HAL
open science

Archives, fragments Web et diasporas. Pour une exploration désagrégée de corpus d'archives Web liées aux représentations en ligne des diasporas

Quentin Lobbé

▶ **To cite this version:**

Quentin Lobbé. Archives, fragments Web et diasporas. Pour une exploration désagrégée de corpus d'archives Web liées aux représentations en ligne des diasporas. Web. Université Paris-Saclay, 2018. Français. NNT : . tel-01988596v1

HAL Id: tel-01988596

https://inria.hal.science/tel-01988596v1

Submitted on 27 Dec 2018 (v1), last revised 21 Jan 2019 (v2)

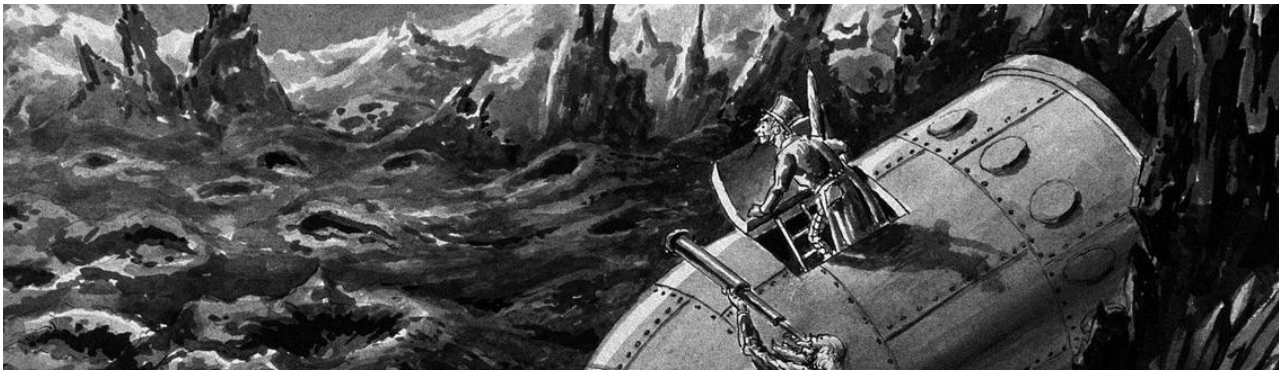
HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Quentin Lobbé

Archives, fragments Web et diasporas

Pour une exploration désagrégée de corpus d'archives Web liées aux représentations en ligne des diasporas



Thèse de doctorat de l'Université Paris-Saclay, préparée à Télécom ParisTech, école doctorale des Sciences et Technologies de l'Information et de la Communication (STIC) pour la spécialité Données, Connaissances, Apprentissage et Interaction.

Présentée par **Quentin Lobbé** le 9 novembre 2018
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, enseignante chercheuse à Télécom ParisTech

Soutenue publiquement à Paris le 9 novembre 2018, devant un jury composé de :
Bruno Bachimont (Rapporteur), professeur à Sorbonne Université
Marc Spaniol (Rapporteur), professeur à l'Université de Caen Basse-Normandie
Anat Ben-David, professeure à l'Open University of Israel
Valérie Schafer, professeure à l'Université du Luxembourg
Bruno Defude, professeur à Télécom SudParis
Julien Masanès, CEO à Internet Memory Research

Titre : Archives, fragments Web et diasporas. Pour une exploration désagrégée de corpus d'archives Web liées aux représentations en ligne des diasporas

Mots clés : archives Web, fragments web, diasporas, traces numériques des migrations

Résumé : Le Web est un environnement éphémère. Alors que de nouveaux sites Web émergent chaque jour, il arrive que certaines communautés disparaissent entièrement de la surface de la toile, ne laissant derrière elles que des traces incomplètes voire inexistantes. Face à la volatilité du Web vivant, plusieurs initiatives d'archivage cherchent malgré tout à préserver la mémoire du Web passé. Mais aujourd'hui, force est de constater qu'un mystère demeure : Pourquoi, alors qu'elles n'ont jamais été aussi vastes et aussi nombreuses, les archives Web ne font-elles pas déjà l'objet de multiples recherches historiques ? Initialement construites pour inscrire la mémoire de la toile sur un support durable, ces archives ne doivent pourtant pas être considérées comme une représentation fidèle du Web vivant. Elles sont les traces directes des outils de collecte qui les arrachent à leur temporalité d'origine. Partant de là, cette thèse ambitionne de redonner aux chercheurs les moyens théoriques et techniques d'une plus grande maniabilité du Web passé, en définissant une nouvelle unité d'exploration des archives Web : le fragment Web, un sous-ensemble cohérent et auto-suffisant d'une page Web. Pour ce faire, nous nous inscrirons dans l'héritage des travaux pionniers de l'Atlas e-Diasporas qui permit, dans les années 2000, de cartographier et d'archiver plusieurs milliers de sites Web migrants. Source principale de données à partir desquelles nous déploierons nos réflexions, c'est à travers l'angle particulier des représentations en ligne des diasporas que nous chercherons à explorer les archives Web de l'Atlas.

~

Title : Archives, Web fragments and diasporas. For a disaggregated exploration of web archives related to online representations of diasporas

Keywords : Web archives, Web fragments, diasporas, digital traces of migrations

Abstract : The Web is an unsteady environment. As Web sites emerge every days, whole communities may fade away over time by leaving too few or incomplete traces on the living Web. Facing this phenomenon, several archiving initiatives try to preserve the memory of the Web. But today, a mystery remains : While they have never been so vast and numerous, why are the Web archives not already the subject of many historical researches ? In reality, Web archives should not be considered as a faithful representation of the living Web. In fact, they are the direct traces of the archiving tools that tear them away from their original temporality. Thus, this thesis aims to give researchers the theoretical and technical means for a greater manageability of the Web archives, by defining a new unit of exploration : the Web fragment, a coherent and self-sufficient subset of a Web page. To that end, we will follow the pioneering work of the e-Diasporas Atlas which allowed, in the 2000s, to map and archive thousands of migrant Web sites. Main source of data from which we will unfold our reflections, it is through the particular angle of online representations of diasporas that we will explore the Web archives of the Atlas.

« 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

Merci à Dana et à Pierre qui ont été ma boussole bicéphale à travers cette odyssée, me guidant avec enthousiasme, confiance et douceur des plages endiablées de L.A. aux collines venteuses de Jérusalem ; aux membres du jury et aux rapporteurs pour leurs retours précieux et avisés sur le texte que vous tenez entre vos mains ; à Stéphane sans qui je ne me serais jamais lancé sur cette voie, me destinant plutôt à yoyoter indéfiniment rue du Sentier ; à Thomas, Jérôme, Claude et Zeynep pour m'avoir initié à l'art secret de l'archivage du Web ; à Anat, Adam, Vered et Tzipy qui m'ont donné l'impulsion heureuse de l'écriture il y a un an déjà ; à Mathieu, maître bâtisseur, maître penseur ; à Muriel et Christophe devenus au fil du temps experts ès *plus gros serveur* ; à Johan, Katja, Nasr, Ahmed, Bence et Jorieke qui ont permis à l'autre Calais d'exister par le tracé ; à Chaima, Marie et à l'ensemble de mes élèves des mois de juin dont les contributions au Chapitre 6 furent essentielles ; à Max boss final de la `C212-C214` ; à Hana, Atef, Achref et Mohammad pour avoir tenu à faire de notre bureau partagé l'une des meilleures pâtisseries du quartier ; à Antoine, Albert, Jean-Louis, Mauro et Fabian, sages parmi les sages, piliers du labo ; à Camille, Mostafa, Katerina, Marc et Heitor qui savent ce que cela fait d'être de l'autre côté, (véritables) piliers du labo ; à mes chers collègues doctorants et non moins grands amis : Marie, Maroua, Miyoung, Oana, Hélène, Léa, Pierre-Alexandre, Mikaël, Luis, Julien, Jacob, Jonathan, Thomas und Thomas, derniers remparts face au recul de la science et ((véritables) véritables) piliers du labo ; à mon gang d'indéboulonnables casse-cous : à Louise et Perrine mes sœurs d'âme, à Igor et Noé mes fiers cannistes, ainsi qu'aux compagnon.e.s d'aventures gérômoises ; à ma grand-mère, à ma meilleure mère et à ma meilleure sœur qui forment cette sainte trinité maison capable, à elle seule, de m'offrir jusqu'à plus soif l'inestimable dose de bonheur nécessaire à mon quotidien.
Et puisque le voyage arrive maintenant à son terme, je dédie cette thèse à ceux qui ne sont plus là pour en voir la fin : à mon grand-père, à mon père.

Table des matières

Introduction 9

Chapitre 1 Les Représentations en ligne des diasporas 17

1.1 *Généalogie du Web* 18

1.2 *Dispersés mais présents sur la toile* 26

1.3 *L'Atlas e-Diasporas : cartographier les sites migrants* 32

Chapitre 2 Préserver le Web et les e-Diasporas 43

2.1 *Vingt ans d'archivage du Web* 44

2.2 *Sélectionner, collecter et consulter les corpus* 53

2.3 *Les archives Web de l'Atlas e-Diasporas* 67

Chapitre 3 Traces discrétisées et temporalité figée 75

3.1 *Détruire pour mieux archiver* 76

3.2 *Un temps sans durée* 78

3.3 *Construire un moteur d'exploration d'archives Web* 82

3.4 *Les archives ne sont pas des traces directes du Web* 91

Chapitre 4 Fragmenter les archives Web 95

4.1 *En dessous des pages Web* 96

4.2 *Le fragment Web : définition* 105

4.3 *Scraping et méthodologie d'extraction* 105

4.4 *Penser une exploration désagrégée* 118

4.5 *Intégration au moteur d'exploration* 123

Chapitre 5 | Exploration de collectifs migrants éteints 131

5.1 À la recherche de l'étonnement 132

5.2 Les traces d'une mutation numérique 136

5.3 Un soulèvement en ligne éphémère 151

5.4 Mener des recherches historiques depuis les archives Web 174

Chapitre 6 | Au Delà des Archives Web 183

6.1 Les archives de navigation Web de la BPI 184

6.2 Les dynamiques de l'accueil des réfugiés (2015-2017) 194

Conclusion 207

Bibliographie 211

Introduction

Le Web est un environnement instable, éphémère. Né il y a moins de 30 ans, il se présente tout autant comme un flux continu d'information qu'un territoire en perpétuelle expansion. Alors que de nouveaux sites Web émergent jour après jour, il arrive régulièrement que des communautés disparaissent entièrement de la surface de la toile, ne laissant derrière elles que des traces incomplètes voire inexistantes.

Face à la volatilité du Web vivant, plusieurs initiatives d'archivage cherchent malgré tout à préserver la mémoire du Web passé. En effet, depuis le milieu des années 90, des pionniers archivent régulièrement une part significative des ressources publiées en ligne (des pages, des liens hypertextes, du contenu multimédia ...), permettant ainsi la sauvegarde de cet héritage numérique pour les générations futures. Mais aujourd'hui, force est de constater qu'un mystère demeure : Pourquoi, alors qu'elles n'ont jamais été aussi vastes et aussi nombreuses, les archives Web ne font-elles pas déjà l'objet de multiples recherches historiques ? Pourquoi l'usage des archives ne s'est-il-pas démocratisé ?

Cet état de fait est à ce point paradoxal que le Web est devenu une composante essentielle de nos vies quotidiennes. Le recours à ses archives devrait être, sinon un reflex, au moins une possibilité raisonnable pour qui souhaite interroger les traces du Web passé : Quel a été le devenir de tel ou tel collectif en ligne au cours des dix dernières années ? Pourquoi certains groupes d'internautes ont-ils-été impactés par un événement socio-politique donné ? De quelle manière cette technologie s'est-elle imposée sur la toile malgré la concurrence en place ? etc.

Les sujets de recherche ne manquent pas et l'intérêt scientifique pour les archives Web reste vivace. En effet, ces corpus sont de véritables mines d'or ne demandant qu'à être explorées par des historiens, par des sociologues ou par de simples curieux. Mais le chemin pour y arriver est semé d'embuches et il existe, encore aujourd'hui, des freins majeurs à la possibilité de jeter un regard en arrière sur le Web et sur son histoire.

Le problème, selon nous, se niche au sein même de la structure des corpus archivés et se répercute en cascade dans les différents systèmes

d'exploration existants. À l'origine, les archives Web furent construites pour inscrire la mémoire de la toile sur un support durable. Pourtant, ces archives ne doivent pas être considérées comme une représentation fidèle du Web vivant. Elles sont, malgré elles, les traces directes des outils de collecte programmés par les archivistes.

En effet, l'unité principale d'exploration des collectages reste la page Web. Tous les systèmes d'archivage sont, ainsi, construits au-dessus des pages qu'ils capturent et qu'ils associent à une date de téléchargement. De fait, cette date devient le seul marqueur temporel permettant d'interroger les corpus préservés. Arrachées à leur temporalité d'origine, les archives Web nous reviennent alors comme des objets discrétisés et figés, sans lien direct avec la réalité du Web vivant dont elles sont, pourtant, censées être le reflet.

Ceci étant exposé, nous voulons, dans cette thèse, redonner aux chercheurs les moyens théoriques et techniques d'une plus grande maniabilité du Web passé, et ce, en définissant une nouvelle unité d'exploration des archives Web : le fragment Web, un sous-ensemble cohérent et auto-suffisant d'une page Web.

Le fragment Web a la particularité d'être associé à une date d'édition (c.-à-d., la date a laquelle il a été mis en ligne) plutôt qu'aux seules dates de téléchargement. Ce faisant, les archives Web quittent, enfin, le temps des robots collecteurs pour retrouver le temps du Web tel qu'il a été. Et dans un même mouvement, c'est l'humain que nous cherchons à replacer au cœur de l'analyse des archives en témoignant directement, grâce au fragment Web, du geste des personnes qui auront, par le passé, publié, partagé ou commenté un contenu en ligne.

Aussi, le fragment Web introduit une nouvelle grammaire analytique qui permet aux chercheurs d'explorer les archives Web de manière désagrégée et libérée de la pesanteur des formats existants. Les corpus collectés passent alors sur la table de montage, où ils sont découpés, déplacés et mis en relation avec d'autres éléments épars du Web passé. Cette souplesse méthodologique ouvre la voie à des explorations à grande échelle, susceptibles de retracer l'histoire en ligne et hors-ligne d'un moment singulier. Et s'il semble ambitieux de vouloir traiter ici de l'ensemble des mutations du Web ou des transformations qu'il a induites dans nos sociétés, nous pensons néanmoins pouvoir nous focaliser sur un problème plus générique : comment rendre compte, au présent, d'un événement et de sa genèse en le restituant dans la double temporalité du Web et du réel ?

Cependant les enjeux soulevés par l'exploration des archives du Web dépassent bien souvent le cadre strict d'une seule et unique discipline. Aussi, pour interroger l'histoire du Web passé, il ne faudra pas hésiter à croiser les regards, les perspectives de recherches et à s'ouvrir aux controverses. Ainsi, bien que soutenue en informatique, la présente

thèse se situera, par moment, aux frontières de l'histoire et de la sociologie.

Pour le chercheur en informatique, la nature même des corpus d'archives Web est, en soi, un défi de taille. Leur volume et leur étendue posent un problème fondamental d'ingénierie. Comment trouver efficacement, parmi plusieurs millions de versions de pages archivées, l'unique document correspondant à une requête donnée ?

Souvent trop larges et trop éparses, les archives Web peuvent manquer de cohérence interne voire thématique, et, malheureusement, ces soucis sont difficilement résolubles a posteriori. En effet, un chercheur doit souvent faire face à un corpus déjà constitué ou terminé (cela sera le cas de cette thèse). Il se contente alors des archives telles qu'il les découvre sans pouvoir influer directement sur la forme des collectages.

Là encore, ce qui sous-tend toutes nos interrogations est une réflexion profonde sur la structure des archives Web. Quels problèmes posent les formats courants d'archivage ? Doit-on explorer les archives en l'état ? Ou peut-on imaginer une ou plusieurs transformations capables de faciliter leur exploration ?

Pour l'historien, les corpus d'archives Web demeurent des espaces encore majoritairement inexplorés (tout au moins à grande échelle). Aussi, peut-on imaginer que le Web passé soit déjà si distant de nous qu'il nous serait aujourd'hui impossible de l'appréhender dans toute sa complexité ? Existe-il des phénomènes de conservation différentielle qui entraîneraient une sous-représentation de certains types de sites Web ? N'oublions pas qu'une forme d'obsolescence technique rend désormais très compliquée la reconstruction à l'identique du Web des années 90 (par exemple) et, par là même, son analyse *in situ*.

Ce faisant, il s'agit peut-être pour nous de considérer un site ou un ensemble de pages archivées comme un chantier de fouilles à part entière. Comment étudier l'histoire d'un site Web en avançant par strates et couches successives jusqu'à remonter à ses origines et à sa mise en ligne initiale ? Se pose alors la question des indicateurs : La trace que nous avons sous les yeux est-elle historiquement significative ? Comment la dater ? Dans quel contexte a-elle été produite puis sauvegardée ? Et inversement, qu'elles ressources n'ont pas été (en conscience ou non) préservées par les archivistes ?

Pour le sociologue, explorer un corpus d'archives Web est l'occasion d'étudier les multiples dynamiques à l'œuvre sur la toile. En effet, mettre un site en ligne, l'alimenter et le faire vivre est une forme de présence tout à fait singulière sur le Web. Là où d'autres naviguent de page en page et réagissent face à du contenu déjà existant, certaines

personnes choisissent plutôt de mettre en ligne des ressources originales. Ces espaces créés de toutes pièces deviennent des lieux d'échanges, de connections et de partages. Ce faisant, il existe un geste singulier à interroger derrière chaque site Web. Aussi, ces derniers doivent être considérés comme des acteurs évoluant au sein de réseaux d'internautes, de pages Web, de robots d'indexation, d'institutions...

Par exemple, la dynamique des liens de citation produits sur le Web est une piste de réflexion pour qui souhaite questionner une forme de géopolitique de l'hypertexte. Quels enjeux hors-ligne déterminent un lien de reconnaissance en ligne ? Comment un acteur conserve-t-il une position d'autorité au sein d'un collectif de sites nouvellement créé ? Quelle forme d'organisation *IRL* [1] préfigure la naissance d'une communauté d'internautes ? etc.

1. Expression anglaise (*in real life*) désignant généralement la vie en dehors du Web.

Finalement, la question est, ici, de faire du Web un espace de recherche capable d'enrichir une enquête qualitative ou d'être lui même renforcé par une série d'entretiens. Comment faire se confronter et mettre en relation le récit d'une personne donnée avec ses traces Web telles qu'elles ont été archivées ? Peut-on réorienter l'exploration automatique d'un corpus d'archives Web sur la base d'indices issus d'observations de terrain ?

Sur ce dernier point et de manière transversale à tous les enjeux précédemment cités, demeure la question de l'interdisciplinarité qu'il nous faudra prendre à bras-le-corps. En effet, il s'agira pour cette thèse de se positionner dans un dialogue constant entre chaque discipline afin de conjuguer ensemble les problématiques des unes et des autres. Notre méthodologie d'exploration des archives Web se verra ainsi fortement influencée par ce besoin de pouvoir passer, à volonté ou suivant les contraintes du moment, d'un ensemble de traitements automatiques à grande échelle vers une série de validations focalisées et manuelles.

Pour ce faire, nous nous inscrirons dans l'héritage des travaux pionniers de l'Atlas e-Diasporas (Diminescu, 2012a) qui mobilisa, à la fin des années 2000, des ingénieurs, des sociologues et des historiens autour de l'étude des formes de représentation en ligne des diasporas. L'Atlas, dirigé et conçu par D. Diminescu, permit l'observation et la cartographie de plus de 9000 sites Web migrants [2]. Créés et maintenus par des personnes migrantes, ces sites se connectaient les uns aux autres par des liens de citation hypertexte formant, ainsi, de véritables réseaux : les e-Diasporas, des collectifs migrants en ligne qui s'organisaient d'abord et avant tout sur le Web.

2. `http://www.e-diasporas.fr/`

Mais devant faire face à la disparition totale ou partielle de tout ou partie de ces sites, il a rapidement été décidé de les archiver pendant quatre années durant, de 2010 à 2014 (de manière hebdomadaire ou mensuelle), et ce, afin de permettre la tenue de recherches futures. C'est

ainsi que les archives Web de l'Atlas e-Diasporas se présentent aujourd'hui à nous comme source principale de données à partir desquelles nous déploierons les réflexions portées par cette thèse. De fait, c'est à travers l'angle particulier des usages et des activités en ligne des populations migrantes que nous chercherons à explorer ces corpus.

Partant de là, le Chapitre 1 commencera par dresser un historique des origines d'Internet et du Web. Nous découvrirons alors la manière avec laquelle les Technologies de l'Information et de la Communication (TIC) sont devenues, pour les sociologues, des angles possibles d'analyse des migrations récentes. Faisant figure d'aboutissement scientifique dans ce domaine, nous présenterons alors les différentes étapes de la construction de l'Atlas e-Diaspora en terminant par l'étude de son corpus marocain qui servira de fil conducteur à cette thèse.

