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

Introduction générale

Ici l'intro de la thèse.

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

Retour aux origines du Web

Le migrant connecté

Le Web, espace de communication et d'organisation

L'Atlas e-Diasporas

Chapitre 3

| 20 ans d'archivage du Web

Les pionniers

Internet Archive et le pre-Unesco

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)

Constituer des corpus d'archives

Méthodologie d'acquisition

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

Un format unique ?

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

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

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

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

Crawl blindness

Cohérence

Duplicata

Construire un moteur d'exploration d'archive

Extraction et enrichissement

Définition du schéma d'indexation

Détection d'événements

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 désagrégation 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* sont indissociables des archives Web telles que nous les connaissons. Liés organiquement à la structure même des divers formats d'archivage, ils en sont les artéfacts directs et demeurent des obstacles majeurs pour qui souhaite mener à bien une exploration du Web archivé. S'il est tout à fait possible, depuis la Way-Back Machine par exemple, de différencier *à la main* la date d'archivage d'une page Web visitée de la date supposée de sa création ou de la date de publication réelle de tel ou tel contenu, cette tâche deviendra rapidement fastidieuse voire titanesque à mesure que grandira notre périmètre d'exploration. En associant les contenus archivés à la temporalité seule du crawler (voir **Figure X**), les archives Web se détachent de réalité historique des sites et des pages Web dont elles sont censées être le reflet fidèle. Ainsi, un site Web pourra être sur-représenté (ou inversement sous-représenté : voir **Figure X**) au sein d'un corpus d'archives au cours d'une période potentiellement plus liée à la programmation du crawler qu'à l'activité effective du site et des ses pages.

L'exploration fine et complète de notre corpus d'archives sous-tend à toute compréhension de l'évolution de l'e-Diasporas marocaine telle que nous la formulions en Introduction. Tout l'enjeu de ce chapitre sera de déterminer un moyen d'affranchir les archives Web du lien qui les unissent aux crawlers. Il s'agira ici de retrouver, dans le code HTML des corpus archivés, les indices de la création d'une page ou de la publication d'un contenu et d'articuler notre analyse autour de ces nouvelles données : descendre au delà du niveau des Web.

L'essentiel de ce chapitre sera donc consacré à ce déplacement, à ce changement d'échelle analytique (ou de strate au sens de N. Brügger [Brügger, 2009]) et à ces conséquences. Au geste du crawler, nous substituerons le geste de l'auteur et du lecteur dont le passage sur les pages Web archivés aura laissé des traces qu'il nous faudra exploiter. Sur ce point, nous proposerons l'introduction d'une nouvelle unité d'exploration des archives Web : le *fragment Web*. Nous présenterons, dans un premier temps, la genèse et les inspirations qui sous-tendent et motivent la définition du *fragement Web*. Puis, nous revien-

drons en miroir sur les modalités techniques et théoriques d'un moteur d'exploration d'archives Web telles que présentées dans le Chapitre 4, mais cette fois-ci, revisitées à l'aune du *fragment Web*. Cette mise à niveau servira de base aux explorations présentées plus avant dans le Chapitre 6.

5.1 Vers une nouvelle unité d'exploration

Au delà de la simple volonté de s'affranchir des limitations du format actuel des archives Web, notre proposition de fragmenter les corpus d'archives est portée par l'idée selon laquelle toute archive est une matière destinée à être déconstruire, désagréger ou ré-arrangée en vue de la questionner et d'inventer l'histoire. Une archive ne parle pas seule [Ketelaar, 2006], elle n'est jamais fermée, jamais complète, mais se tient toujours prête à être réinterprétée par une nouvelle génération. Mais gardons néanmoins à l'esprit, comme le rappel J. Derrida[1], que le document d'origine (dans notre cas la page Web archivée) 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.

Le *fragment Web*, tel que nous le présenterons dans la suite de ce chapitre, ne sera donc pas une version modifiée d'une page Web archivée, mais bien une nouvelle entité issue de la page Web déconstruite et utilisable en parallèle des modèles d'exploration d'archives déjà existants.

[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"*

Jacques Derrida. Mal d'archive. *Paris, Galilee*, page 371, 1995

Découper, déplacer, monter

Nous évoquions déjà, dans le chapitre précédent, le personnage de Funes imaginé par J. L. Borges qui, dans la fable, est condamné à ne plus jamais rien oublier au détriment de ses propres capacités à penser. Funes se redécouvre ainsi sans cesse, n'arrivant pas à se créer des souvenirs, à se raconter de mémoire sa propre histoire[2]. Pour mémoriser 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 film de nos histoires telles que nous nous en rappelons. En cela, la posture de l'historien face à un document archivé doit pouvoir se rapprocher de celle du monter de cinéma face à une matière filmée. Leurs outils sont semblables. Lorsqu'il invente l'histoire, l'historien déconstruit, isole ou rapproche les archives entre elles. Dans son court métrage de 1993 *"Je Vous Salue, Sarajevo"*[3], J. L. Godard fragmente, isole et fait se confronter des inserts éclatés d'une photographie du reporteur de guerre R. Haviv à la manière d'un collage-poème ou d'un cinétract[4]. 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,

