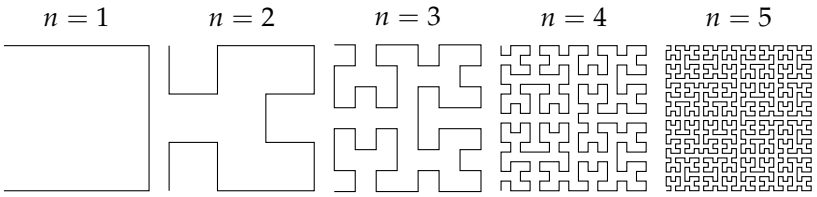

Figure 8.3: Hilbert curves of various degrees n. *Notice that this figure only takes up the main textblock width.*

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.

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.

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.

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.

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.

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.

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.

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.

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