Si l'archivage du Web n'est pas au cœur, à proprement parler, de nos travaux (nous explorons les archives après leur constitution), il nous semble néanmoins important de comprendre les modalités de préservation du Web à travers le monde. Le Chapitre 2 reviendra, ainsi, sur les vingt premières années de cet effort d'archivage, essayant de faire un tour d'horizon des diverses techniques de collecte, de stockage et de consultation du Web passé. Ce faisant, nous complèterons la description du réseau marocain de l'Atlas e-Diasporas, considéré cette fois-ci comme un corpus d'archives Web à explorer.

Arrachées au Web vivant par des robots collecteurs, les archives du Web passé basculent dans une nouvelle temporalité, induisant la présence de différents biais de compréhension que le Chapitre 3 cherchera à décrire : cécité de collecte, cohérence entre pages, contenu dupliqué... Ce sera ainsi l'occasion de porter un regard critique sur les archives Web telles que nous les connaissons. Nous présenterons ensuite, d'un point de vue technique, notre moteur d'exploration des archives Web, support des analyses à venir.

Cela aboutira, dans le Chapitre 4, à la définition d'une stratégie d'analyse capable de s'affranchir de l'héritage pesant des outils de collecte, ou tout au moins, d'en atténuer les effets. Nous changerons ainsi d'unité d'exploration, en présentant notre principale contribution : le fragment Web. Nous porterons notre réflexion sur la question de la datation des corpus archivés, en associant à chaque fragment une date d'édition. Un cas simple de détection d'événements dans les archives Web nous permettra d'en faire la démonstration.

Arrivés au Chapitre 5, nous plongerons enfin au cœur de nos archives en menant deux explorations successives. Nous chercherons, en premier lieu, à questionner le devenir d'un collectif de blogueurs marocains de l'étranger. Entre 2008 et 2018, ces internautes ont progressivement déserté leur blogosphère pour migrer vers des plateformes de

réseaux sociaux. Nous suivrons, ensuite, l'impact d'une mobilisation [3], liée aux vagues contestataires du Printemps arabe, sur les membres d'une communauté centrale de l'e-Diaspora marocaine : le forum en ligne *yabiladi.com*. Enfin, tirant les leçons de ces deux explorations, nous introduirons la notion de moments pivots du Web, comme une contribution à l'analyse du Web en tant qu'objet de recherches historiques.

3. Il s'agira des grandes manifestations marocaines du 20 février 2011.

Nous terminerons cette thèse par un aperçu, dans le Chapitre 6, de la diversité des formes d'archives Web et de traces numériques des migrations qu'il nous aura été donné d'explorer. En étudiant les logs de navigation de l'Internet Libre de la Bibliothèque Publique d'Information du Centre Pompidou (BPI), nous chercherons à comprendre les particularités d'un Web de première nécessité, destiné à un public précaire. Puis nous examinerons les archives du programme d'hébergement des personnes exilées Comme à la Maison (CALM), afin de saisir les dynamiques de l'accueil des réfugiés en France, chez des particuliers, à la fin de l'été 2015.

Principales contributions

Publications

— Lobbé, Q. (2018), Where the dead blogs are : a disaggregated exploration of Web archives to reveal extinct online collectives, 20e International Conference on Asia-Pacific Digital Libraries (ICADL), 2018

— Lobbé, Q. (2018), Revealing historical events out of Web archives, 22e International Conference on Theory and Practice of Digital Libraries (TPDL), Demonstration paper, 2018

— Lobbé, Q. (2016), *Cartographier les jungles*, Plein Droit n°110, 2016

Communications

— Lobbé, Q. (2018), Where the dead blogs are [4] : a disaggregated exploration of Web archives to reveal extinct collectives, Web Archiving : Best Practices for Digital Cultural Heritage, Jerusalem, 2018

— Diminescu, D., Jacomy, M. et Lobbé, Q. (2018), About E-Diasporas [5], International Forum on Migration Statistics, Session Migration Traceabillity, Paris, 2018

— Lobbé, Q. (2017), Introducing Web Fragments [6] : An exploration of web archives beyond the webpages, Medialab's research seminar, SciencesPo Paris, 2017

— Lobbé, Q. (2016), Of Maps and Refugees [7], DiasporasLab Workshop, University of California, Los Angeles, 2016

4. http://qlobbe.net/ressources/where-the-dead-blogs-are.pdf

5. http://qlobbe.net/ressources/IFMS_2018.pdf

6. http://qlobbe.net/ressources/introducing_web_fragments.pdf

7. http://qlobbe.net/ressources/of_maps_and_refugees.pdf

Cours

— Lobbé, Q. (2016), Voyage au cœur d'un index Lucene [8] : comprendre les moteurs de recherche de l'intérieur, Télécom ParisTech, 2016

8. http://qlobbe.net/ressources/search.pdf

Codes et logiciels

— Archive-miner [9] : moteur d'extraction et de transformation d'archives Web DAFF en documents Solr (Java, Spark)

— Archive-search [10] : moteur de recherche d'archives Web (Java, Bash, Solr)

— Rivelaine [11] : bibliothèque d'extraction de fragments Web (Scala, Javascript)

— Peastee [12] : interface de visualisation d'archives Web (Javascript, HTML)

9. https://github.com/lobbeque/archive-miner

10. https://github.com/lobbeque/archive-search

11. https://github.com/lobbeque/rivelaine

12. https://github.com/lobbeque/peastee

Chapitre 1

Les Représentations en ligne des diasporas

Web et diasporas sont étroitement mêlés. Le premier est rapidement devenu un support de représentation et d'organisation au service des secondes. Les diasporas se drapent ainsi de sur-couches numériques, serties de sites Web migrants, de comptes Twitter et de groupes Facebook liés entre eux, connectés les uns aux autres et formant de véritables communautés diasporiques en ligne, dynamiques et cohérentes.

Car le Web connecte et rapproche. C'est une toile qui, depuis ses origines, a été construite pour mettre en relation des ressources, des personnes et des espaces. Révolution technique, le Web s'inscrit dans la longue histoire de l'informatique, prolongeant et s'appuyant sur l'une des inventions militaro-scientifiques majeure des années 70 (Hafner et Lyon, 1998) : Internet.

Néanmoins, Internet et le Web ne doivent pas être confondus, l'un servant d'ossature à l'autre. Aussi, la première partie de ce chapitre, sera l'occasion pour nous de revenir sur la naissance du Web, au tout début des années 90. Initialement pensé comme un territoire virtuel à explorer, comme une galaxie de sites parmi lesquels naviguer librement, le Web se mue pourtant progressivement en un flux continu d'informations, qu'il convient de suivre avec assiduité. L'internaute s'affiliant à un panel de canaux qu'il affectionne, passant d'un silo de données à un autre.

Mais, toujours en mouvement, le Web reprend rapidement la course de son évolution et devient mobile. Il se glisse aujourd'hui dans nos poches, on l'emporte avec nous, partout, tout le temps. Ubique, il apprend de nous et s'adapte à nos envies. Il sélectionne les ressources que nous consultons, dresse des cartes personnalisées des espaces qu'il nous faut visiter et nous suggère des listes de personnes à qui nous lier. Plutôt que de nous connecter aux autres, le Web nous rapproche désormais de nous même et des nôtres, cloisonnés dans des bulles de filtrages aux parois réfléchissantes.

Être présent sur la toile, y refléter ce que l'on est ou changer d'identité, retrouver ceux qui nous sont proches ou se connecter à d'autres

plus éloignés, sont des motifs récurant du Web. Les populations migrantes ont ainsi été parmi les premières à s'approprier et à s'emparer de ces mécanismes de représentation à distance. De fait, à la fin des années 90, le développement des Technologies de l'Information et de la Communication (TIC) révolutionne le rapport au temps, à l'espace et aux frontières. L'avènement conjoint du Web grand public, des messageries électroniques et des téléphones mobiles offre de nouveaux outils de communication et d'organisation aux populations migrantes et aux collectifs dispersés.

Aussi, la seconde partie de ce chapitre, nous permettra de saisir le changement de paradigme scientifique qui s'opéra, alors, autour de la figure du migrant comme objet d'étude. Considéré, jusque là, à travers le prisme du double déraciné (Sayad, 2000), à la fois absent dans son pays d'origine et difficilement intégré à son pays d'accueil, les TIC permettent, au contraire, de voir et de comprendre le migrant comme l'élément moteur d'une économie du lien et de la présence à distance (Diminescu, 2008).

Ce faisant, l'intérêt sociologique ou historique pour les formes de représentations en ligne des diasporas s'est rapidement développé. Sur la toile, les populations migrantes s'organisent collectivement à travers des réseaux de blogs ou sur des forums de discussion. Les personnes s'y rassemblent et font groupe autour de valeurs culturelles communes et partagées. Le Web devient alors un terrain favorable à l'établissement de points de rencontres, considérés comme des espaces de connections ou de re-connections avec les siens, tout autant que des lieux de conservation d'une mémoire dispersée, voire comme de véritables outils au service d'actions militantes.

Ce faisant, nous terminerons ce chapitre par la présentation de l'Atlas e-Diasporas (Diminescu, 2012a), fruit du travail coordonné d'un collectif de chercheurs, d'ingénieurs et de designers, visant à explorer, documenter et cartographier des réseaux de sites Web migrants. Créés ou maintenus par des personnes migrantes [1], ces sites Web sont connectés les uns aux autres, formant ce que nous appelons une e-Diaspora suivant la nomenclature introduite par l'Atlas. Aussi et plus concrètement, nos réflexions à venir seront basées sur l'exploration de ces e-Diasporas et de leurs archives Web [2].

1. De manière générale, tous les sites liés explicitement à la vie d'une diaspora donnée ont été cartographiés (sites personnels, d'associations ou d'ambassades).

2. Archives sur lesquelles nous reviendrons en détail (Section 2.3).

1.1 Généalogie du Web

L'informatique est la science des machines programmables et du traitement automatisé de l'information. Alors que l'origine des premières machines à calculer remonte à l'antiquité (abaques grecs [3], bouliers chinois...), le traitement automatisé de tâches élémentaires n'apparait qu'avec l'inauguration du métier à tisser programmable, créé par

3. https://fr.wikipedia.org/wiki/Abaque_(calcul)

J.M. Jacquard, au début du XIXe siècle. Suites finies et non ambigües d'instructions codées sur des cartes perforées, l'invention de Jacquard préfigure la **programmation algorithmique** à venir qui, de A. Lovelace[4] à la machine de Turing (1935), s'attèlera alors à la résolution de problèmes de plus en plus complexes. Parallèlement, les ordinateurs ne cessent de se perfectionner et de voir leurs capacités de calculs augmenter (Lazard et Mounier-Kuhn, 2016).

4. En 1843, A. Lovelace crée ce qui est considéré comme le premier programme informatique, à savoir un algorithme de calcul des nombres de Bernouilli.

Faire dialoguer les machines

En pleine guerre froide, J.C.R. Licklider, alors directeur de l'agence américaine Advanced Research Projects Agency (ARPA), théorise un changement majeur : le passage de l'âge du calcul par ordinateur à celui des communications par machines interposées (Licklider et Taylor, 1968). Profitant de la création de la NASA[5], lancée pour contrer l'URSS sur le terrain de la conquête spatiale, Licklider réoriente les travaux de l'ARPA pour laquelle il négocie une plus grande indépendance budgétaire et scientifique (Tréguer, 2017). Ce faisant, ses équipes se focalisent sur la possibilité de faire communiquer des ordinateurs entre eux et à distance par transfert d'information. Alors qu'il existait déjà des circuits de mutualisation des forces de calcul, Licklider cherche plutôt à privilégier la conception d'un réseau informatique pensé, d'abord, comme un vecteur de circulation de l'information, tel qu'il le prophétise :

5. National Aeronautics and Space Administration

« *Life will be happier for the on-line individual because the people with whom one interacts most strongly will be selected more by commonality of interests and goals than by accidents of proximity. Second, communication will be more effective and productive, and therefore more enjoyable. Third, much communication and interaction will be with programs and programmed models, which will be (a) highly responsive, (b) supplementary to one's own capabilities, rather than competitive, and (c) capable of representing progressively more complex ideas without necessarily displaying all the levels of their structure at the same time – and which will therefore be both challenging and rewarding. And, fourth, there will be plenty of opportunity for everyone (who can afford a console) to find his calling, for the whole world of information, with all its fields and disciplines, will be open to him – with programs ready to guide him or to help him explore.* » — (Licklider et Taylor, 1968, p.21)

Convaincus par Licklider, de jeunes chercheurs et ingénieurs rejoignent rapidement l'ARPA et planchent sur un premier prototype. L'un d'entre eux, P. Baran propose de bâtir un **réseau distribué** (Figure 1.1, c) à l'opposé des modèles centralisés ou décentralisés plus classiques[6] (Figure 1.1, a & b). Le modèle distribué tend, en effet, à éviter la création de *hubs* ou de points d'étranglements fragilisant la structure globale du réseau et préfère, au contraire, un maillage redondant où chaque nœud (c.-à-d., une machine) est relié à son voisinage direct.

6. Modèles notamment utilisés pour les réseaux téléphoniques.

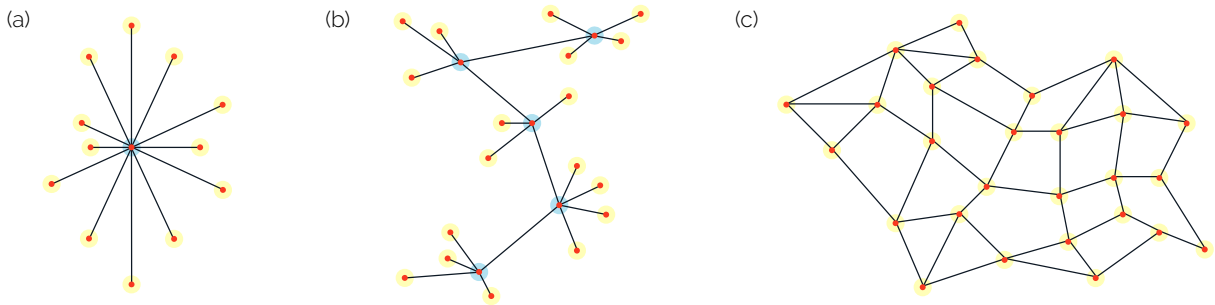


Figure 1.1: Trois modèles de réseaux : centralisé (a), décentralisé (b), distribué (c)

En 1969, après quelques années d'efforts, un premier réseau baptisé **Arpanet** voit le jour. Un test grandeur nature est effectué dans la foulée. Des messages sont alors échangés à distance entre les machines de quatre universités de l'Ouest américain[7]. Mais il faut attendre octobre 1972 pour que l'Arpanet soit publiquement présenté lors de l'International Conference on Computer Communication (ICCC), à Washington. Le réseau démontre alors sa capacité à faire communiquer pour la première fois des ordinateurs entre côtes Est et Ouest des États-Unis (Hafner et Lyon, 1998, p.175-186).

7. L'*University of California, Los Angeles* (UCLA), l'*University of California, Santa Barbara* (UCSB), la *Leland Stanford Junior University* (laboratoire SRI) et l'*University of Utah*.

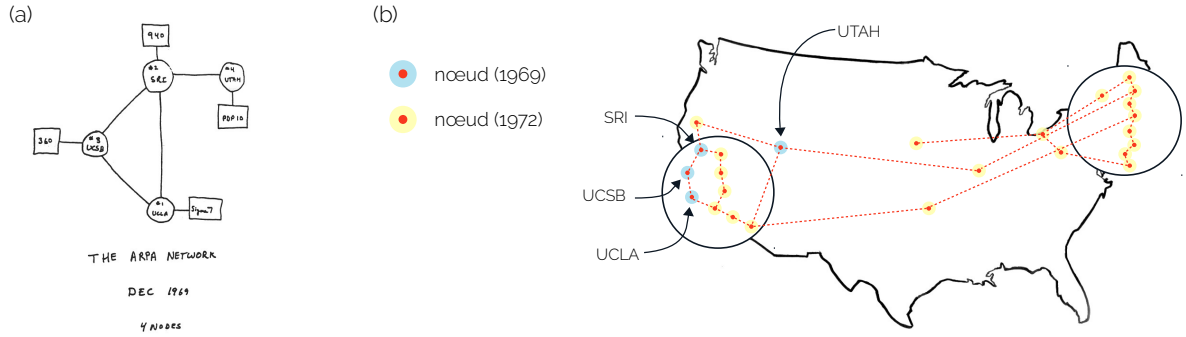


Figure 1.2: Schéma de la première version de l'Arpanet (P. Baran, 1969, a) et structure de l'Arpanet en 1972 (b)

Mais au delà de la pure démonstration technique, c'est une communauté entière qui se rassemble ce jour là à l'ICCC et, parmi elle, des personnes qui ne s'étaient encore jamais rencontrées hors du réseau. Et c'est là l'une des grandes réussites symboliques de l'Arpanet que d'avoir fait émerger les premières communautés en ligne sous la forme, à cette époque, de listes de discussions et de spécifications[8]. Le réseau met en relation, il rapproche les chercheurs malgré la distance.

8. Notamment à travers la liste de discussion et de spécifications *Request for comment* (RFC), `https://fr.wikipedia.org/wiki/Request_for_comments`

Néanmoins, pour communiquer avec n'importe quel interlocuteur, humain ou robot, il faut d'abord établir un protocole (un langage par exemple) et trouver le moyen technique de transmettre le message que l'on cherche à émettre (par la parole ou par le geste). Aussi, le génie des équipes de l'ARPA est d'avoir implémenté pour la première fois à grande échelle un système de communication robuste : la **communication par paquet**.

Contrairement à la communication téléphonique, ce modèle ne nécessite pas l'établissement d'un circuit électrique continu entre deux points fixes. En effet, sur l'Arpanet chaque message est divisé en un ensemble de paquets distincts [9], envoyés les uns à la suite des autres de manière discontinue. C'est au récepteur de reconstituer l'intégralité du message grâce aux informations contenues dans son en-tête (Figure 1.3).

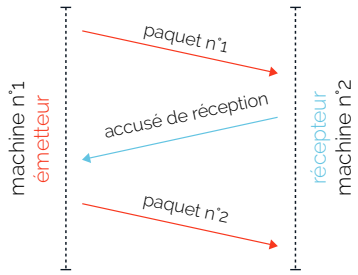


FIGURE 1.3: Fonctionnement générale de la communication par paquets

9. Système antérieur à l'Arpanet `https://fr.wikipedia.org/wiki/Commutation_de_paquets`.

S'appuyant sur le succès de sa démonstration, l'Arpanet s'étend tout au long des années 70, accueillant de nouvelles machines voire des réseaux fermés d'ordinateurs déjà existants [10]. Mais face à cette croissance continue se pose rapidement la question des modalités de communication et d'intégration de machines diverses et variées. Comment les faire parler la même langue à l'échelle globale tout en respectant leur spécificités techniques locales ?

10. Notamment le réseaux de l'International Network Working Group (INWG)

Pour résoudre ce problème, et ce dès 1983, l'Arpanet fait du **TCP/IP** [11] son principal protocole de communication, désormais adossé à une **architecture en couches**. Ce faisant, chaque nœud du réseau est alors identifié par une adresse IP unique permettant d'acheminer précisément les paquets de données vers un récepteur [12] (Figure 1.4).

11. De l'anglais, Transmission Control Protocol/Internet Protocol, `https://fr.wikipedia.org/wiki/Suite_des_protocoles_Internet`

12. `https://fr.wikipedia.org/wiki/Adresse_IP`

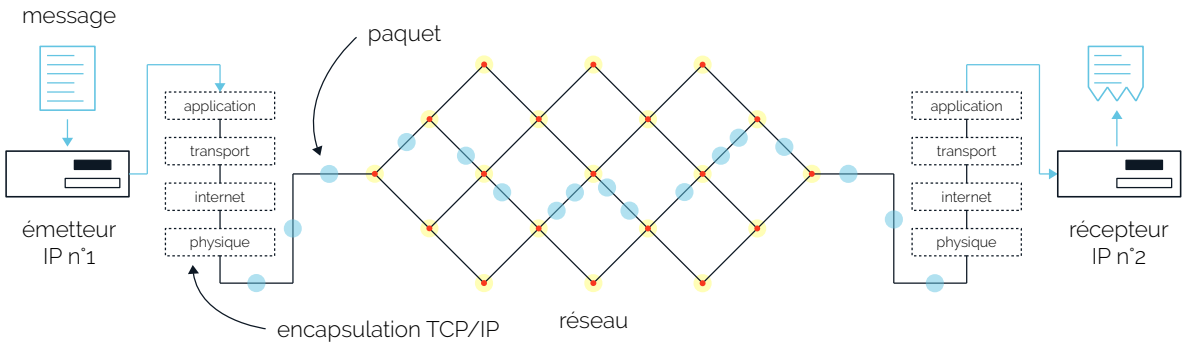


FIGURE 1.4: Cheminement d'un message à travers un réseau de type Arpanet, suivant le protocole TCP/IP

Les messages, quant à eux, cheminent à travers les différentes couches du réseau qui se voient, chacune, attribuer un rôle différent dans le transport de l'information. L'idée générale est de permettre à n'importe qui de développer de nouveaux protocoles applicatifs (comme un système de messagerie ou de discussion en temps réel), tout en conservant

une grande autonomie et sans avoir à se soucier de la livraison des messages. Ainsi, l'adoption du TCP/IP dynamise le développement international du réseau, en lui offrant une plasticité jusque là inédite. En effet et c'est là la clé de sa réussite, l'Arpanet est désormais capable de s'interconnecter avec d'autres réseaux existants [13].