[2] " *Ils nous laissent entrevoir ou déduire le monde vertigineux de Funes. Celui-ci, ne l'oublions pas, était presque incapable d'idées générales, platoniques. [...] Son propre visage dans la glace, ses propres mains, le surprenaient chaque fois.*"

J.L. Borges. *Fictions*. Collection Folio. Editions Gallimard, 1974

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

[4] Voir `https://fr.wikipedia.org/wiki/Cin%C3%A9tract`

une ville de son temps. Le film finit par dévoiler entière, l'image dans toute son horreur. Décomposer pour mieux recomposer.

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.

Ceux sont ces outils de montage que nous souhaitons, par l'intermédiaire du *fragment Web* donner aux explorateurs d'archives Web. Avoir en main des éléments fragmentés de pages Web éloignés, que nous pourrions rapprocher à souhait afin de traiter plus largement d'un moment particulier de l'histoire du Web.

Les strates d'analyse du Web

Les strate du web ou la nécessité de descendre sous le niveau de la page Web

La question de la datation d'une page archiver

Ici on introduit une échelle de datation et on se questionne sur la meilleur façon de dater une archive web

Désagréger pour changer de temporalité

Là on fait l'expérience edition date vs crawler date et l'on découvre que le fragment web peut nous permettre d'échapper à la tyrannie du crawl et à structure temporelle linéaire. Et Godard et

Le fragment Web : 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 :

Scraping et méthodologie d'extraction

Extraire de l'information issue d'une page Web

Là on parle de scraping et on fait une revue de l'état de l'art et l'on parle de readability ...

Implémentation technique

Là on parle de rivelaine et de la fonction distance ...

Exemples et discussions

Là on parle de l'automatique vs le fait à la main avec le truc firefox

Penser une exploration désagrégée

Atténuer les "crawl blindness"

Cohérence relative entre archives

Dédupliquer les corpus

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

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

Les traces d'une mutation numérique

D'une communauté vibrante de blogs ...

Là on raconte l'état des blogs en 2008

... 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

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

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

Remettre l'humain au cœur des archives

Fouiller les archives du Web profond

Les traces nativement numérique

Vers une sociologie numérique des migrations

Chapitre 8

Ressources

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

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}
```

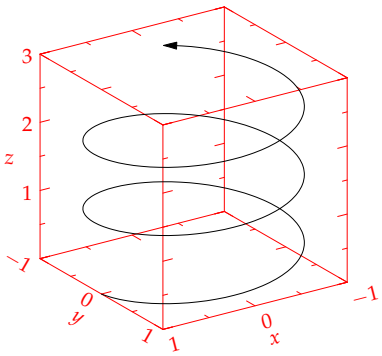

Figure 8.1: This is a margin figure. The helix is defined by $x = \cos(2\pi z)$, $y = \sin(2\pi z)$, and $z = [0, 2.7]$. The figure was drawn using Asymptote (`http://asymptote.sf.net/`).

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.

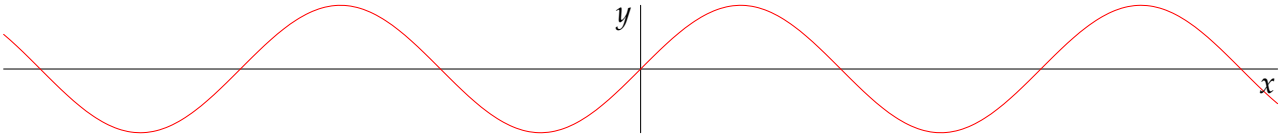

Figure 8.2: This graph shows $y = \sin x$ from about $x = [-10, 10]$. *Notice that this figure takes up the full page width.*

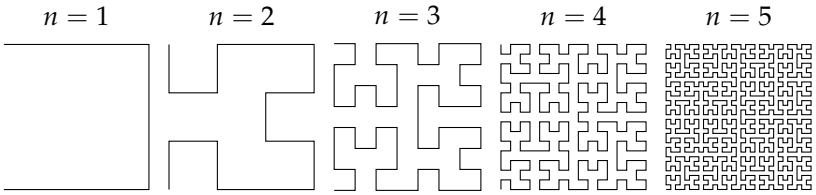

Figure 8.3: Hilbert curves of various degrees n. *Notice that this figure only takes up the main textblock width.*

Chapitre 9

Conclusion

| Bibliography

J. Baschet. *Défaire la tyrannie du présent: Temporalités émergentes et futurs inédits*. L'horizon des possibles. Editions La Découverte, 2018. ISBN 978-2-7071-9734-4.

J.L. Borges. *Fictions*. Collection Folio. Editions Gallimard, 1974.

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

Jacques Derrida. Mal d'archive. *Paris, Galilee*, page 371, 1995.

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