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 désagragation se fait dans le modèle de données mais également dans la façon de conduire sont exploration.

Chapitre 5

| Fragmenter les Archives Web

Où l'on parle des différentes strates du Web, de l'intuition de descendre sous le niveau de la page, de la définition du fragment Web, de la méthodologie d'extraction et de la redéfinition d'un modèle d'analyse désagrégé des archives Web

Vers une nouvelle unité d'exploration

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.

Le fragment Web

Là c'est la définition

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] Edward R. Tufte. *Beautiful Evidence*. Graphics Press, LLC, first edition, May 2006. ISBN 0-9613921-7-7; and Edward R. Tufte. *Envisioning Information*. Graphics Press, Cheshire, Connecticut, 1990. ISBN 0-9613921-1-8

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}
```

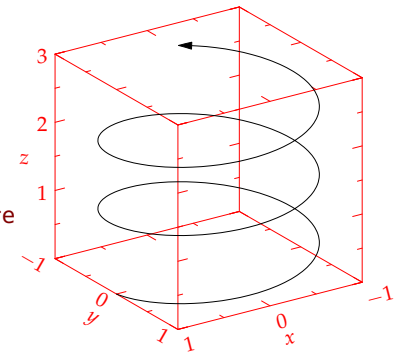

Figure 8.1: This is a margin figure. The helix is defined by $x = \cos(2\pi z)$, $y = \sin(2\pi z)$, and $z = [0, 2.7]$. The figure was drawn using Asymptote (`http://asymptote.sf.net/`).

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.

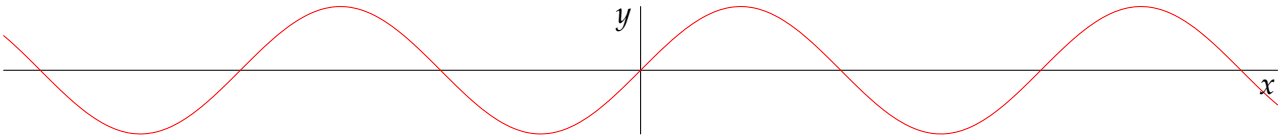

Figure 8.2: This graph shows $y = \sin x$ from about $x = [-10, 10]$. *Notice that this figure takes up the full page width.*

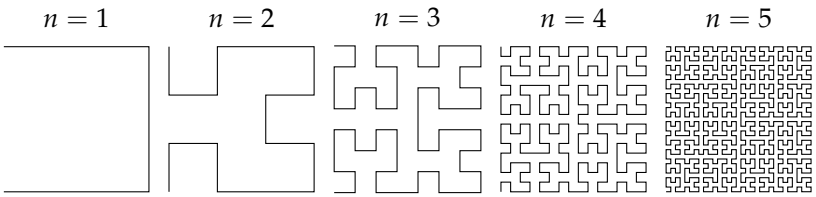

Figure 8.3: Hilbert curves of various degrees n. *Notice that this figure only takes up the main textblock width.*

Chapitre 9

Conclusion

Bibliography

Edward R. Tufte. *Envisioning Information*. Graphics Press, Cheshire, Connecticut, 1990. ISBN 0-9613921-1-8.

Edward R. Tufte. *Beautiful Evidence*. Graphics Press, LLC, first edition, May 2006. ISBN 0-9613921-7-7.