13. L'Arpanet n'est pas le 1er réseau à avoir existé. D'autres *internets* étaient à cette époque déjà en place. Mais l'Arpanet est le 1er réseau par lequel tous pouvaient s'interconnecter.

L'Arpanet vieillissant laisse alors progressivement sa place à un **Internet** unifié et devenu global. Le vieux rêve de Licklider a pris forme, les machines et les hommes communiquent d'un continent à l'autre et les réseaux pénètrent déjà différentes couches de la société : des administrations aux industries, en passant par le monde académique.

L'invention du World Wide Web

Au tournant des années 90, T. Berners Lee et R. Cailliau, alors chercheurs au CERN [14], planchent sur un nouveau système de partage et de gestion de documents informatiques. Leur intuition est la suivante : appliquer le principe de l'**hypertexte** aux capacités de circulation d'information de l'Internet. Avant cela, l'Internet servait principalement à envoyer des messages électroniques, à partager des fichiers (système *FTP*), à communiquer entre serveurs (système *Telnet*), à échanger des articles au sein de communautés (système *Usenet*)...

14. Désignation courante de l'Organisation européenne pour la recherche nucléaire, à Genève.

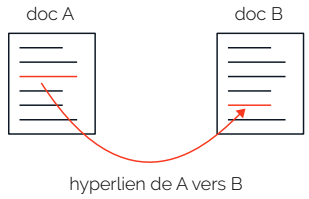


Figure 1.5: Un hyperlien entre deux documents A et B

Théorisé par T. Nelson au milieu des années 60 (Nelson, 1967), l'hypertexte est un modèle de mise en relation de documents numériques par liens de citation ou références (c.-à-d., les hyperliens). Chaque document ou ensemble de documents hypertextes contient des unités d'information (texte, image...) reliées les unes aux autres par des hyperliens (Figure 1.5). L'hypertexte [15] est donc tout autant un système d'organisation de documents qu'un principe de navigation dans un espace numérique. En effet, un utilisateur peut cheminer d'une ressource à l'autre, depuis un terminal informatique, en suivant de manière non linéaire les liens de son choix.

15. D'autres systèmes hypertextes existaient alors comme Hypercard (années 80) ou Gopher (contemporain du Web). Ce qui, au final, a fait le succès du Web fut sa grande souplesse et l'utilisation d'une licence libre, contrairement à Gopher.

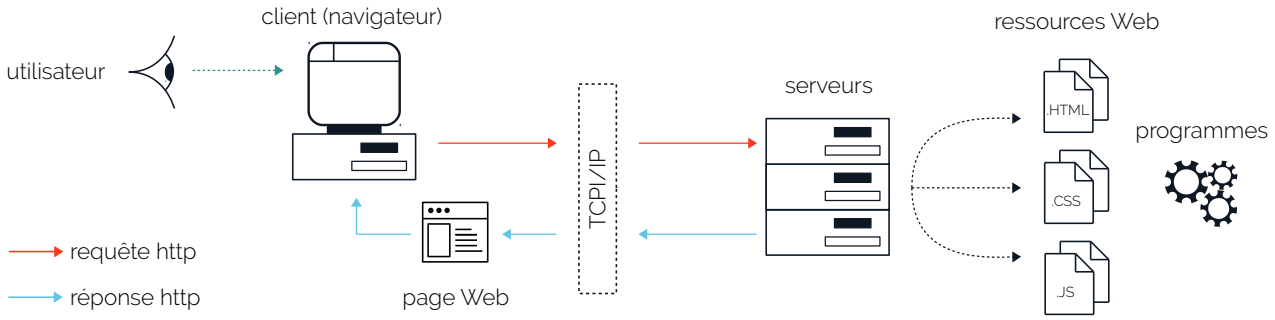


Figure 1.6: Communication HTTP entre un client et un serveur sur le Web

Ainsi, le World Wide Web (Web), tel que l'inventent Berners Lee et Cailliau (CERN, 1993), se veut être un système d'écriture, de lecture et de partage de documents numériques sur Internet tout autant qu'un espace de navigation hypertexte. En cela, le Web est une application d'Internet [16] permettant à un **client** de consulter une ressource située à un point distant du réseau et mise à disposition par un **serveur**. Pour faire dialoguer clients et serveurs, le Web utilise le protocole de communication **HTTP** [17]. Aussi, un client voulant accéder à un document en ligne, envoie une requête HTTP au serveur qui en a la charge. Sa requête est alors transférée à travers le réseau Internet, et ce, jusqu'au serveur. Si ce dernier le peut ou le veut, il donne accès à la ressource Web via l'envoi d'une réponse HTTP dédiée (Figure 1.6). En pratique, l'internaute n'est pas directement considéré comme étant le client, ce rôle est attribué à son **navigateur Web** [18] à qui revient également la tâche de charger et d'afficher à l'écran le document à consulter.

16. Il se développe et s'étend dans la couche applicative du réseau.

17. De l'anglais *Hypertext Transfer Protocol*, notons qu'il existe une version sécurisée de ce protocole : le https.

18. Logiciel conçu pour naviguer sur le Web et afficher son contenu.

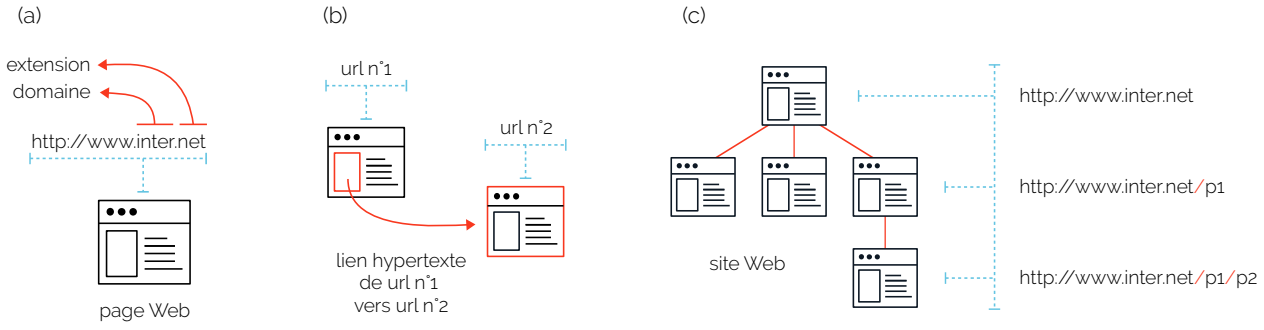


FIGURE 1.7: De la page au site Web

Sur le Web, chaque document est identifié par une adresse unique appelée **URL** [19]. Une URL est constituée du préfix *http://* [20], suivi d'un nom de domaine indiquant le serveur web auquel le navigateur doit s'adresser. Ce nom de domaine est pourvu d'une extension (*.fr*, *.com*, *.net*...) désignant un ensemble plus large de serveurs (Figure 1.7, a). Partant de là, un hyperlien est toujours construit autour de l'URL du document pointé, seule l'adresse de la cible est ainsi renseignée (Figure 1.7, b).

L'unité de consultation et de navigation de base du Web est la **page Web** [21]. Aussi, nous nous tiendrons, dans la suite de ce manuscrit, à l'usage courant qui tend à assimiler tout document du Web à une page, bien que les images ou vidéos en ligne soient adressables par leurs propres URL. Sur ce point, chaque document du Web possède un marqueur de type, le type MIME [22], indiquant plus ou moins précisément sa nature : text, image, audio...

19. De l'anglais *Uniform Resource Locator*

20. Indiquant le protocole de communication. On utilise *https ://* pour la version sécurisée.

21. Nous discuterons des langages de programmation conçus pour représenter les pages Web en Section 4.3

22. `https://fr.wikipedia.org/wiki/Type_de_médias`

Plusieurs pages peuvent être rassemblées en une seule et même entité de plus haut niveau, abstraite et arbitraire : le **site Web** (Figure 1.7, c). De fait, un site est un espace hypertexte cohérent du Web, construit comme un arbre de documents partant d'une page principale (la *front page*) et dont les ramifications peuvent être suivies jusqu'aux pages les plus éloignées. La forme des URL de chaque page d'un site Web peut traduire cette arborescence (Figure 1.7, c).

Toile, silos et bulles

Le premier navigateur Web est développé par les équipes du CERN en 1992. Ce *navigateur en ligne de commande*[23], qui ne dispose pas encore d'interface graphique à proprement parler, permet, en entrant une URL donnée, d'accéder à une page Web et d'afficher le texte qu'elle contient. Le Web est alors considéré comme une **toile**[24] pouvant théoriquement s'étendre dans toutes les directions de développement et formant, de fait, un réseau rhizomatique[25] de sites et de pages sans véritable hiérarchie interne.

23. http://line-mode.cern.ch/

24. Traduction littérale de World Wide Web

25. https://frama.link/zXhEvzna

Pourtant, le Web est plus structuré qu'il n'y parait. Plusieurs modèles le décrivent ainsi comme une galaxie possédant, en son centre, un noyau très dense de sites Web fortement connectés les uns aux autres (Broder *et al.*, 2000). Ces sites font **autorité** car, suivant une loi de puissance[26], ils attirent dans leur seule direction une grande partie des liens de citations émis par les autres pages (Barabási *et al.*, 2000). C'est en passant par ces points centraux du réseau que l'on accède généralement aux parties les plus éloignées du Web formées, par endroit, de grappes lointaines de sites ou d'ilot de pages isolées (Kumar *et al.*, 2000).

26. https://fr.wikipedia.org/wiki/Loi_de_puissance

Pour l'internaute, à moins de connaitre d'avance une URL à visiter, l'exploration du Web se fait en cheminant de page en page, via un navigateur. Aussi et afin d'améliorer ce processus de découverte, un nouveau type de site apparait : les moteurs de recherche. Deux d'entre eux, Yahoo ! et Lycos sont lancés conjointement dès 1994 et préfigurent déjà l'arrivée d'un grand nombre de concurrents. Le premier, Yahoo ! est un registre indexant les URL de plusieurs centaines ou milliers de sites Web que l'on peut alors parcourir suivant un arbre thématique. Lycos, quant à lui, introduit un modèle hybride, à la fois répertoire et réel moteur de recherche basé sur le contenu des ressources à découvrir. Le succès de ces deux pionniers ouvre la voie aux moteurs de première génération que sont Infoseek, HotBot et Altavista (le plus populaire). L'accès à un document ne se fait plus par le choix d'une URL ou d'une suite d'URL, mais en se fiant à une liste de résultats retournés par un algorithme.

C'est avec l'arrivée de Google, en 1998, que se dessinent de nouvelles formes de navigation en ligne, donnant une prime aux sites Web po-

pulaires et centraux avec la mesure dite du PageRank[27] (Page *et al.*, 1999). Google réintroduit, par ailleurs, une forme de simplicité dans la recherche de ressources sur le Web, et ce, sans oublier d'être efficace. En effet, en distribuant les tâches de calcul parmi plusieurs milliers de machines (Brin et Page, 1998), les méta-données, le texte puis les images des pages Web sont automatiquement et rapidement analysés par les systèmes de Google. Face la toile, l'internaute peut alors être vu comme un navigateur ayant à sa disposition un certains nombre d'outils conçus pour accélérer ou spécialiser ses recherches (Figure 1.8, a).

27. https://fr.wikipedia.org/wiki/PageRank

Dans la foulée, l'invention des systèmes de syndication permet à l'internaute de ne plus aller chercher l'information au hasard de sa navigation, mais de la suivre directement depuis des applications dédiées. Les Google Reader[28] et autres lecteurs de flux RSS ou Atom[29] agrègent les contenus publiés sur des blogs, sur des forums ou sur des sites d'actualités en un ou plusieurs canaux unifiés. Le Web n'est plus alors seulement un territoire à explorer, il se présente également comme un flux continu d'informations et de ressources à suivre.

28. https://fr.wikipedia.org/wiki/Google_Reader

29. Formats de données utilisés pour la syndication à du contenu Web https://fr.wikipedia.org/wiki/RSS et https://fr.wikipedia.org/wiki/Atom

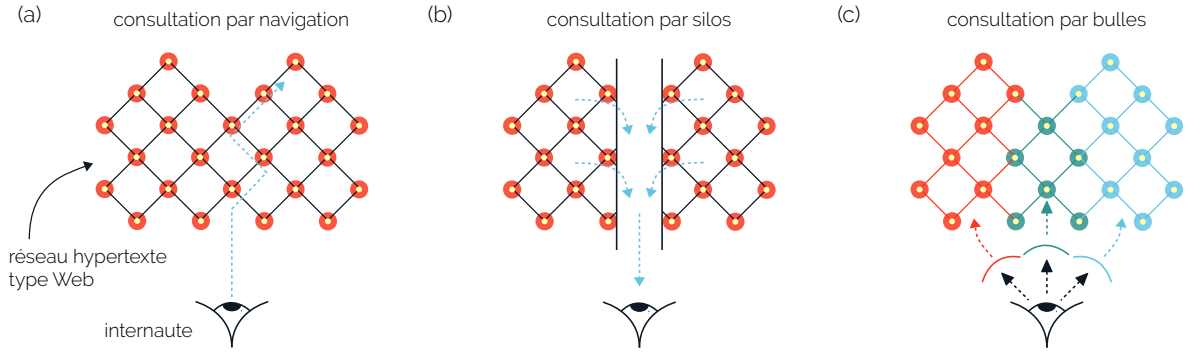


FIGURE 1.8: Différents modes de consultation d'une information dans un réseau hypertexte (type Web)

Poussant cette logique plus avant, la vague des réseaux sociaux comme Myspace (2003), Facebook (2004) ou encore Twitter (2006) cherche à installer des espaces de partages et de connections à l'intérieur même du Web, voire *en marge* de celui-ci (Ellison, 2007). En effet, tout est fait pour que l'internaute consulte le Web depuis l'intérieur de ces plateformes, de manière indirecte via le contenu partagé par des amis en ligne ou par des algorithmes de recommandation. Ces réseaux sociaux deviennent alors de véritables médias alternatifs dont le modèle économique est basé sur la vente d'espaces publicitaires en ligne (Loveluck, 2015). Ce faisant, la navigation sur le Web s'oriente de plus en plus vers des circuits fermés, l'information se découvrant et s'échangeant depuis des **silos** de données (Figure 1.8, b) auxquels il est obligatoire de se connecter (Helmond, 2015).

Au tournant des années 2010, le Web devient mobile et peut être

consulté autrement que depuis l'ordinateur du salon. Grâce au smartphone, le Web nous suit partout, tout le temps. Nous le glissons dans nos poches et il s'adapte en conséquence, la forme et le contenu des sites variant en fonction de l'appareil depuis lequel on navigue ou par rapport à la qualité de notre connexion réseau [30]. De plus, certaines applications ou certaines plateformes de réseaux sociaux sont créées à dessein et uniquement pour être utilisées depuis un téléphone portable ou via l'écran d'une télévision. Ainsi, le Web se personnalise, il apprend de nous et des traces que nous déposons sur la toile.

30. https://fr.wikipedia.org/wiki/Site_web_adaptatif

La plasticité du Web va alors grandissante à mesure que les systèmes qui y prospèrent nous dépossèdent de nos choix de consultations (Cardon, 2015). Les filtres cherchent à se faire oublier de nous. Ils sont automatiquement réajustés suivant l'évolution de nos profils en ligne et déterminés par des algorithmes toujours plus performants. Nous expérimentons tous désormais un Web différent. Il n'y a plus une seule toile, mais une multitude de versions adaptées à nos habitudes de navigation. Les résultats des moteurs de recherche ou l'information qui circule sur les pages Facebook sont personnalisés suivant chacune de nos particularités (Robertson *et al.*, 2018). Nous cheminons maintenant sur le Web enfermés dans des **bulles** de filtrage [31] (Figure 1.8, c) façonnées à notre image (Gillani *et al.*, 2018).

31. L'expression *chambre d'écho* est également utilisée.

1.2 Dispersés mais présents sur la toile

Lors de la leçon inaugurale [32] de sa chaire *Migrations et sociétés*, donnée au Collège de France en avril 2018, F. Héran choisit de rappeler, en préambule, que les migrations ont de tout temps été ancrées dans nos vies. Par-delà les crises et les épisodes spectaculaires qui polarisent l'attention [33], les migrations restent et demeurent des composantes essentielles des dynamiques qui animent nos sociétés, comme nous le rappellent les faits suivants :

32. Voir la leçon : https://frama.link/CmK5e2ex

33. Voir le travail réalisé sur la base de données CALM en rapport avec la *crise des réfugiés*, en Europe à l'été 2015 (Section 6.2).

Pour l'ONU [34], le terme **migrant** désigne « *toute personne qui a résidé dans un pays étranger pendant plus d'une année, quelles que soient les causes, volontaires ou involontaires, du mouvement, et quels que soient les moyens, réguliers ou irréguliers, utilisés pour migrer* ». En Europe et plus précisément au sein de l'Union Européenne (UE), le flux de migrants est estimé selon l'INSEE [35] à près de 2,4 millions de personnes pour l'année 2015. Cet ordre de grandeur fait de l'Europe un continent important d'immigration, supérieur même à des terres historiques de migrations comme les États-Unis.

34. https://refugeesmigrants.un.org/fr/définitions

35. https://www.insee.fr/fr/statistiques/1521331

De son côté, la France est un pays d'immigration depuis le milieu du XIXème siècle, époque où la main d'œuvre étrangère était sollicitée pour venir travailler dans l'industrie minière et ferroviaire. Aussi, durable-

ment marquée par le passé colonial français, l'immigration hexagonale est aujourd'hui constituée à 44% de personnes originaires du Maghreb et d'Afrique subsaharienne et à 3% de l'ancienne Indochine[36]. En 2018, les migrants de la première génération[37] représentent ainsi 11% de la population française, la seconde génération[38] comptant, quant à elle, pour 13% du total. De fait, un quart de la population française est aujourd'hui liée à l'immigration.

36. Projection pour l'année 2018, basée sur les statistiques 2014 de l'INSEE `https://www.insee.fr/fr/statistiques/3303358?sommaire=3353488`

37. C.-à-d., les personnes nées à l'étranger, même naturalisées française par la suite.

38. Personne descendant d'un ou deux parents migrants.

Par ailleurs, Héran souligne que le fait économique (se déplacer pour trouver du travail) n'est plus une composante première des migrations modernes. En effet, à partir des années 70, une décorrélation s'opère entre motifs économiques (croissance/récession du pays d'accueil) et flux d'immigration. De nos jours, les motifs qui poussent une personne à se déplacer sont bien plus nombreux : études, regroupement familial, changement climatique, bouleversements politiques, fuir sous contrainte un pays, un régime ou tout simplement rechercher une vie meilleur.

Mais souvent perçues comme trop froides, les statistiques sur les migrations échouent à traduire et à transcrire l'infinie complexité de ces dispersions individuelles et collectives. Ce faisant, Héran clôt sa leçon en rappelant que les entretiens sociologiques, les récits de migrants et les analyses de traces (parlées, écrites...) sont autant de contrepoints capables de venir contextualiser, densifier, critiquer ou interroger les chiffres manipulés par les démographes, les administrations et les politiques.

Aussi, cette section sera pour nous l'occasion de présenter comment de nouveaux types de traces (les traces nativement numériques et notamment celle laissées sur le Web) sont devenues, depuis la fin des années 90, objets de compréhension des migrations, passant d'une science de l'absence à l'étude de multiples formes de présences.

Absence et présence

Au cours des années 70 et 80, le sociologue A. Sayad réalise plusieurs séries d'enquêtes de terrain, s'entretenant le plus souvent avec des immigrés algériens venus travailler en France dix ou vingt ans plus tôt (Sayad, 2000). Partant de l'étude de ces récits individuels, sa méthode consiste à avancer par croisements et rapprochements successifs : croisements de sujets interviewés, rapprochements d'époques, confrontation entre générations... Il amène ainsi les personnes, avec qui il s'entretient, à restituer au présent des récits d'immigrations passées, révélant par là-même certains regrets face aux illusions initiales de l'émigré, voire une forme de souffrance a posteriori chez l'immigrant. Il met également en avant les déterminations individuelles et collectives liées à la mémoire récente de la Guerre d'Algérie et de la colonisation française, montrant, de fait, l'existence de décalages générationnels structurant

les représentations et les pratiques des personnes rencontrées.

Une distance nait ainsi au cœur de la figure de l'émigré/immigré : une distance physique, sentimentale, familiale et spirituelle vis-à-vis de son pays d'origine d'une part, et une distance culturelle, linguistique, religieuse voire sociale et politique vis-à-vis du pays d'accueil d'autre part. Le migrant est alors vu comme un déraciné, pris dans un tiraillement continu entre présence et absence :

« *L'émigration, pour ne pas être pure "absence", appelle une manière d'ubiquité impossible, une manière d'être qui affecte les modalités de l'absence qu'elle entraîne (de même qu'elle affecte les modalités de la présence par laquelle se matérialise l'immigration) : continuer à "être présent en dépit de l'absence", à être "présent même absent et même là où on est absent" – ce qui revient à "n'être que partiellement absent là où on est absent" – c'est le sort ou le paradoxe de l'émigré – et, corrélativement, à "ne pas être totalement présent là où on est présent, ce qui revient à être absent en dépit de la présence", à être "absent (partiellement) même présent et même là où on est présent" – c'est la condition ou le paradoxe de l'immigré. Le risque pour l'émigré et pour l'immigré qu'il est aussi est que ces formes incomplètes d'absence et de présence finissent, tôt ou tard, par s'accomplir intégralement : la présence "physique" et seulement physique de l'immigré finira par devenir une présence "morale" aussi (par le corps et par l'esprit ; par l'actuel et par le futur ; par le travail et par l'engendrement, c'est-à-dire le sang ; par le fait et par le droit) ; corrélativement, l'absence matérielle et seulement matérielle de l'émigré finira par devenir une absence "morale" (et "spirituelle"), une absence consommée, une rupture accomplie avec la communauté.* » — (Sayad, 2000, p.225-226)

Pourtant, les derniers travaux de Sayad, à la fin des années 80, révèlent déjà les prémisses d'un changement à venir. De nouvelles formes de présences à distance, qu'il détecte dans l'« *usage inattendu de l'enregistrement sur magnétophone* » (Sayad, 1985), permettent d'entretenir un lien discontinu entre deux personnes. Ainsi, décryptant l'envoi de messages sur cassettes magnétiques entre une mère et son jeune fils émigré, Sayad note cette phrase prononcée par la mère :

« *Ah ! Celle-là [la mallette contenant le magnétophone], c'est ma consolation ! Elle me cautérise (les plaies). Depuis qu'on a inventé ces cassettes, si je pouvais je ne m'en séparerais jamais. Je leur parle comme je te parle, comme je vous parle à tous, comme s'il était là devant moi.* » — (Sayad, 1985, p.12)

La technique commence alors à rendre possible l'ubiquité recherchée par le migrant. En effet, dans les années 90, le perfectionnement et la démocratisation des Technologies de l'Information et de la Communication (TIC) révolutionnent notre rapport à l'espace, au temps et aux frontières. Les téléphones portables, les systèmes de messagerie électroniques et le Web naissant deviennent rapidement, dans les mains des personnes migrantes, des outils de communication et d'organisation individuelle ou collective.

D'un point de vue épistémologique, l'usage des TIC s'installe comme un nouvel angle d'analyse possible des migrations, voire comme une composante essentielle de la figure moderne du **migrant connecté**. Pro-

longeant la réflexion de Sayad, D. Diminescu propose, à travers un manifeste (Diminescu, 2008), une rupture dans la manière de penser l'image du déraciné :

« *un autre principe organisateur émerge : mobilité et connectivité forment désormais un ensemble de base dans la définition du migrant du XXI siècle. Ensemble ils agissent comme un vecteur qui assure et conduit les lignes de continuité dans la vie des migrants et dans les rapports que ceux-ci entretiennent avec leur environnement d'origine, d'accueil ou parcouru. Hier : immigrer et couper les racines ; aujourd'hui : circuler et garder le contact. Cette évolution semble marquer un nouvel âge dans l'histoire des migrations : l'âge du migrant connecté* » — (Diminescu, 2008, p.3)

Le migrant s'inscrit désormais dans une culture de liens qu'il construit lui même et qu'il entretient dans sa mobilité (géographique, sociale, sur les réseaux...). Il se déplace, noue des alliances à l'extérieur de sa communauté mais sans pour autant se détacher de cette dernière, la somme des liens virtuels permettant d'être toujours présent à distance, ici ou là-bas. Réactivables, ces liens s'inscrivent dans une dynamique faite de ruptures et de moments de continuité qu'il nous faut apprendre à analyser.

Ainsi, ces nouvelles formes de présence à distance font rapidement l'objet de recherches dédiées (Diminescu, 2002). Les traces qu'elles génèrent, de par les systèmes techniques sur lesquels elles s'appuient ou via les modalités de stockage qu'elles induisent, engendrent une hybridation scientifique. Des projets pionniers et interdisciplinaires sont alors mis en place dans les années 2000, alliant de fait, méthodes d'analyses sociologiques et techniques d'exploration informatiques. Ces travaux s'incarnent aujourd'hui dans le champ dit des *digital migration studies* [39], domaine de recherche internationalement reconnu (Leurs et Ponzanesi, 2018) et pouvant conjuguer anthropologie (Zijlstra et Liempt, 2017), psychologie (Chen, 2010), sociologie (Damian et Van Ingen, 2014)...

39. Études numériques des migrations

Des traces nativement numériques

Les pratiques communicationnelles des migrants, portées et accentuées par l'usage des TIC, forment de vastes corpus de données à explorer. Captées directement depuis des flux d'informations *live* ou récupérées après coup dans des systèmes de stockage, ces données peuvent aider à comprendre le fonctionnement de certains réseaux transnationaux, tels que des systèmes de transfert d'argent par téléphone portable (Bounie *et al.*, 2010), ou encore servir à mesurer l'intégration des migrants dans les pays d'accueil. Ce sont également des vecteurs de compréhension des modalités de surveillance déployées par les institutions pour contrôler les étrangers (Amoore, 2006). Nous pouvons enfin nous servir de ces traces pour reconstituer le parcours individuel d'une personne passant

d'une frontière à l'autre, à la manière d'un journal de bord, et ce, en supplément d'une série d'entretiens (Diminescu, 2016).

Définies, structurées et finement datées par des systèmes informatiques et automatiques, ces traces témoignent du quotidien des migrations ou de grands événements polarisants. Ce sont des empreintes (Rogers, 2009) qui peuvent être déposées en conscience sur la toile (un message sur un blog), arrachées de force et inscrites dans une base de données (au passage d'une frontière) ou encore dérobées par divers systèmes espions et capteurs à l'insu des personnes concernées (vidéo-surveillance). **Nativement numériques**, certaines de ces traces sont produites exclusivement par et depuis les TIC, ne devant ainsi pas être confondues avec d'autres formes de traces numérisées a posteriori (textes et images scannés, transcription au format numérique de documents historiques passés...) et qui sont l'apanage du champ plus large des Humanités Numériques [40].

Le présente thèse est ainsi quasi-exclusivement concernée par l'exploration de ces traces dites nativement numériques [41]. Aussi, plutôt que de réévaluer et de mettre à jour des méthodes de recherche existantes, ces traces nécessitent le développement d'outils conceptuels et méthodologiques nouveaux [42], à la frontière entre sociologie et ingénierie (Barats, 2016).

À titre d'exemples, travailler sur ce type de traces revient à analyser des séries de courriers électroniques, des conversations en ligne via des systèmes de messagerie (Dekker *et al.*, 2018), la couverture d'un territoire par un réseau mobile (Sánchez-Querubín et Rogers, 2018), des échanges vidéo en direct [43], des traces GPS d'embarcations (Heller et Pezzani, 2014), des messages vocaux postés sur Telegram [44], ou encore des *selfies* partagés sur Instagram (Risam, 2018), etc.

40. https://fr.wikipedia.org/wiki/Humanités_numériques

41. Voir le Chapitre 6 pour un tour d'horizon plus vaste des multiples types de traces nativement numériques.

42. Voir, par exemple, les outils d'analyse développés par les équipes du MédiaLab http://tools.medialab.sciences-po.fr/ ou de DMI https://wiki.digitalmethods.net/Dmi/ToolDatabase.

43. Voir le projet Ubi-screen réalisé par I. Grosse http://www.isabellegrosse.net/ubiscreen/.

44. Utilisation de l'application Telegram (https://telegram.org/), voir https://frama.link/JnjBqw8u.

Être ensemble sur le Web

Au début des années 2000, le Web est déjà pourvoyeur d'une grande partie des traces numériques sur lesquelles peuvent être menées des investigations. Sites et pages Web (Tyner et Kuhlke, 2000), blogs (Park et Thelwall, 2008), forums en ligne (Van den Bos et Nell, 2006) et autres bases de données sont autant de documents succeptibles de faire l'objet de recherches. Ce faisant, le groupe de travail TIC-Migration, coordonné par D. Diminescu, se constitue dans le but d'analyser certains sites Web retenus pour leurs liens directs avec les migrations. En effet, ces nouveaux supports d'information commencent à devenir des formes de représentation à part entière pour les communautés dispersées.

Pour S. Dufoix, le terme **diaspora** renvoie à l'idée de dispersion d'un peuple ou d'une communauté à travers le monde. Mais il peut également désigner les multiples formes d'organisations et de repré-

sentations (ethniques, nationales ou religieuses) de ces même communautés (Dufoix, 2003). Ainsi, toute diaspora cherche à maintenir une mémoire collective, pour créer des repères et un imaginaire commun (Bernal, 2006). Et s'il faut essayer de demeurer ensemble malgré la distance, cela passe souvent par la création d'espaces (physiques, en ligne, sur papier, par la musique...) et de moments dédiés à l'animation de la communauté. On y évoque d'égale manière les grands bouleversements historiques et les petits événements du quotidien (Koukoutsaki-Monnier, 2012) qui jalonnent la vie du groupe.

De fait, les sites Web deviennent progressivement de nouveaux points de rencontres à distance, où chaque communauté décide de son mode de communication privilégié (par vidéo pour les tziganes, par messages vocaux pour les kabyles...). De natures variables, ces espaces en ligne prennent la forme de carrefours où l'on se croise, de vitrines pour l'histoire de la communauté (Whitaker, 2004), de lieux de reconstruction et de partage de repères communs (Sabancioglu, 2011), ou encore de porte-voix au service de revendications politiques (Kumar, 2018).

Combinant leurs analyses, les premiers travaux des chercheurs du groupe TIC-Migration apportent rapidement de nouvelles pistes de réflexion (Laflaquière *et al.*, 2005). Selon C. Scopi (Scopsi, 2009), certains sites Web[45] sont créés autour de valeurs culturelles communes et maintenus exclusivement par des communautés en ligne de migrants, tels que les RME (Ressortissants Marocains de l'Étranger) ou les Viet Keu (Vietnamien de l'étranger). L'analyse de sites communautaires bulgares, menée par R. Soultanova, révèle pour sa part la transformation progressive de sites personnels de migrants[46] en véritables portails ouverts sur la diaspora et fédérant un collectif régulier d'utilisateurs. Pour S. Gangloff, le Web peut aussi servir de point d'appui[47], afin de sensibiliser l'opinion publique sur le devenir des ressortissants turques de l'étranger. De son côté, T. Guignard voit, dans les messages postés sur des sites sénégalais[48], un nouvel espace d'échange nord/sud, les membres de la diaspora sénégalaise souhaitant, grâce au Web, renouer avec leur culture d'origine.

Le profil des migrants faisant vivre quotidiennement les diasporas en ligne peut aussi faire l'objet d'études. En effet, la mise en place et la maintenance de ces sites appellent (tout au moins dans les premières années du Web) un important bagage technique, comme le décrit M. Nedelcu en documentant la création du portail roumain *thebans.com* par un jeune ingénieur émigré au Canada (Nedelcu, 2003). Enfin, la toile devient parfois un outil militant, servant à organiser une action ou à maintenir vivace le souvenir d'une lutte. Sur ce point, le site *pajol*[49] fait figure de précurseur. Créé en 1996, il retrace l'histoire de l'occupation de l'église Saint-Bernard à Paris par le mouvement des sans-papiers, et

45. *mondeberbere.com*, *originesvietnam.com*...

46. *bgcanada.com*

47. *info-turk.be*, *bleublancturc.com*...

48. *homeview.sn*, *lesoleil.sn*...

49. `http://www.bok.net/pajol/index2.html`

ce jusqu'à leur expulsion par les forces de l'ordre en aout de la même année [50].

50. https://fr.wikipedia.org/wiki/Mouvement_des_sans-papiers_à_Paris_en_1996

Pris indépendamment les uns des autres, ces sites ont tous un élément ou une information importante à nous révéler. Pourtant, le Web peut aussi s'appréhender à une échelle plus large et bien plus complexe, faite de rapports de forces, de rapprochements et d'éloignements, d'amitié et d'ignorance. Car le Web est une toile et les formes de représentations en ligne des diasporas n'échappent pas à la tentation de faire réseau. Bien au contraire, les sites migrants se connectent les uns aux autres, via des liens de citation hypertextes et forment ainsi les nouveaux territoires de la dispersion. Pour les explorer, le collectif TIC-Migration s'agrandit et se lance, à la fin des années 2000, dans une campagne de cartographie de plusieurs milliers de sites migrants, aujourd'hui réunis au sein de l'Atlas e-Diasporas.

1.3 L'Atlas e-Diasporas : cartographier les sites migrants

Dirigé et conçu par D. Diminescu, l'Atlas e-Diasporas est un effort collectif de recensement, de cartographie et d'analyse de sites Web migrants (Diminescu, 2012a). La recherche et le développement sur les aspects techniques de l'Atlas sont, quant à eux, réalisés par M. Jacomy.

Un **site Web migrant** est un site créé ou maintenu par une ou plusieurs personnes migrantes ou qui est en lien direct avec ces dernières. De manière plus générale, diaspora ou migration doivent être des composantes essentielles de ces sites. Cela peut être un site personnel, un blog, le site d'une association, un portail ou un forum, un site institutionnel ou toute autre support de contenu Web similaire. L'usage de ces sites par des migrants n'est pas un critère de sélection. En effet, un site consulté par un migrant (un média, un réseau social...) n'est pas nécessairement un site migrant. De fait, la prise en compte du caractère *migrant* d'un site s'appuie, d'abord et avant tout, sur le contenu de ce dernier et sur sa production de liens de citations. Enfin, un site migrant n'est pas obligatoirement localisé dans un pays étranger [51], il peut également être hébergé depuis le pays d'origine de son fondateur.

51. Sur la question de localisation d'un site, voir la discussion en Section 5.2.

Actant l'existence de nouvelles formes de représentations en ligne des diasporas et s'appuyant sur la démocratisation du Web au grand public [52], l'Atlas agrège les travaux conjoints de plus de 80 chercheurs, ingénieurs et designers dispersés dans le monde. Ce faisant et après plusieurs années de recherches, ce sont plus de 9000 sites Web qui se voient ainsi documentés et répartis dans pas moins de 30 e-Diasporas [53].

52. Le passage du Web 1.0 au Web 2.0, avec l'apparition des blogs et des réseaux sociaux, fait du Web un espace accessible à des personnes peu ou pas spécialistes des TIC.

53. Voir la liste complète http://www.e-diasporas.fr/#workingpapers

Une **e-Diaspora** n'est pas seulement un outil d'analyse, c'est un véritable concept, fruit du travail collectif de l'ensemble des contributeurs de l'Atlas (Diminescu, 2012b). L'e-Diaspora est l'analogie rapportée

au Web de la notion classique de diaspora, soit des personnes issues d'une même culture ou territoire, dispersées de par le monde, mais maintenant malgré tout des liens, faisant réseau. Une e-Diaspora est un collectif migrant qui s'organise, d'abord et avant tout, sur le Web, une communauté dont les pratiques sont déterminées et portées par les possibilités d'interactions numériques qu'offre le Web. Une e-Diaspora est un collectif dispersé, hétérogène et instable, sa forme changeant chaque fois qu'un nouvel élément est intégré. Les grandes directions de son développement ne sont pas figées, mais reconfigurées à mesure que le collectif évolue. Ces transformations sont le fait de processus individuels et non collectifs, un nouveau membre est intégré ou s'intègre à la l'e-Diaspora en ajoutant ou en supprimant un simple lien de citation hypertexte sur son site.

En effet, d'un point de vue pratique, une e-Diaspora est un réseau orienté de sites Web, liés les uns aux autres par des liens de citations hypertextes. Le sens de ces liens a une importance puisqu'il permet, en partie, de déterminer le rôle joué par un individu dans la communauté : un site établi de longue date et faisant autorité dans le collectif sera pointé par nombre de membres (Figure 1.9, a) ; au contraire, un site cherchant à s'intégrer à une e-Diaspora commencera par citer des membres existants (Figure 1.9, b) dans l'idée de se faire reconnaitre, accepter... Selon le point de l'observateur, un lien hypertexte sera soit considéré comme **sortant** (il est émis par le site témoin) ou comme **entrant** (il pointe le site témoin depuis un autre site).

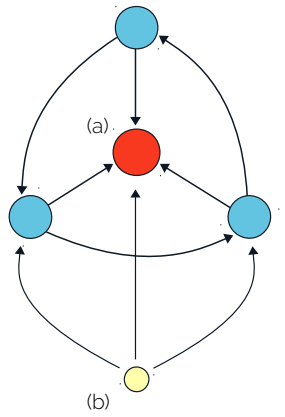


FIGURE 1.9: Exemple de réseau orienté, (a) fait autorité, (b) cherche à intégrer la communauté

Cette manière d'analyser la valeur d'un lien de citation permet de définir les contours d'une géopolitique de l'hypertexte (Rogers *et al.*, 2013), les e-Diasporas étant à la fois *online* et *offline*. Aussi, ce que nous voyons sur le Web peut tout à fait être le reflet d'une situation qui trouve ses origines hors-ligne. L'analyse des différentes e-Diasporas trace des ponts entre le Web et le réel. Chaque e-Diaspora doit être étudiée tout autant au regard de ce qu'elle est en ligne (c.-à-d., un réseau de sites Web), que vis-à-vis des connaissances accumulées sur la diaspora hors-ligne et dont elle est le pendant numérique.

Outils et méthodologie

Explorer, documenter, cartographier et visualiser un tel corpus de sites migrants nécessite le développement d'outils dédiés à chacune de ces tâches. Pour ce faire, les ingénieurs et les chercheurs investis dans la création de l'Atlas définissent ensemble une chaîne inédite d'analyse socio-technique (Figure 1.10).

À chaque e-Diaspora est d'abord assigné un chercheur spécialiste des communautés concernées (un sociologue, un historien...), sa tâche initiale consistant à **explorer** le Web afin de sélectionner un premier en-

semble de sites migrants. L'unité de base de toute exploration est donc le site Web, et non la page ou tout autre ressource postée sur la toile. Ce faisant, le chercheur se sert de son expertise et de ses connaissances du terrain pour identifier les **sites sources** depuis lesquels il constitue petit à petit son corpus. Partant de ces sites sources, le chercheur utilise le logiciel Navicrawler[54] pour cheminer à travers le Web, en suivant les liens de citations sortants et en sélectionnant à la volée ce qu'il considère être des sites migrants.

54. Outil de collecte de sites Web, conçu dans le cadre de l'Atlas e-Diaspora, présenté comme une extension au navigateur Firefox (`https://frama.link/q3HWA2JU`).

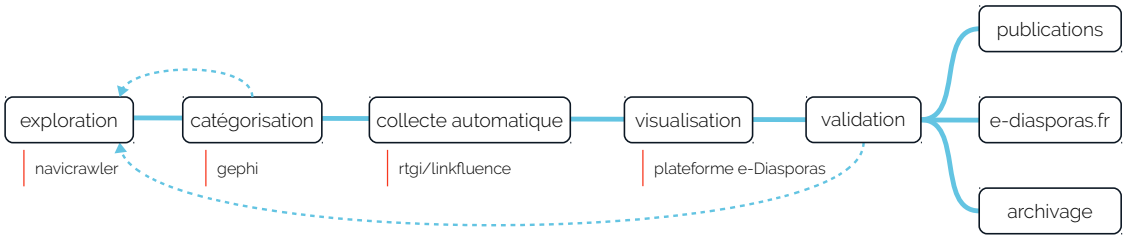


Figure 1.10: Chaîne d'analyse socio-technique de l'Atlas e-Diasporas (source : M. Jacomy)

Une fois satisfait de son corpus, lorsque son exploration converge enfin, le chercheur exporte sa collecte dans le logiciel de visualisation de graphes Gephi[55]. Là le réseau est spatialisé (une place est attribuée dans l'espace à chaque site) et un système de détection de communautés est appliqué[56]. Mais ajoutons que durant l'exploration du Web avec Navicrawler, le chercheur peut, s'il le souhaite, ajouter une ou plusieurs catégories aux sites visités (type de sites, langue, thématique...), et ce, en se basant sur le contenu affiché à l'écran.

55. Outil open-source de spatialisation et d'analyse de graphes, créé spécifiquement dans le cadre de l'Atlas, développé par M. Jacomy, M. Bastian et S. Heymann `https://gephi.org/`

56. Voir la Section 5.3 pour une discussion sur la détection de communautés dans Gephi

Aussi, le passage par Gephi est l'occasion d'affiner cette première **catégorisation** et de la questionner au regard de la spatialisation du réseau et des communautés de sites détectées. Ce faisant et au terme de cette étape, chaque corpus est ainsi constitué d'un ensemble de sites migrants et de liens de citations enrichis par le savoir du chercheur et les indices révélés par les différents logiciels utilisés.

À ce niveau, les corpus sont confiés à la société Linkfluence[57] qui réalise une **collecte** large et automatique du Web afin de consolider la forme et la structure des futures e-Diasporas. Il s'agit également de détecter et de capturer les **sites frontières**. En effet, ces sites ne sont pas directement rattachés à une e-Diaspora (c.-à-d., sélectionnés par un chercheur), mais peuvent permettre d'enrichir la compréhension du réseau par la connaissance de son voisinage direct.

57. `https://linkfluence.com/fr/`

Les réseaux de sites sont, ensuite, **visualisés** sur la plateforme en ligne de l'Atlas[58]. Par convention, les e-Diasporas sont visualisées sous la forme de graphes dans lesquels les sites Web sont représentés par des cercles pleins et les liens de citations sont figurés par des lignes (fléchées ou non).

58. `http://www.e-diasporas.fr/`

S'ensuit l'étape de **validation** qui marque le moment important où chercheurs et ingénieurs se confrontent aux réseaux, les annotent, les corrigent et commencent à s'en servir comme outil de travail sur le terrain (Figure 1.11). Car une e-Diaspora (comme toute forme de visualisation) n'est pas une fin en soi, mais un moyen d'enrichir et de questionner, par le prisme du Web, notre connaissance des communautés concernées. C'est là tout le caractère inédit de la méthodologie développée pour l'Atlas : faire des allers-retours constants entre données du Web et savoir des chercheurs, entre traitements automatiques et validations manuelles. Il s'agit, à tout moment, de pouvoir revenir à l'étude de la dimension sociale des traces observées.

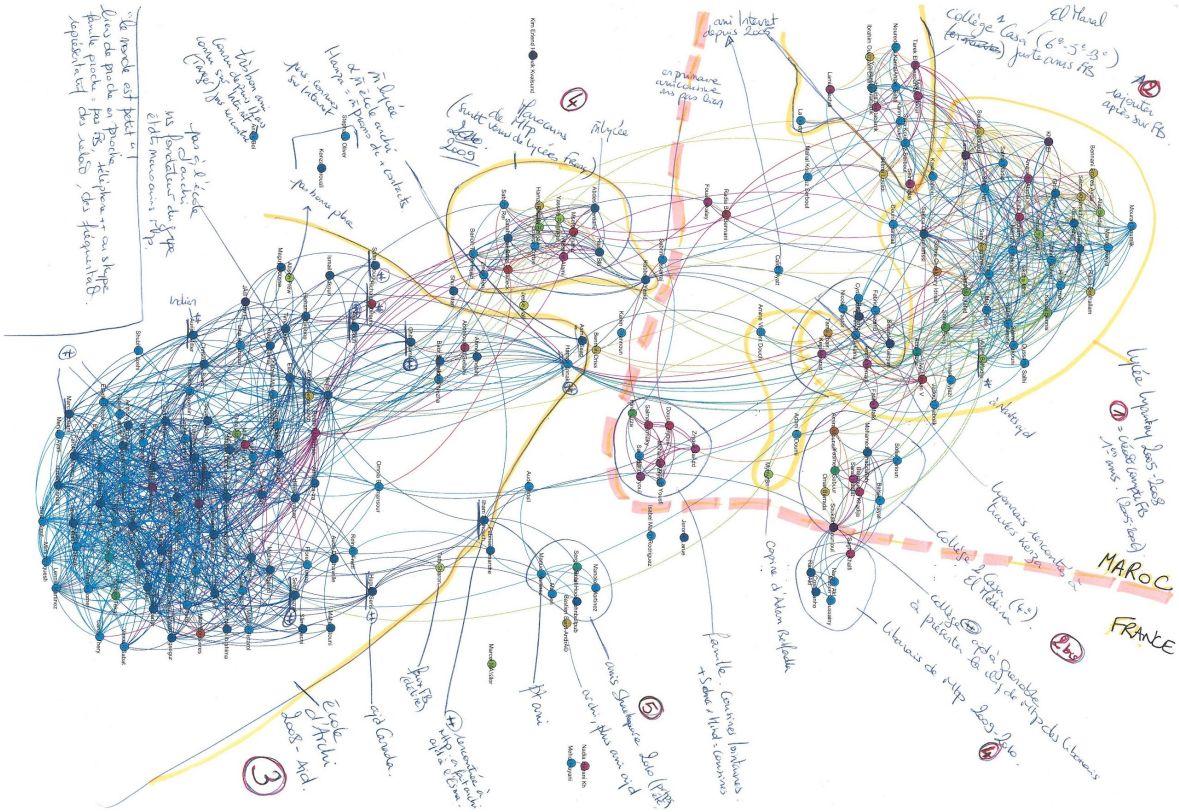


FIGURE 1.11: Annotation de l'e-Diaspora des étudiants Marocains sur Facebook, par S. Marchandise (Marchandise, 2014)

Résultats et interprétations

Chaque e-Diaspora peut être analysée suivant deux grands niveaux de lecture : **1)** le niveau topographique nous permet de considérer les connexions et les proximités entre les acteurs cartographiés. Certains

réseaux se présentent ainsi comme un ensemble de grappes d'entités dispersées (e-Diaspora indienne, Figure 1.12, b), d'autres (e-Diaspora palestinienne, Figure 1.12, a) restent plutôt centrés autour de sites faisant autorité (Ben-David, 2012) ; **2)** le niveau quantitatif/qualitatif est une restitution des informations obtenues lors de l'exploration des sites par les chercheurs et lors de confrontations avec leurs travaux sur le terrain. Ce niveau regroupe toutes les connaissances venues enrichir la constitution des réseaux.

Les différentes parties de l'Atlas sont à l'image de leurs diasporas de rattachement. Les réseaux permettent de saisir la complexité de la dispersion de chaque communauté, tout autant que leur formes particulières de représentation et d'organisation en ligne. Les e-Diasporas brossent ainsi un portrait singulier du Web et de sa prise en main par les populations migrantes : sites Web, forums en ligne (e-Diaspora marocaine [59]), blogs (e-Diaspora tunisienne [60]) et réseaux sociaux (e-Diaspora des étudiants marocains sur Facebook [61]) sont tous représentés, certains même depuis le début des années 90. De plus et bien que majoritairement hébergés sur des serveurs aux États-Unis, le lieu depuis lequel on alimente un site est souvent affiché de manière visible (e-Diaspora népalaise [62]), réinjectant de fait une dimension géographique à la lecture de ces réseaux.

59. https://frama.link/vWpTQ3gb
60. https://frama.link/JtnjZJf0
61. https://frama.link/QyjKMuqB
62. https://frama.link/FSP7-6bx

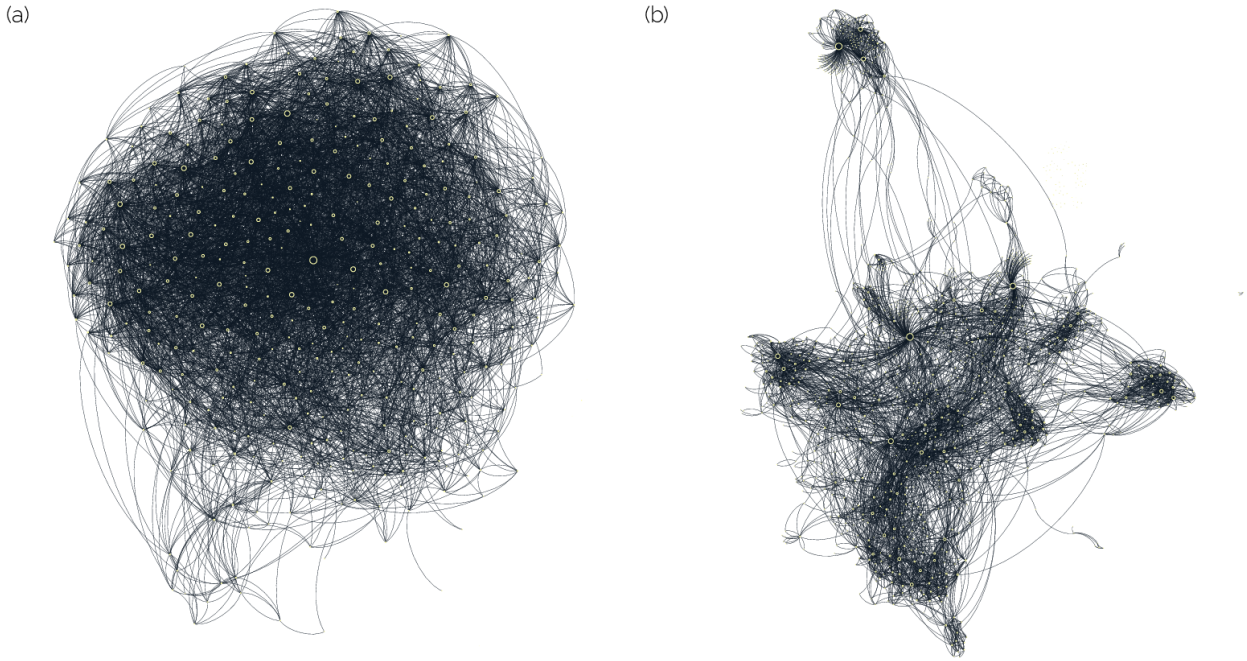


Figure 1.12: L'e-Diaspora palestinienne (a, par A. Ben-David) et l'e-Diaspora indienne (b, par E. Leclerc)

Le Web devient alors un espace transnational où peuvent croitre des

imaginaires collectifs jouant, parfois, un rôle performatif pour certaines diasporas en formation (Bruslé, 2012). Lire une e-Diaspora permet également de comprendre les relations qu'entretiennent les émigrés avec leur pays d'origine et en particulier la proximité entre sites d'expatriés et sites étatiques. Dans le cas de l'e-Diaspora française [63], ces liens sont forts et réciproques (Berthomière, 2013), contrairement à l'e-Diaspora mexicaine [64] où la connexion ne se fait que dans un sens (de l'état vers la diaspora). Une lecture plus fine peut enfin être engagée en associant des sites à des communautés régionales ou religieuses comme le propose E. Leclerc pour la e-Diaspora indienne (Leclerc, 2012) révélant, par la spatialisation du réseau, des tensions réelles entre chaque groupe d'activistes.

63. https://frama.link/_K718ZHp

64. https://frama.link/JyWsTG1V

Pour terminer, rappelons que toute e-Diaspora est forgée en premier lieu par un chercheur et par sa connaissance personnelle du terrain. Ce faisant, il pourrait être légitime de remettre en cause la partialité et la subjectivité des réseaux résultants. Néanmoins, en comparant deux corpus palestiniens [65], réalisés en parallèle par deux chercheurs différents, on se rend compte que ces derniers se recoupent et se recroisent en grande partie.

65. e-Diaspora palestinienne n°1 : https://frama.link/Qr3MbQe3,e-Diaspora palestinienne n°2 : https://frama.link/6jXd-ndV

Les chercheurs sont essentiels à la constitution de l'Atlas, car toute l'analyse ne peut pas être déléguée aux machines et aux traitements automatiques. À titre d'exemple, nous savons que le Web est un mélange d'éléments textuels, audios, vidéos ou logiciels. Or il peut arriver que certains indices clés d'une analyse ne soient disponibles qu'à travers des images ou des bannières. Aussi, sachant que le traitement automatique de ces dernières est très couteux (Ben-David *et al.*, 2018) et que certaines communautés utilisent énormément d'éléments visuels sur leurs sites, l'interprétation humaine doit rester au cœur de la chaîne méthodologique.

L'e-Diaspora Marocaine

Dans la suite de cette thèse, nous nous intéresserons plus particulièrement à l'e-Diaspora marocaine [66] et à son corpus de sites et de blogs, sélectionnés par D. Diminescu et M. Renault en 2008. Support de nos explorations à venir, l'e-Diaspora marocaine présente plusieurs avantages permettant de l'intégrer facilement comme base de travail. Tout d'abord, nous conservons une proximité directe avec les chercheurs ayant cartographié ce réseau, pouvoir échanger régulièrement sur la base d'hypothèses ou de résultats étant un avantage non négligeable. Par ailleurs, lors de la cartographie initiale des sites, D. Diminescu et M. Renault avaient établi des contacts avec les créateurs de certains blogs marocains. Il pourrait, par exemple, être intéressant de réactiver ces connexions pour de futurs entretiens.

66. http ://www.e-diasporas.fr/wp/moroccan.html

L'e-Diaspora marocaine a, ensuite, la particularité d'agréger un grand nombre de sites Web écrits en français. Bien que nous prévoyons un certain nombre de traitements automatiques (applicables à un large panel de langues), il ne faut pas négliger la part importante de travail manuel à réaliser. Être locuteur de la langue d'un site reste donc un avantage. Enfin, cette e-Diaspora peut faire l'objet d'une lecture topographique relativement claire, présentant différents espaces et différentes communautés de sites, amenant chacun et chacune vers plusieurs pistes de réflexion.

Le corpus marocain est composé de 156 sites Web et de 397 liens de citations hypertextes. Majoritairement écrits en français et en anglais, les sites se répartissent en trois grandes catégories : **1)** sites d'associations et d'ONG (50 sites), **2)** blogs (47 sites), **3)** sites institutionnels (22 sites). Leurs thématiques sont diverses et variées, mélangeant développement, citoyenneté et politique ou encore sujets intimistes et billets d'humeur (voir la Figure 1.13 pour un comptage complet).

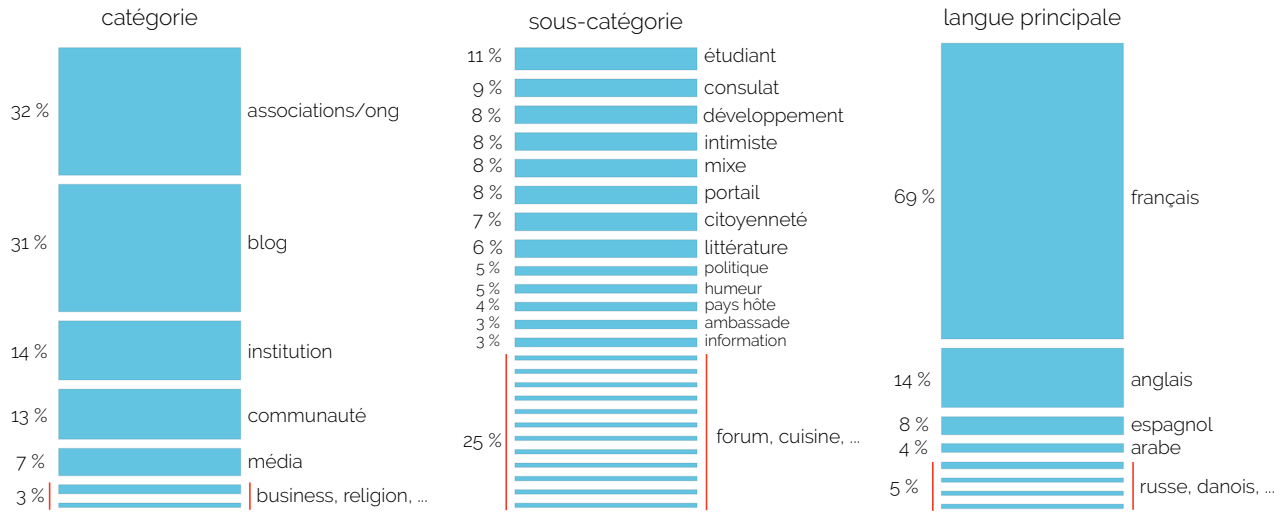


Figure 1.13: Répartition des sites de l'e-Diaspora marocaine par catégorie, sous-catégorie et langue principale (source : Atlas e-Diaspora `https://frama.link/Pyt3Aq2x`)

La Figure 1.14 représente la carte de l'e-Diaspora marocaine sous la forme d'un graphe : les nœuds sont les sites, les arcs sont les liens hypertextes et la taille des nœuds est proportionnelle au degré de chaque site[67]. D'un point de vue de la topologie du réseau et de sa seule structure, nous pouvons observer un premier groupe de sites (Figure 1.14, a), non liés entre eux et très isolés du reste du graphe. Ce sont principalement des sites d'associations qui, au regard de leur production de liens de citations, semblent s'ignorer mutuellement. Ce faisant, la carte commence à nous interroger et à tracer des hypothèse entre le Web et le réel. Cette méconnaissance entre sites associatifs

67. Le degré peut ici être compris comme une mesure d'importance. Voir la Figure 2.6 et la Section 2.2 pour une explication plus détaillée de la notion de degré.

est elle uniquement visible sur le Web ou peut elle être vérifiée sur le terrain ? Pourquoi les associations refuseraient-elles de faire collectif ?

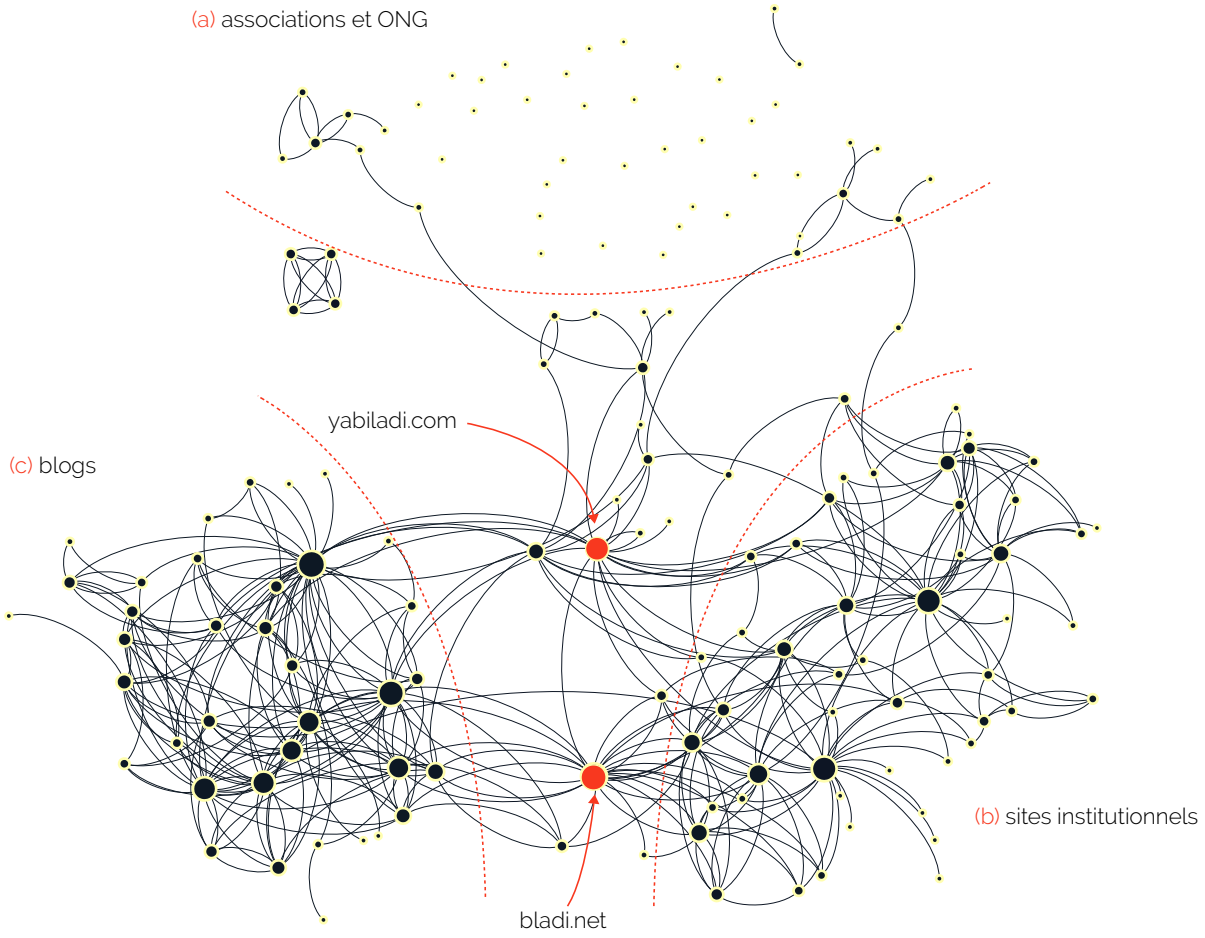


FIGURE 1.14: Cartographie de l'e-Diaspora marocaine (Par D. Diminescu & M. Renault)

Deux autres groupes de sites sont également très clairement visibles (Figure 1.14, b & c), mais présentent un profil différent. Très fortement liées entre elles, chacune de ces communautés de sites semble en revanche détachée du reste du graphe, toujours un peu à l'écart. Le premier groupe (Figure 1.14, b) est ainsi composé de sites institutionnels (consulats, ambassades...), formant un ensemble dense et homogène. Mais, théoriquement construits par l'État marocain à destination de ses ressortissants vivant à l'étranger, il est frappant de voir que ces sites ne sont pas très bien connectés au reste du réseau et encore moins à la seconde communauté de sites. En effet, ce dernier ensemble (Figure 1.14, c) témoigne tout particulièrement de la présence en ligne des migrants marocains, puisqu'il est principalement constitué de blogs personnels d'émigrés.

Les sites institutionnels donnent, ainsi, l'impression d'échouer dans leur volonté (pourtant affichée) de structurer et d'organiser la vie de la diaspora, en polarisant l'attention autour d'eux (tout au moins sur le Web). Ils ne sont pas centraux, ce rôle est déjà occupé par deux autres sites (Figure 1.14, points rouges), faisant le pont entre chacun des trois groupes de l'e-Diaspora marocaine. Cet état de fait est-il nouveau ? Peut-être que ces sites viennent tout juste de s'installer ? L'État marocain est-il au courant de cette situation ? Pourquoi les blogs ne se connectent-ils pas aux sites institutionnels, est-ce de l'ignorance, voire de l'indifférence ou est-ce fait en conscience ?

De leur côté, les sites *yabiladi.com* et *bladi.net* occupent une position d'autorité étant, de fait, plutôt bien cités. Ce sont des portails communautaires, bien établis dans l'écosystème et dans l'histoire longue de l'e-Diaspora marocaine [68]. Leurs forums en ligne attirent notamment plusieurs centaines de visiteurs chaque jour. Enfin, l'utilisation de la langue anglaise permet de faire un dernier lien [69] entre les groupes institutionnels et l'ensemble des blogs, plusieurs sites anglophones étant ainsi situés autour de *bladi.net*

68. Le site *yabiladi.com* et le groupe de blogs migrants feront l'objet d'explorations dédiées au Chapitre 5 de cette thèse.

69. Voir la carte de la distribution des sites par langue `https://frama.link/2hFDT02Q`

***

Ces premières interrogations ne sont que des exemples pris parmi l'ensemble des problématiques soulevées par l'Atlas. Chaque nouvelle observation attentive d'une e-Diaspora amenant son lot de pistes à explorer, soit qu'elles portent sur une dimension singulière d'une population migrante, soit qu'elles interrogent plus largement le Web en tant qu'écosystème. Par exemple, les communautés de sites, révélées par l'e-Diaspora marocaine, suivent-elles un comportement particulier à ce seul réseau ? Ou peut-on dire que la forte connectivité des blogs entre eux est un fait général du Web ?

Aussi, les e-Diasporas doivent être comprises et analysées vis-à-vis de trois aspects : **1)** le Web et ses règles propres, **2)** les connaissances rassemblées en amont et en lien avec la diaspora concernée, **3)** les retours et les confrontations à postériori avec le terrain sociologique ou historique. Mais une grande famille de questions semble pourtant difficile à adresser, et ce partant uniquement des e-Diasporas telles que nous les avons présentées. Comment prendre en compte le temps long, l'éphémère, les dynamiques et autres événements dans l'analyse de ces réseaux ?

Car les e-Diasporas ne sont, au final, que des images fixes, des captures instantanées d'une petite partie du Web. Or, les pages et les sites évoluent et se transforment en continu, rendant rapidement obsolètes les cartes réalisées pour l'Atlas. Ainsi, lorsqu'après dix années d'attente, nous revenons enfin visiter chacun des sites collectés, force est de constater qu'une grande majorité d'entre eux ont disparu, ont

été effacés ou ne sont plus ce qu'ils étaient.

Mais anticipant le caractère éphémère du Web, les chercheurs coordonnant les travaux de l'Atlas décident d'ajouter une dernière brique à leur chaîne méthodologique (Figure 1.10) : l'**archivage**. Cette volonté de garder une trace des sites observés était même antérieure à la création de l'Atlas puisque la question de l'archivage fut posée dès 2004. Aussi et afin de préserver la tenue de recherches futures, portant sur l'histoire et l'évolution de chacune des e-Diasporas, une campagne d'archivage des sites cartographiés est lancée en 2010. Ce faisant, cette thèse et les chapitres à venir auront pour principal objet l'exploration au présent des archives Web de l'Atlas e-Diasporas.

Chapitre 2

Préserver le Web et les e-Diasporas

Face à la disparition totale ou partielle des sites Web recensés par l'atlas e-Diasporas (Section 1.3), il a rapidement été décidé de mettre à l'abri cet héritage numérique. De le préserver. Aussi et dès mars 2010, se met en place, sous la responsabilité technique des équipes de l'Institut Nationale de l'Audiovisuelle (INA), une vaste campagne d'archivage des quelques 9000 sites migrants alors cartographiés. Ces archives Web, aujourd'hui constituées [1], peuvent enfin faire l'objet de recherches et d'explorations. Grâce à ce travail d'archivage, nous pouvons questionner au présent les traces du Web passé.

1. L'archivage s'est officiellement terminé en septembre 2014.

Mais de part la nature même du médium, le Web appelle la mise en place d'un archivage particulier, basé sur des techniques de collectage dédiées. Comment archiver ce qui est, tout autant, un flux continu d'informations qu'un territoire en perpétuelle expansion ? Quels compromis ont du faire les archivistes pour garantir une collecte représentative ? Jusqu'à quel point les archives du Web passé sont-elles fidèles au Web vivant ?

Dans ce chapitre, nous prendrons le point de vue des archivistes. 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. En effet, ces diverses initiatives capturent et stockent, jour après jour, plusieurs centaines de milliers de pages Web. Par exemple, Internet Archive a réalisée, depuis son lancement en 1996, près de 650 000 000 000 captures d'objets Web (pages, images, vidéos, etc.), cela pour un total de 40 Po d'archives. Mais la réussite écrasante d'Internet Archive ne cache-elle pas un mouvement en perte de vitesse ? Qu'en est-il dans le reste du monde ? Où en sont les travaux de recherche portant directement sur les archives Web ?

Pour comprendre la manière avec laquelle ces archives sont constituées, nous introduirons ensuite les principales méthodes de sélection, collecte et stockage des corpus à préserver. Nous présenterons, enfin, les contours des archives e-Diasporas à proprement parler. Leurs particularités et leurs caractéristiques. Leur durée et leur 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 3) qui motiveront la présentation de notre principale contribution (au Chapitre 4). Ce faisant, les éléments que nous nous apprêtons à présenter s'appuient principalement sur la lecture de l'ouvrage de J. Masanès : *Web Archiving* (Masanès, 2006b) qui reste, encore aujourd'hui, une référence. Nous compléterons et mettrons à jour, au besoin, ces informations.

2.1 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 [2], les acteurs français de la pratique. Alors qu'étaient évoquées 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 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 [3].

2. http://www.bnf.fr/fr/professionnels/anx_journees_pro_2016/a.jp_161122_23_archivage_web.html

3. http://multimedia.bnf.fr/video/prof/161123_10_dl_web.mp4

Cette querelle initiale et son dénouement (la cotutelle du dépôt légal) 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.

Quand la technique devient mémoire

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 [4]. 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.

4. 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 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, p.60). 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 cours 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 [5] en France, par l'ordonnance royale du 28 décembre 1537, à des fins de préservation culturelle mais également de contrôle politique.

Les siècles 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 [6], 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 [7] 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) :

— En 1971, le projet Gutenberg commence à collecter des copies

5. « *Nous avons délibéré de faire retirer, mettre et assembler en notre librairie toutes les œuvres 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 et Guilbaud, 1960).

6. Cinématographie, photographie, radiodiffusion, etc.

7. Bases de données, logiciels, interfaces, etc.

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 grec ancien et toujours subsistant.

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

Nous appelons **initiative d'archivage** tout projet d'archivage mené ou piloté par une personne seule, un collectif, une institution, etc. L'archivage du Web débute à la fin des années 90, et plus précisément en 1996 lorsque 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 1.1). La National Library of Australia est ainsi à l'initiative du projet Pandora [8] qui vise à archiver les publications en ligne australiennes sur la base d'une collecte sélective et continue de sites Web australiens. La Swedish Royal Library, quant à elle, lance le projet Kulturarw3 [9] qui s'essaye à une collecte *intégrale* et espacée dans le temps des sites du Web suédois (Arvidson *et al.*, 2000).

8. `http://pandora.nla.gov.au/`

9. `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 d'une organisation à but non lucratif afin de rendre accessible au plus grand nombre le passé du Web [10]. 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.

10. `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 [11] (Hallgrinsson et Bang, 2003). En 2003, la publication par l'UNESCO de la *Charte sur la conservation du patrimoine numérique* (UNESCO, 2003) marque un nouveau tournant pour l'archivage du Web en reconnaissant la valeur universelle d'une telle démarche et l'urgence face à la disparition potentielle de tout, ou d'une partie, de l'héritage numérique mondial :

11. *Nordic Web Archive*

« *Le patrimoine numérique mondial risque d'être perdu pour la postérité. Les facteurs qui peuvent contribuer à sa perte sont l'obsolescence rapide du matériel et des logiciels qui servent à le créer, les incertitudes concernant les financements, la responsabilité et les méthodes de la maintenance et de la conservation et l'absence de législation favorable à sa préservation. L'évolution des attitudes n'a pas suivi celle des technologies. L'évolution numérique a été trop rapide et trop coûteuse pour que les pouvoirs publics et les institutions élaborent en temps voulu et en connaissance de cause des stratégies de conservation. La menace qui plane sur le potentiel économique, social, intellectuel et culturel du patrimoine, pierre angulaire de l'avenir, n'a pas été pleinement saisie.* » — Charte sur la conservation du patrimoine numérique, Article 3, (UNESCO, 2003)

Pour de nombreuses bibliothèques nationales, la charte de l'UNESCO fait l'effet d'un accélérateur (Figure 2.1). Les institutions sont encouragées dès 2003 à archiver leur Web national (Gomes *et al.*, 2011). Mais notons ici que la notion de Web national reste discutable (Abiteboul *et al.*, 2002), il s'agira souvent de crawler le Web en fonction d'une extension de nom de domaine donnée (.fr, .jp, .uk, etc.), extension qui ne couvre pas exhaustivement l'ensemble des sites associés à un domaine national précis, elle est plutôt à considérer comme une borne inférieure de celui-ci (Koehler, 1999). Le cas des corpus de l'Atlas e-Diasporas en est un très bon contre-exemple, nombre de sites migrants possédant une extension générique (.com, .net) ou correspondant au pays d'accueil plutôt qu'au pays d'origine (Leclerc, 2012).

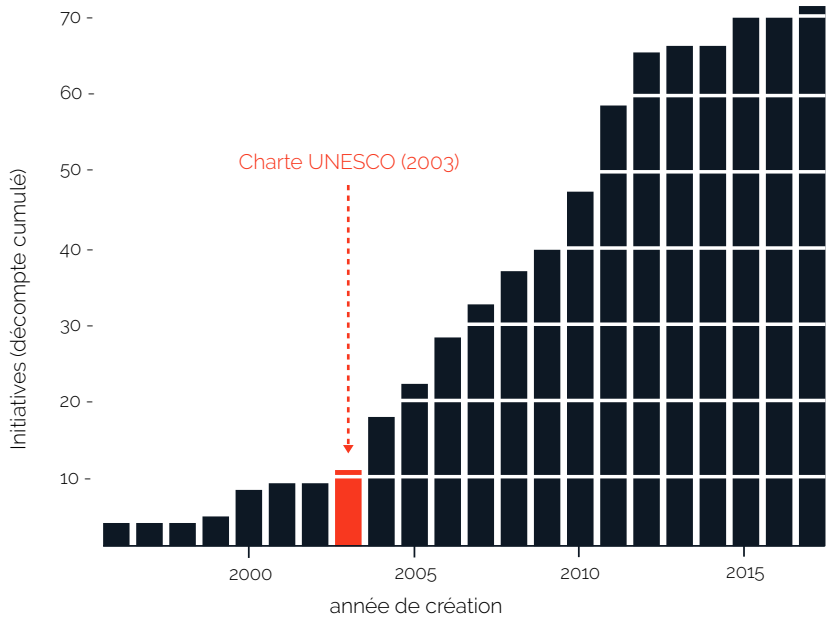


FIGURE 2.1: Évolution cumulée du nombre d'initiatives d'archivage du Web par année de création (sources : (Gomes *et al.*, 2011) & `https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives`)

Ces nouveaux acteurs de l'archivage du Web peuvent être classés suivant la terminologie introduite par J. Masanès (Masanès, 2006b, p.76),

entre initiatives publiques ou privées, poursuivant un but lucratif ou non. L'accès aux corpus archivés peut être entièrement public ou restreint et limité, en ligne ou physique (machine de consultation dans une bibliothèque). Par exemple, Internet Archive est une initiative à but non lucratif, avec un accès public à l'ensemble de ses corpus en ligne [12] depuis 2001.

12. La Wayback Machine est officiellement lancée en 2001. Avant cette date les corpus d'Internet Archives n'étaient pas accessibles au public. En 2002, une copie intégrale des archives est consultable à la Bibliotheca Alexandrina, en Égypte.

Une autre manière de catégoriser ces initiatives est de regarder la nature des corpus archivés. Nous avons déjà évoqué les corpus territoriaux censés capturer les contours d'un Web national. Cette notion peut être également transposée à plus fine échelle : celle d'une région ou d'une ville (Boudrez et Van den Eynde, 2002). Un corpus d'archives peut être conçu pour cibler une thématique donnée, souvent centrée sur des événements politiques (Voerman *et al.*, 2002; Schneider *et al.*, 2003) : élections, référendums, etc. Certaines initiatives s'affranchissent même de barrières géographiques devenues contraignantes en préservant des sites de domaines nationaux étrangers (Gomes *et al.*, 2009).

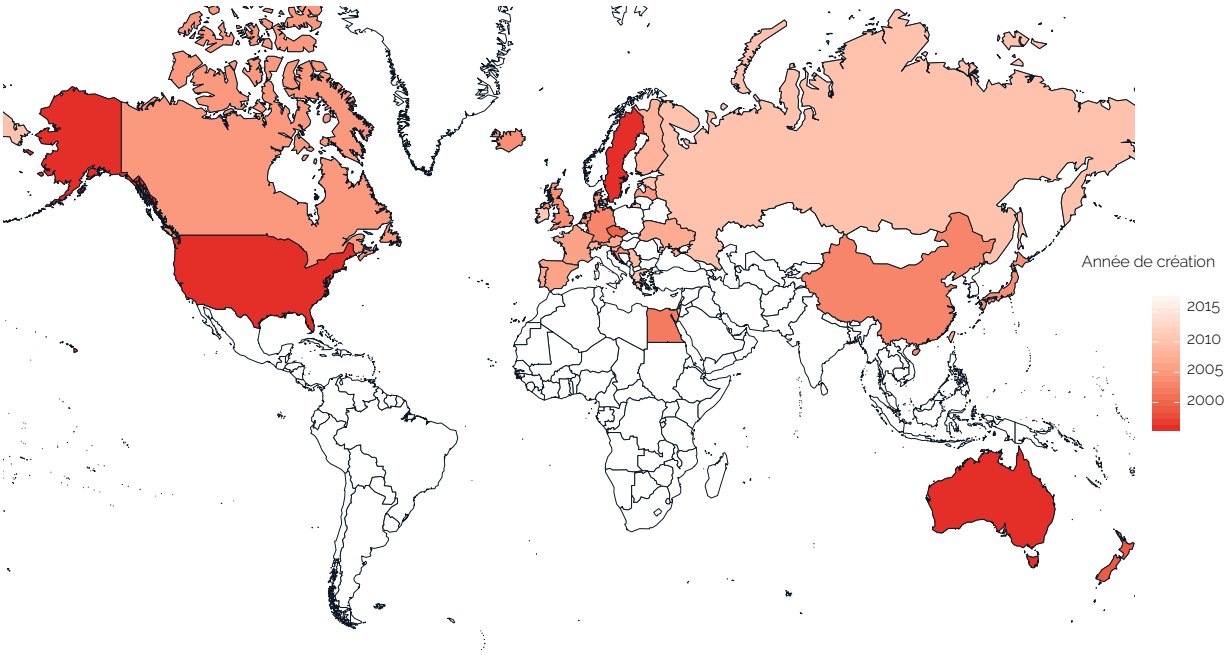


Figure 2.2: Carte des initiatives d'archivage du Web par pays et années de création (sources : (Gomes *et al.*, 2011) & `https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives`)

Enfin, il est possible de voir les corpus d'archives Web par rapport à l'utilisation que l'on en fait. Certains sont ouvertement tournés vers la consultation publique (Internet Archive, à nouveau), d'autres sont constitués à des fins universitaires (on pense au corpus japonais WARP [13] de la National Diet Library). La British Library, de son côté, fait de ses archives Web une utilisation détournée, voire cachée, en char-

13. `http://warp.da.ndl.go.jp/search/`

geant les versions passées des pages Web des sites gouvernementaux (par exemple, *gov.uk*) si celles ci sont momentanément ou définitivement inaccessibles [14]. L'un des plus gros corpus d'archives Web reste en revanche celui détenu par Google qui permet d'accéder depuis le cache [15] de son moteur de recherche à une version précédemment collectée d'une page. Notons, pour terminer ce tour d'horizon, que même si elles sont nombreuses de part le monde, les initiatives d'archivage du Web sont historiquement et géographiquement le fait d'états occidentaux (Figure 2.2). Les continents sud-américain et africain (hormis la Bibliotheca Alexandrina) sont absents de ce paysage, rendant encore plus précieux les corpus transnationaux archivés par l'Internet Archive et l'Atlas e-Diasporas. Notons, enfin, que certains corpus peuvent être dédiés à un ou plusieurs types précis de ressources Web. L'initiative Software Heritage [16] archive les logiciels et portions de codes informatiques régulièrement publiés en ligne.

En vingt années d'existence, les archives du Web ont agrégé autour d'elles une communauté de chercheurs et d'ingénieurs participant à sa promotion. L'International Internet Preservation Consortium (IIPC) est fondé en 2003 dans l'idée de proposer des rapports et des suivis réguliers de l'archivage [17] et des workshops sont organisés (l'IWAW, International Web Archiving Workshops). Mais le fait est de constater que la dynamique visible à la fin des années 2000 est en train de se tasser. En 2017, une seule et unique initiative a vu le jour (en Belgique autour du projet Promise [18]) et les vastes projets de recherche et d'exploitation des archives Web que sont ARCOMEM (Risse *et al.*, 2014), LAWA (Spaniol et Weikum, 2012) et LIWA (Denev *et al.*, 2009) n'ont pas trouvé de successeurs. Enfin, rappelons que la position centrale d'Internet Archive dans cet écosystème ne la met pas à l'abri d'une possible disparition. En 2016, B. Kahle prévoyait (notamment en réaction à l'élection de D. Trump à la tête des États Unis) de déplacer une nouvelle copie intégrale d'Internet Archive au Canada [19]. Et n'oublions pas que, jusqu'à présent, la survie d'Internet Archive est et reste étroitement liée à son seul fondateur.

Mais alors que les institutions semblent s'en détourner, le futur des archives Web viendra peut être de ceux que M. Graham, directeur de la WayBack Machine, nomme les « *rogue archivists* » [20]. S'inscrivant dans les pas d'A. Swartz [21], les rogues archivits sont des activistes et libristes militants s'appropriant politiquement la question des archives. Soit qu'ils considèrent le Web et son contenu comme un commun de l'Humanité (Coriat, 2015), soit qu'ils voient dans les archives un moyen de faire perdurer la mémoire de minorités opprimées (De Kosnik, 2016). Ils administrent de manière autonome certains des 7000 robots archivistes qui alimentent quotidiennement Internet Archive. Encore mineure, au regard des volumes globaux d'archives Web, il est néanmoins possible

14. https://www.bl.uk/collection-guides/uk-web-archive

15. https://fr.wikipedia.org/wiki/Mémoire_cache

16. https://www.softwareheritage.org/?lang=fr

17. http://internetmemory.org/images/uploads/Web_Archiving_Survey.pdf

18. https://promise.hypotheses.org/

19. http://blog.archive.org/2016/11/29/help-us-keep-the-archive-free-accessible-and-private/

20. https://youtu.be/33_fnPwaEM0

21. En 2011, A. Schwartz télécharge la base de données de l'éditeur JSTOR afin de « *libérer* » plusieurs millions d'articles scientifiques payants, dont une part importante appartenait au domaine public (voir *The Internet Own Boy* réalisé par B. Knappenberger en 2014 https://archive.org/details/TheInternetsOwnBoyEsp). Les suites judiciaires de cette affaire mèneront au suicide de Swartz. Ses nombreuses contributions, à la fois sur des aspects techniques du Web et sur la question plus politique de l'accès universel aux connaissances, restent considérables.

de déceler la trace de leurs contributions dans les travaux d'A. Ben David (Ben-David et Amram, 2018) qui révèle que les archives du Web nord-coréen (présentes dans le corpus d'Internet Archive) n'ont pas été collectées par des systèmes institutionnels mais par des indépendants, non assujettis à la géopolitique des proxys.

Le cas des archives Web Françaises

En France, l'archivage du Web est l'aboutissement singulier de dix années d'expérimentations techniques et de construction d'un cadre législatif inédit. Aujourd'hui, deux institutions se partagent le périmètre du dépôt légal du Web : la Bibliothèque Nationale de France (BNF) et l'Institut National de l'Audiovisuel (INA).

Crée par François 1er, le **dépôt légal** est l'obligation pour tout éditeur (imprimeur, producteur, importateur, etc.) de déposer chaque document dont il a la charge (en France) à la BNF ou auprès de l'organisme le plus adapté à la nature particulière de ce document. Tout ce qui se publie et s'édite en France est donc directement collecté par la BNF. L'INA, quant à elle, administre les archives de la radio et de la télévision. L'institut fut initialement créé pour en faire une exploitation commerciale destinée aux professionnels de l'audiovisuel. L'État français est l'un des premiers états au monde a avoir posé la question des conditions de la mémoire culturelle et patrimoniale du Web. Le Web devait rentrer dans le périmètre du dépôt légal et c'est ainsi que furent posées les bases d'un futur **dépôt légal du Web**.

Les tractations commencent officiellement en 2001. En s'appuyant sur la directive européenne 2001/29/EC [22], dite *Information Society Directive*, l'Assemblée Nationale ouvre au débat la discussion du *Projet de loi sur la société de l'information* [23]. Cette loi vise à adapter le droit français aux NTIC en matière de libertés de communication, de commerce en ligne, mais également de droit d'auteur. De ces débats découle, en 2006, l'adoption de la loi DADVSI [24] relative *au droit d'auteur et aux droits voisins dans la société de l'information* qui définit le cadre légal des archives Web à venir :

22. https://en.wikipedia.org/wiki/Copyright_Directive

23. http://www.assemblee-nationale.fr/11/projets/pl3143.asp

24. https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000266350

« *Les logiciels et les bases de données sont soumis à l'obligation de dépôt légal dès lors qu'ils sont mis à disposition d'un public par la diffusion d'un support matériel, quelle que soit la nature de ce support. Sont également soumis au dépôt légal les signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique.* » — Loi DADVSI, Article 21

La BNF et l'INA souhaitant toutes deux se voir confier le plein contrôle du dépôt légal de Web par l'État, les équipes de J. Masanès (BNF) et de B. Bachimont (INA) se lancent l'une comme l'autre dans la course

à l'archivage dès le tournant des années 2000, c'est-à-dire bien en amont de tout arbitrage politique. Comme nous le verrons dans la section suivante (Section 2.2), la masse de travail à mettre en place pour débuter une collecte est considérable. L'histoire de l'archive du Web en France est donc tout autant l'aboutissement d'une volonté politique que le fruit d'années de recherches et développements. Une profonde réflexion théorique est ainsi menée sur la nature même du Web en tant que matière à préserver et à restituer après collecte (Bachimont, 2009).

Pourtant d'un point de vue purement technique, les deux institutions suivent des directions divergentes : la BNF et J.Masanès s'associent à la définition du format d'archivage WARC, l'INA et T. Drugeon créent le format DAFF et développent un crawler indépendant [25]. L'INA lance rapidement sa collecte de sites Web de manière expérimentale (Bachimont *et al.*, 2005), alors que l'État s'oriente vers un partage du dépôt légal : une solution à deux têtes. Le cadre de cette partition est défini par le décret du 19 Décembre 2011 [26] qui établit que :

— La BNF archivera l'ensemble du domaine national français et d'outre-mer au moins une fois par an. Par là, sont identifiés tous les sites Web en .fr ainsi qu'une liste blanche de sites en .com, .org, .net... « *édités par des personnes physiques ou morales domiciliées en France* ».

— L'INA archivera un sous-ensemble thématique du Web français centré sur les sites dit *médias* (sites des services et des médias audiovisuels, Web TV et Web radios, programmes radio et TV, professionnels de l'audiovisuel, etc.). La fréquence de collecte sera variable et adaptée à la nature même des mises à jour de ces sites (chaque jour, semaine, mois, etc.)

Si le périmètre de la collecte de la BNF reste classique au regard de ses consœurs mondiales, le corpus archivé par l'INA, lui, est tout à fait singulier (Mussou, 2012). La collecte se concentre sur un jeu de 14.000 sites Web médias (sélectionnés manuellement). Seulement 30% d'entre eux ont une extension .fr contre 50% en .com (Drugeon, 2005). Par ailleurs, l'INA intègre à son corpus des sources vidéos (de youtube et dailymotion dès 2010), des flux RSS, des Tweets (depuis 2014), etc. Les deux institutions offrent également la possibilité à des chercheurs de constituer des corpus tiers et portés sur une thématique précise : l'ANR Web90 est montée en partenariat avec la BNF [27] et porte sur l'étude des premières années du Web français (Schafer et Thierry, 2016), l'INA réalise une collecte dédiée aux attentats de Paris fin 2015 [28]. L'INA est enfin la seule des deux institutions à avoir encore aujourd'hui une équipe technique dédiée à la recherche et à l'exploitation de ses archives Web.

Les corpus de la BNF et de l'INA se veulent donc complémentaires. Ils appartiennent à la catégorie des initiatives publiques mais n'offrant

25. Nous discuterons dans la Section 2.2 des aspects techniques des divers formats d'archivage.

26. https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000025002022&categorieLien=id

27. https://web90.hypotheses.org/

28. https://asap.hypotheses.org/173#more-173

Figure 2.3: Demande d'accréditation pour accéder aux zones de consultations des archives Web à la BNF (site François Mitterand)

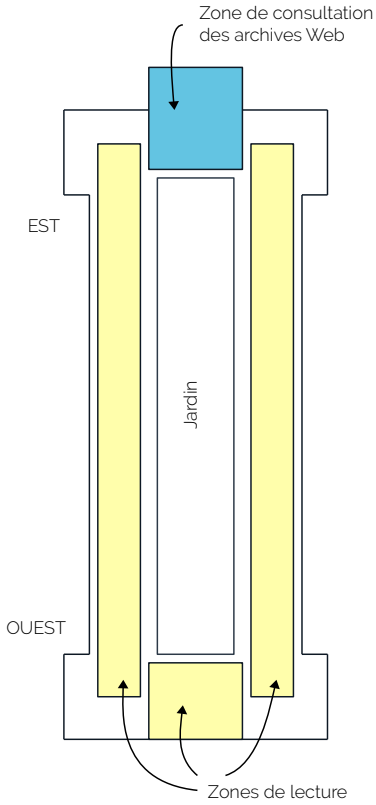


Figure 2.4: Localisation (en bleu) des postes de consultation des archives Web à la BNF (Rez-de-Jardin, site François Mitterand)

qu'un accès physique aux contenus archivés : il n'y a pas de portail en ligne de consultation des archives. Le chercheur doit se déplacer dans l'un des 31 centres locaux de l'INA ou, s'il est a Paris, il reste possible d'accéder aux deux corpus depuis la BNF.

~

Je reprends ici le *je* pour parler d'une expérience personnelle. Courant 2017, je me suis rendu à la BNF (site François Mitterrand) afin d'y consulter les archives de l'INA et de la BNF et tester les modalités d'accès aux corpus. Ce récit ne vaut pas généralité, mais doit être pris comme un témoignage d'une exploration d'une heure trente dans la bibliothèque avant de trouver les archives. Je pensais tout d'abord (en me fiant aux indications du site Web) qu'il était possible de consulter les archives depuis le réseau Wifi du lieu.

Après divers échecs, les bibliothécaires m'ont progressivement fait passer d'interlocuteur en interlocuteur jusqu'à finalement me faire accéder (moyennant une demande d'accréditation, Figure 2.3) à l'une des salles du Rez-de-Jardin de la BNF (Figure 2.4). Là les archives de l'INA ne sont consultables que depuis une poignée de postes labellisés *Inathèque*. Les archives de la BNF, elles, sont accessibles depuis l'ensemble des machines de la zone. Une fois connecté, un moteur de recherche classique nous permet de faire des recherches par URL (pour la BNF) et plein texte (pour l'INA), il n'est en revanche pas possible de sauvegarder ses recherches ou de les exporter d'une quelconque manière. Je me suis donc servi de mon téléphone pour photographier les pages Web qui m'intéressaient.

J'ai été surpris de voir que la consultation des archives Web à Paris relève d'une véritable expédition et je remercie les bibliothécaires d'avoir finalement su me guider. Mais s'il ne faut pas faire grief de leur méconnaissance, force est de constater que nous ne devons pas être très nombreux à consulter les archives Web là bas et que celles ci ne sont pas particulièrement mis en avant auprès du public et du personnel.

~

Les conditions d'accès restreintes aux corpus de l'INA et de la BNF ne jouent pas en la faveur d'une démocratisation de leur exploitation. Auprès du grand public d'une part, mais également vis-à-vis de potentiels chercheurs voulant questionner les archives. S'il reste tout à fait possible de venir étudier une liste prédéfinie d'URL et de sites (en lecture seule), les modalités d'accès n'encouragent pas à l'exploration des corpus ni à la possibilité de mener des recherches larges et/ou automatisées (il n'y a pas d'API[29] par exemple).

29. Point d'entrée à une source de données depuis l'extérieur de son environnement de stockage.

2.2 Sélectionner, collecter et consulter les corpus

D'un point de vue purement technique, il a fallu penser de toute pièce de nouveaux processus d'archivage. Alors que la préservation du Web a rapidement été considérée comme une nécessité, des équipes de pionniers ont rassemblé des connaissances éparses et les ont conjuguées pour créer les outils de sélection, collecte et fouille des futurs corpus d'archives Web.

Un objet éphémère et multiforme

À l'époque où le Web ne contenait qu'une poignée de pages et de sites, la question de son auto-préservation[30] fut posée. Le Web s'archivait-il déjà de lui même ? Pourquoi faire intervenir un archiviste ?

Lorsqu'un nouveau contenu est publié, il est possible de reléguer en bas de page les éléments plus anciens, à la manière d'une pile. Les **content management systems** (CMS[31]), apparus avec l'essor des blogs, avaient ainsi pour vocation de stabiliser ce processus de création et suppression de contenus en ligne, en reléguant les publications passées dans une section dédiée. Rien n'empêche également un site Web d'être copié dans son intégralité puis redéployé sur de nouveaux serveurs pour le préserver. Le numérique facilite et rend possible la reproduction à l'infini de ses objets. Ainsi, une page Web aurait théoriquement pu ne jamais disparaître de la toile. Mais il fut montré, dès le début des années 2000, que, malgré ces possibilités, le Web restait un milieu hautement instable. La disparition de contenu est un phénomène inhérent au Web.

La durée de vie d'un site Web, peut être calculée par rapport à la mesure de sa demi-vie, soit la durée qu'il faut pour que la moitié de son contenu (ici ramené au nombre de pages[32]) disparaisse du Web (Koehler, 2004). Dès 1999, la demi-vie moyenne d'un site Web est ramenée à 50 jours (Cho et Garcia-Molina, 1999), estimation qui doit être relativisée par rapport au contexte de publication du site et à la nature de ses pages (McDonnell *et al.*, 1999; Fetterly *et al.*, 2003). De même, 80% des pages Web collectées entre 2003 et 2004 par les archives du Web japonais ont été effacées en moins d'un an du Web vivant (Toyoda et Kitsuregawa, 2006). Pour préserver le Web il faut donc intervenir et archiver avant qu'une page ne soit détruite. Ainsi, avant toute collecte, l'archiviste doit prendre en compte les changements susceptibles d'intervenir sur une page ciblée, afin de minimiser la perte d'information.

Les changements subis par une page Web au cours de son existence sont multiples (Douglis *et al.*, 1998; Adar *et al.*, 2009), allant de la modification de son contenu jusqu'à une évolution de la structure des liens qui la relie au reste du site. La fréquence de changement

30. traduit ici de l'anglais *self preserving* (Spinellis, 2003)

31. Système permettant de gérer automatiquement et de manière dynamique le contenu d'un site Web, comme Wordpress, Drupal, Joomla !, etc.

32. La demi-vie peut être appliquée à d'autres objets numériques comme des bases de données ou des documents scannés. Cette mesure est elle même dérivée des techniques de datation des atomes.

d'une page peut être estimée et prédite en s'appuyant sur des versions précédemment archivées (Chawathe et Garcia-Molina, 1997; Khoury *et al.*, 2007). Il est possible d'affiner l'estimation de cette fréquence en catégorisant les changements par types (structurels, sémantiques ou cosmétiques) (Yadav *et al.*, 2007). Plutôt que de considérer chaque page indépendamment les unes des autres, les changements peuvent être détectés à l'échelle d'un site ou d'un réseau. On s'appuiera alors sur la présence de liens hypertextes (Liu *et al.*, 2000) ou sur des relations hiérarchiques plus marquées (Lim et Ng, 2001). Dans la suite de ce manuscrit, nous nous limiterons à considérer comme changements les seuls actes de création, de modification ou de suppression d'une partie ou de l'ensemble d'une page Web.

Il faut finalement attendre qu'il soit archivé, pour pouvoir considérer le Web comme un support d'informations capable de s'auto-préserver. Ce n'est qu'une fois les corpus d'archives rendus accessibles depuis le Web lui même (Brügger, 2009), que l'on peut considérer qu'il garde en lui la trace (mesurée et mesurable) de ses états passés.

Par ailleurs et à la différence d'autres types de documents à archiver, une page Web possède ce que J. Masanès (Masanès, 2006b, p.47) nomme une **double cardinalité**.

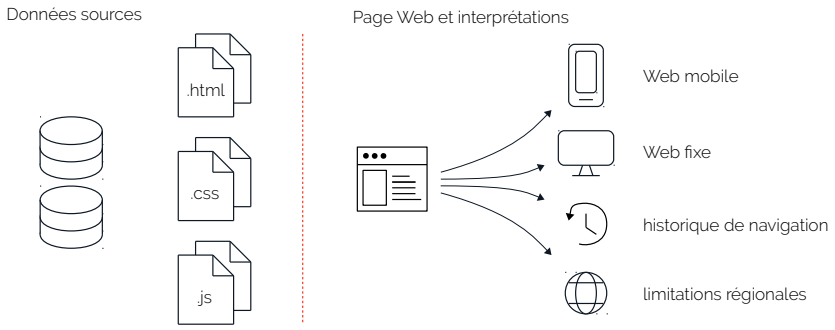


Figure 2.5: La double cardinalité d'une ressource Web, d'après (Masanès, 2006b)

La cardinalité est le nombre d'instances en circulation d'un artéfact donné : un musée conservera des pièces uniques et originales, une librairie mettra à disposition de ses visiteurs des copies. La cardinalité donne toute sa valeur à un objet archivé et influence les techniques de préservation. Jusqu'à l'invention de l'imprimerie, pour archiver un livre il fallait le copier à la main. L'original était conservé dans un lieu donné et les copies envoyées vers d'autres bibliothèques (Canfora, 1990). L'original se perdant parfois, la copie (rectifiée ou annotée) devenait à défaut œuvre de référence. Mais la notion d'original disparait avec l'imprimerie. Le livre dans sa forme est stabilisé, les bibliothèques possédant toute la même version d'un ouvrage devenu reproductible à

l'identique (Febvre et Martin, 2013). Contrairement au livre, les sites et les pages Web ont la singularité de présenter une double cardinalité :

1. les fichiers sources, hébergés sur un serveur donné
2. l'infinité d'accès possibles à cette source

D'une machine à l'autre ou d'un écran à l'autre, une page Web sera toujours vue différemment (Bon, 2014). Soit que la taille de l'écran (ordinateur, smartphone, tablette, etc.) aura modifié son aspect, soit que la qualité de la connexion à Internet n'aura pas permis de tout charger, ou encore que l'historique de navigation aura influencé l'affichage de la page à nos yeux.

C'est pour cela que J. Masanès propose de parler de **ressource Web** (Masanès, 2006b, p.48) pour nommer tout objet Web susceptible d'être archivé. Une ressource Web est un document unique dont la source peut être identifiée précisément mais interprétée d'une infinité de manière possibles. Depuis son navigateur, derrière son écran. Pour l'archiviste se pose alors la question de quoi archiver ? L'original ou toutes les interprétations d'une même page ?

Arrivé à ce point, archiver le Web revient donc à prendre en compte l'ensemble des états successifs d'une ressource Web afin de ne rien rater. Une fois la fréquence d'archivage décidée, la collecte peut être opérée du point de vue de la source ou du point de vue de l'internaute naviguant derrière son écran. Où l'archiviste choisira-il de se positionner, lui et ses outils de collecte ?

Sélection

Toute collecte sur le Web débute par le choix d'un point d'entrée clairement identifié. L'archiviste ne peut se permettre de dériver au hasard du Web pour trouver les sites qui l'intéressent. La **sélection** désigne donc l'ensemble des techniques mises en place pour définir ce ou ces points d'entrée. S'agit-il d'un site Web précis ? D'une liste de pages Web ? D'un masque ou d'un motif d'URL à satisfaire ? À quelle profondeur débuter l'archivage ? Doit-on commencer par collecter la page principale d'un site, la *front page* [33], ou un sous-ensemble de pages contenant un mot clé donné ?

Le principal critère de sélection définissant le périmètre d'un corpus à archiver reste l'extension d'URL. Les .fr, .uk et autres .ma définissent le cadre grossier d'un domaine national sur le Web (Section 2.1). La sélection s'opère alors en validant un masque d'URL ou une heuristique prédéfinie [34]. Il est également possible de dessiner les contours d'une archive en partant d'une liste initiale de sites Web, appelés sites sources et liés à une thématique précise. Il faut pour cela faire appel, en amont de toute collecte, à des experts (sociologues, historiens, etc). L'INA, par

33. Page principale d'un site Web, on peut également parler de page d'accueil.

34. Dans un script cherchant à sélectionner les sites du domaine français, on ne conservera que les noms de domaine qui valident l'expression régulière suivante : $.*\backslash.fr\$$

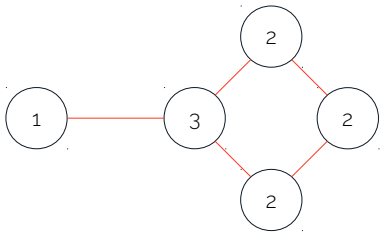


Figure 2.6: Graphe dont les nœuds sont labellisés par degré. En théorie des graphes, le degré $deg(v)$ d'un nœud v correspond au nombre de liens incidents (entrants ou sortants) à ce nœud.

exemple, a procédé par expertise pour identifier les 14 000 sites média de son périmètre d'archivage.

Lors d'une collecte plus large, les archivistes peuvent s'appuyer sur des indices traduisant la valeur d'un site ou d'une page visitée. Le degré d'un site Web (Figure 2.6) permet ainsi d'estimer son autorité au sein d'un environnement hypertexte (Abiteboul *et al.*, 2002). Une autre stratégie consiste à identifier les sites sources par rapport aux habitudes de navigation des internautes. Quels sites sont fréquemment visités ? Quelles pages en particulier ? Il s'agit alors d'exploiter les requêtes adressées à un moteur de recherche en ligne (Pandey et Olston, 2005). Ou encore, de se baser sur les motifs d'accès (AlNoamany *et al.*, 2013) afin de privilégier, au sein d'un même site, l'archivage de séquences de pages fréquemment visitées. Le système de sélection d'Internet Archive, en particulier, utilise cette dernière méthode (Kimpton et Ubois, 2006). Dans la même veine, un historique de navigation personnel pourra faire office de liste de primo-candidats à archiver (Dumais *et al.*, 2016). Cette fonctionnalité a récemment été ajoutée aux archives du Web danois [35]. Mais précisons, qu'aucune stratégie de sélection ne prévaut sur une autre. Elles sont d'ailleurs souvent combinées pour former une chaine complexe en amont de tout collectage (Section 2.3).

35. Une version bêta du système danois est disponible ici : `https://github.com/netarchivesuite/solrwayback`

Par ailleurs, il est possible d'opérer une sélection par *crowd sourcing* en faisant appel à des archivistes tiers. C'est toute l'idée du service payant *Archive-it* [36], lancé en 2006 par Internet Archive, permettant à tout un chacun de se constituer des corpus d'archives Web. Quelques 230 millions d'URL ont ainsi été recueillies entre 2006 et 2007 avant d'être reversées dans le fond d'archives principales de la Wayback Machine. Dans l'idée de démocratiser encore d'avantage leur exploitation, les archives portugaises offrent la possibilité à chaque internaute de suggérer une liste de pages ou de sites Web (pas nécessairement appartenant au domaine portugais .pt) à ajouter aux collectages [37]. L'utilisateur devient ainsi acteur de la préservation du Web.

36. `http://www.archive-it.org`

37. `http://sobre.arquivo.pt/en/collaborate/suggest/`

Terminons en soulignant que le choix d'archiver une page plutôt que l'ensemble d'un site (et inversement) n'est pas trivial. À quelle échelle doit-on archiver ? L'arbitrage est souvent décidé au cas par cas et peut faire l'objet de compromis. Même si l'on demande à Internet Archive de sauvegarder une page précise, le système remontera toujours à la front page du site afin d'en archiver la racine (Kimpton et Ubois, 2006). Ce genre de mécanisme permet d'amender et d'enrichir les points d'entrée après chaque collecte. La découverte de nouveaux sites appelant à réévaluer sans cesse la liste d'origine.

Collecte

Par **collecte** nous désignons l'ensemble des techniques visant à transformer une page du Web vivant en une page archivée. Comme nous l'indiquions précédemment le Web peut, sous certains aspects, être considéré comme capable de s'auto-préserver. Une fois archivés, l'ensemble des éléments collectés restent accessible depuis le Web. Le Web contient en lui même les traces de son passé. Lorsque l'on scanne une pellicule, image par image, pour archiver un film, on fait subir à ce support de mémoire une transformation. De l'analogique au numérique. Dans le cas d'une ressource Web, la transformation induite par la collecte est minime. Il s'agit grossièrement de venir prélever les fichiers d'origines d'une page, sans les altérer et de les dater avant de les réintroduire dans les archives Web.

Or le protocole HTTP qui régit les règles de communication sur le Web, entre client (l'internaute) et serveur (la page), n'autorise qu'un accès unitaire aux ressources Web. Il n'est possible d'accéder au Web qu'une page à la fois. La page Web (identifiée par une URL unique) est en cela l'unité de consultation de base du Web. La collecte doit donc s'effectuer page après page et non par lot, comme illustré ci-dessous :

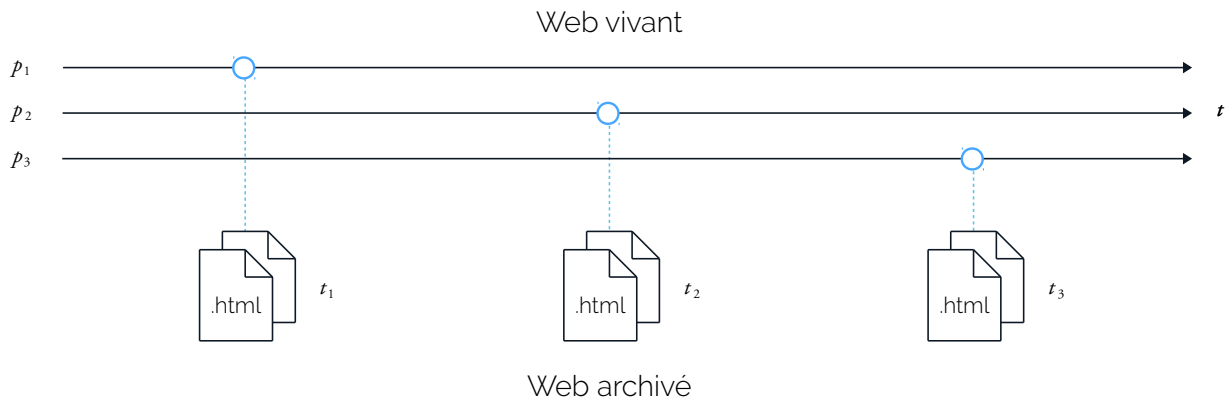


FIGURE 2.7: Archivage du Web vivant page après page, de p_1 à p_3, entre les instants t_1 et t_3

Il existe ainsi trois grandes familles de techniques d'archivage du Web, qui témoignent du déplacement progressif des outils de collecte du serveur vers le client.

La première, nommée **archivage côté serveur**, consiste à collecter les ressources directement depuis le serveur hébergeant un site ou une page Web. Cette technique est plutôt employée par les producteurs de données eux même, s'ils souhaitent archiver l'ensemble des ressources d'une de leurs plateformes Web par exemple. Mais les moyens à mettre en place sont considérables car, à la différence d'une simple

copie, l'archivage côté serveur induit la réplication du site (les fichiers HTML, CSS, etc., d'origine) et de l'entièreté de son environnement de développement et d'hébergement. On utilisera plutôt cette méthode pour archiver le Web dit profond [38].

38. À opposer à la portion visible du Web (ce que l'on voit derrière son écran), le Web profond désigne tout élément qui n'est pas accessible directement depuis un crawler : les formulaires, bases de données, etc., (Lawrence et Giles, 2000). Une discussion sur l'exploration d'archives du Web profond sera menée au chapitre 6.

La seconde approche, dite **archivage des transactions**, se situe à la frontière entre serveur et client. Il s'agit ici de positionner l'outil de collecte au niveau du système d'entrées/sorties du serveur hébergeant le site Web ciblé (Fitch, 2003). Ce que l'on archivera sera le couple [requête, réponse] du client au serveur, soit la demande d'un internaute cherchant à visiter une URL donnée et la page Web telle que retournée par le serveur. Cette forme d'archive dessine une vision non-exhaustive d'un site Web mais néanmoins fidèle à la réalité du flot d'internautes qui le parcourent. En capturant la trace des pages effectivement visitées et la manière toujours unique dont celles-ci sont affichées à l'écran des utilisateurs, cette technique est la seule qui intègre directement l'humain et ses gestes dans les archives Web [39].

39. Ce type d'archive fera l'objet d'une exploration dédiée au Chapitre 6, où nous interrogerons les logs de navigation Web de la Bibliothèque du Centre Pompidou.

La dernière famille, connue sous l'appellation d'**archivage côté client**, est aussi la plus répandue. Ayant acté qu'une ressource Web pouvait être visualisée d'une infinité de manière possible par le client (l'internaute), l'archiviste choisit ici de placer son outil de collecte en lieu et place de l'utilisateur. L'outil devient client et cherche à reproduire les interactions d'un internaute pour accéder au contenu ciblé : la page Web à archiver. Tout l'enjeu est donc de définir et de contrôler l'exhaustivité de ces interactions pour construire une copie fidèle d'une page ou d'un site.

Comme la collecte doit être menée page par page, programmée à l'avance et conduite à échelle large, les archivistes du Web se sont inspirés des crawlers développés pour les moteurs de recherche (Pant *et al.*, 2004) à la fin des années 1990. Un **crawler** est un robot programmé pour parcourir un site ou un ensemble de sites, une page à la fois, en capturant au passage l'ensemble de ses fichiers d'origine. Un crawler, pour bien fonctionner, doit respecter des règles de politesse : éviter les dénis de services (DNS, Serveurs HTTP), les listes noires officielles (par exemple robots.txt) et officieuses (*cloaking*, pièges à robot) [40]. Dans le cadre spécifique des archives du Web, un crawler doit en plus intégrer les contraintes temporelles évoquées précédemment. Il a pour mission de capter l'ensemble des changements intervenant sur une page ou un site cible. Enfin, nous appelons **crawl** une campagne d'archivage menée par un crawler et (par abus de langage) le résultat même de cette campagne.

40. Ce manuscrit n'étant pas spécifiquement centré sur la question des crawlers, il est possible d'en apprendre d'avantage en se référant aux cours de C. Maussang (`https://frama.link/FrFrZ5EC`) ou en se tournant vers des ouvrages dédiés (Chakrabarti, 2002; Mitchell, 2015).

Les caractéristiques d'un corpus d'archives Web sont directement le fait du crawler qui a mené la collecte. Un crawl peut ainsi être conduit de plusieurs manières. Une première possibilité revient à entreprendre une collecte en profondeur d'abord (*depth-first*, Figure 2.8, a). Le crawler

capturera en priorité les pages filles de la page sur laquelle il se trouve. Un autre approche consiste à travailler en largeur d'abord (*breadth-first*, Figure 2.8, b). Le crawler privilégiera cette fois les pages sœurs. Mais ces techniques sont lentes et il faudra prévoir un temps considérable pour parcourir l'entièreté d'un site, or l'archiviste cherchera au contraire à minimiser le temps de capture. Aussi, on peut envisager l'instauration d'une limite en profondeur pour ne pas archiver des pages trop éloignées de la racine du site (Figure 2.8, c).

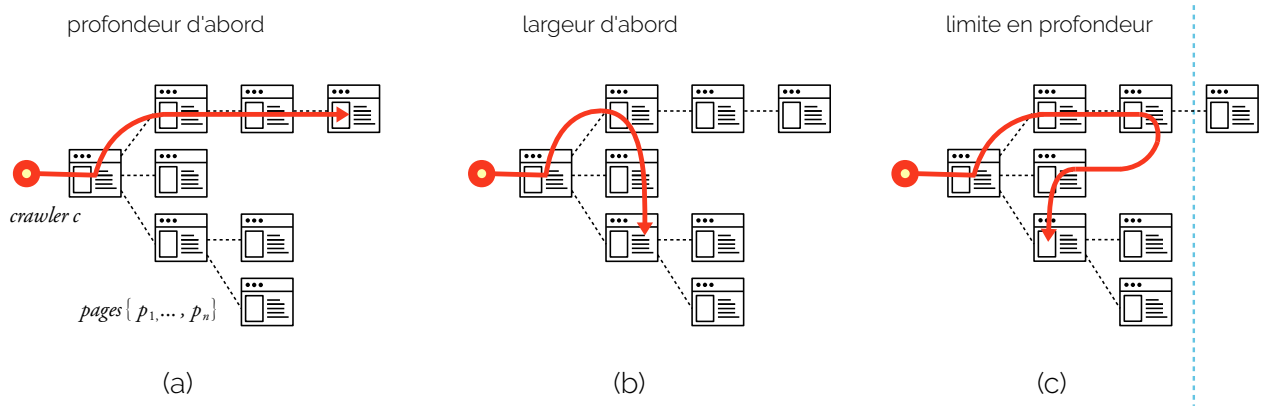


FIGURE 2.8: Différentes stratégies adoptées par un crawler c pour collecter les pages $\{p_1, ...p_n\}$ d'un même site

En pratique, l'archiviste optera plutôt pour un compromis entre largeur et profondeur. Avec la démocratisation des moteurs de recherche en ligne, l'internaute n'est plus obligé de passer par la front page d'un site pour en consulter le contenu. Les profils de navigation se diversifient rapidement (Hölscher et Strube, 2000). Pour identifier les pages pertinentes, les crawlers doivent donc intégrer à leurs programmations divers indicateurs topologiques ou sémantiques. Le pageRank (Page *et al.*, 1999) ou le degré entrant d'un site (Figure 2.9) traduisent tous deux l'importance d'une page crawlée (Cho *et al.*, 1998). Ces mesures peuvent être enrichies au regard de l'historique du crawl ou d'une possible hiérarchie entre pages (Baeza-Yates *et al.*, 2005).

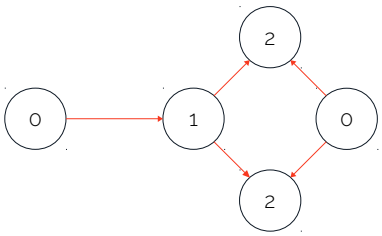


FIGURE 2.9: Graphe dont les nœuds sont étiquetés par degré entrant. En théorie des graphes, le degré $deg^-(v)$ d'un nœud v correspond au nombre de liens incidents entrant à ce nœud.

Nous le verrons lorsque dans le chapitre 3 nous changerons de point de vue, passant de l'archiviste à l'explorateur d'archives, la cohérence est une notion fondamentale. Un crawl doit garantir une forme de cohérence topographique et temporelle vis-à-vis du corpus qu'il cherche à constituer. Sur ce point, il faudra poser la question de l'ordonnancement des sites les uns par rapport aux autres. Certains sites, larges ou volatiles, feront l'objet d'une collecte rapide (*short-term scheduling*) qui mobilisera toutes les ressources du crawler. Pour les autres, le crawl s'inscrira dans le temps long (*long-term scheduling*) et pourra prendre plusieurs jours (Castillo *et al.*, 2004). Néanmoins, certaines contraintes

de politesse peuvent obliger à traiter en parallèle une grande quantité de sites. Mais si cela est possible, ne crawler que lorsqu'un site est le moins susceptible de subir des changements peut garantir une cohésion temporelle au corpus (Saad *et al.*, 2011). S'appuyant sur toutes ces réflexions, les équipes d'Internet Archive présentent en 2004 un crawler open-source, l'Heritrix (Mohr *et al.*, 2004) capable de s'adapter à divers type de collecte : large (*broad crawling*), en continu (*continuous crawling*) ou focalisée (*focused crawling*). Heritrix reste encore aujourd'hui le crawler le plus répandu pour l'archivage du Web.

Face à l'évolution du Web, les crawlers s'adaptent et archivent de nouveaux objets : allant des sites Flash [41] aux vidéos Youtube ou Dailymotion (Pop *et al.*, 2010). Mais alors que les contenus publiés incorporent de plus en plus d'éléments dynamiques, se syndiquer à un flux RSS devient une stratégie à part entière pour collecter de l'information en continu (Oita et Senellart, 2010). Des bibliothèques sont développées pour interpréter les portions de code utilisant du Javascript [42] et les crawlers commencent à se spécialiser pour archiver certains réseaux sociaux. Les interfaces de programmation applicatives (API) s'imposent comme des sources de données auxquelles il convient de se connecter. Si certaines API ouvertes permettent de crawler l'entièreté d'une plateforme [43], d'autres plus limitées obligent les archivistes à faire preuve d'inventivité. Ainsi, Internet Archive possède, depuis Mars 2016, un compte Facebook *charlie.archivist* dont la timeline est régulièrement archivée [44].

41. `https://fr.wikipedia.org/wiki/Adobe_Flash`

42. `https://github.com/ariya/phantomjs`

43. Voir la crawl de Github réalisé en 2010 par F. Cunny et Linkfluence (`http://www.visualcomplexity.com/vc/project.cfm?id=785`).

44. `https://frama.link/7fzU_DbR`

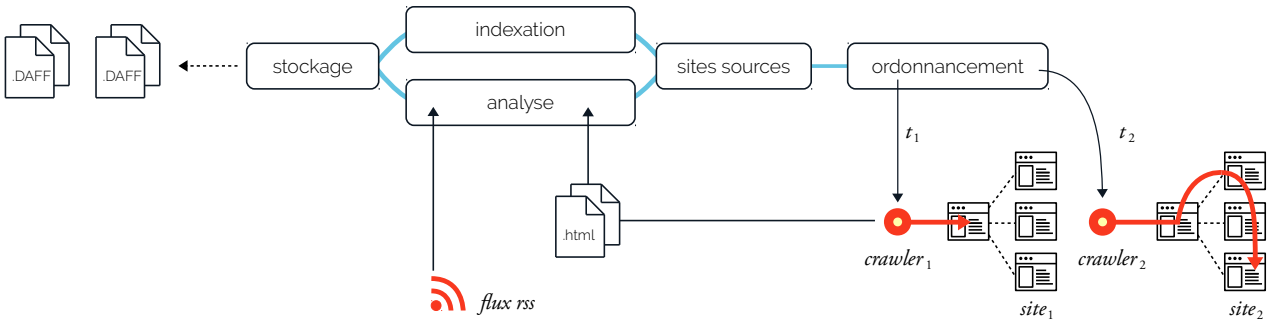


FIGURE 2.10: Fonctionnement général du système de collecte de l'INA

Terminons ce descriptif des techniques de collecte, par une présentation plus poussée du crawler de l'INA. Développé par les équipes de T. Drugeon (Drugeon, 2005), ce crawler fut en charge (de 2010 à 2014) de l'archivage des sites Web de l'Atlas e-Diasporas. Pour commencer (Figure 2.10), un ordonnanceur général (*scheduler*) gère la liste des sites sources auxquels une fréquence de collecte a été associée. En se basant sur cette fréquence, l'ordonnanceur choisit les sites à archiver en priorité

et leur dédie à chacun un crawler (*site crawler*). Plusieurs centaines de crawlers peuvent ainsi être lancés en parallèle. Ces crawlers procèdent à une récolte en largeur d'abord, mais sans sortir du périmètre du site qui leur est alloué. Une fois les pages visitées, le contenu est indexé et stocké sur fichiers. À ce niveau, si des liens hypertextes sortants sont détectés dans une page, les sites pointés par ces derniers sont conservés afin de potentiellement venir enrichir la liste des sites sources. La fréquence de collecte est mise à jour entre deux crawls successifs. Soit en analysant les informations venues de l'agrégateur de flux RSS, soit en comparant l'évolution d'une page archivée d'une version à l'autre. Cette architecture fait du crawler de l'INA un outil extrêmement réactif, adapté à la nature même des sites médias et, de fait, très efficace lorsqu'il s'agit de constituer rapidement des corpus portant sur un événement singulier [45].

45. Voir la collecte réalisée pour les attentats de Paris en 2015 (`https://asap.hypotheses.org/173`)

Stockage

Le **stockage** représente l'ensemble des techniques d'enregistrement d'une ressource Web crawlée. Ainsi, parallèlement à son crawler, l'INA développe son propre format de fichier destiné au stockage des archives Web : le Digital Archive File Format (DAFF). L'INA prend ainsi le contre-pied du reste de la communauté qui, elle, continue de s'en tenir au format Web ARChive (WARC) pour sauvegarder la grande majorité des corpus existants.

C'est en 1996, s'inspirant du format de compression et d'archivage ARC (popularisé à la fin des années 80), qu'Internet Archive définit le ARC_IA [46]. L'idée étant de combiner plusieurs ressources collectées en un seul et même fichier avant de les compresser pour en réduire la taille sur disque.

46. `https://www.loc.gov/preservation/digital/formats/fdd/fdd000235.shtml`

Mais l'ARC_IA évolue rapidement, suivant les avancées des techniques de crawl, et ce, jusqu'à atteindre sa version actuelle : le WARC. Devenu le format standard d'archivage Web en 2009 [47], un fichier d'archives WARC peut être vu comme la concaténation de plusieurs enregistrements (ou blocs), chaque enregistrement correspondant à une ressource Web crawlée [48]. Les informations contenues dans un bloc WARC sont de deux natures : des méta-données et des données. Les **méta-données** (stockées dans l'en-tête du bloc) couvrent toutes les informations relatives au crawl : date de collecte, taille de la ressource, URL de la ressource, ID du bloc... Ces méta-données sont directement suivies des données à proprement parler : soit l'enregistrement brut des fichiers .HTML, .CSS, etc., collectés. Ainsi, chaque fois qu'une page Web est archivée (qu'elle ait évolué ou non depuis le précédent crawl) un bloc est ajouté au fichier WARC courant (Figure 2.11 (a)).

47. `https://www.iso.org/standard/44717.html`

48. Une page, une image... Chez Internet Archive, tout objet associé à une URL unique sera archivé comme ressource Web.

Directement liés au WARC, il est possible d'extraire d'un bloc deux

sous-formats spécialement dédiés à l'exploitation des corpus : les WAT et WET. Un fichier WAT (Web Archive Transformation) ne contient que des méta-données. Contrairement au WET (Web Extracted Text) et à ses dérivés (LGA ou WANE [49]) qui, eux, ne stockent que des éléments de texte issus de la partie données d'un bloc WARC. Pourtant utilisé par la quasi-totalité des initiatives d'archivage, le format WARC introduit un point hautement critiquable : l'existante de redondances dans les fichiers d'archives Web.

49. https://webarchive.jira.com/wiki/spaces/ARS/pages/90997507/Datasets+Available

En effet, entre deux crawls successifs, un bloc WARC sera invariablement crée (que la ressource Web collectée ait évolué ou non). Une page Web stable dans le temps, verra ainsi son contenu archivé autant de fois qu'elle aura été crawlée, conduisant à une consommation d'espace de stockage considérable. C'est donc en partant de l'intuition selon laquelle méta-données et données devraient être stockées séparément (pour ne pas surcharger les corpus) que l'INA a développé le format DAFF.

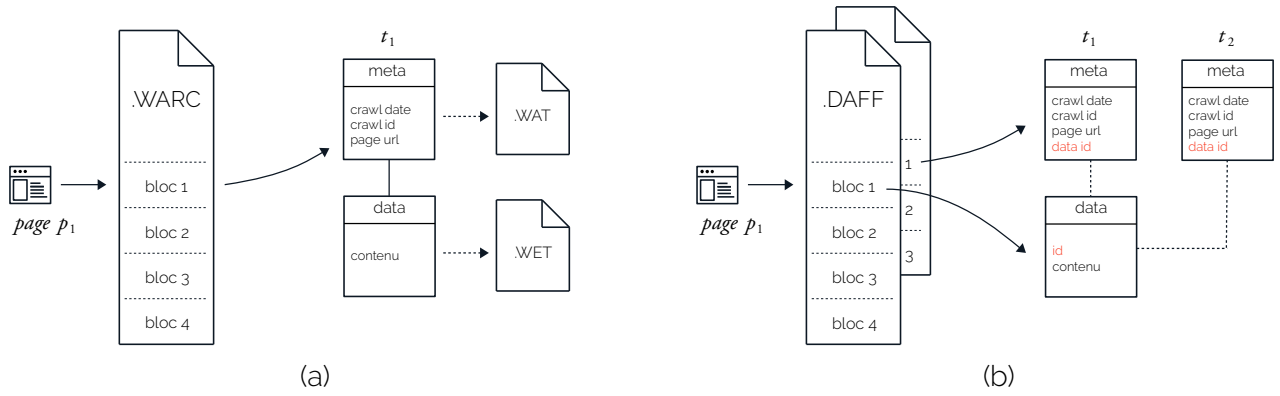


Figure 2.11: Différences entre les formats WARC (a) et DAFF (b)

Une archive DAFF est en réalité l'association de deux fichiers complémentaires : un fichier de méta-données et un fichier de données. Comme pour le WARC, chaque fichier est une suite de blocs correspondant à une ressource Web crawlée. Le fichier de méta-données contient divers champs relatifs à la collecte (voir Table 2.1). Le fichier de données, quant à lui, ne renferme que deux champs : un identifiant unique et le contenu (HTML, CSS, etc.) de la page Web archivée. Avec le DAFF, données et méta-données sont stockées séparément. Ainsi, d'un crawl à l'autre, si la page Web visitée n'a pas évolué, alors son contenu ne sera pas re-téléchargé : seul un nouveau bloc de méta-données sera ajouté pour témoigner du passage du crawler. Chaque bloc de données est donc associé à un ou plusieurs blocs de méta-données (le champ *content* des méta-données correspondant au champ *id* des données). Ce mécanisme

permet, d'une part de ne pas dupliquer inutilement le contenu d'une page Web archivée et, d'autre part, de pouvoir pratiquer rapidement divers calculs statistiques sur le seul fichier de méta-données. Celui-ci étant par nature plus léger qu'un WARC complet, donc moins long à traiter.

Notons que pour tester si une page a évolué depuis sa précédente visite, le crawler de l'INA compare la valeur des champs *id* des blocs de données concernés. En effet l'*id* est une clé SHA-256 résultant du hachage [50] du contenu même de la page Web archivée. On dira donc d'une page Web qu'elle s'est transformée si et seulement si les deux clés successives sont différentes. La nature de ce changement ne sera en revanche pas connue, celui-ci pouvant aller de la refonte entière de la page à la simple suppression d'une virgule.

50. En cryptographie, le hachage consiste à transformer une donnée de taille arbitraire, une image (ou clé) de taille fixe et unique.

Méta données	Champ	Description
obligatoire	id	identifiant unique du bloc
	url	url associée à la ressource
	date	date de téléchargement (ISO 8601)
	content	identifiant unique du bloc de données associé
	status	statut de retour du crawler (ok, request_error, server_error, etc.)
facultatif	crawl_session	identifiant unique de la campagne de crawl
	charset	encodage de la ressource
	type	Type MIME (identifiant du format de donnée de la ressource)
	corpus	nom du corpus d'archives associé
	ip	adresse ip associée au crawl
	level	profondeur du crawl
	page	la ressource est elle une page Web (0 | 1)
	client_country	pays associée à la page
	length	taille du bloc de données associé
	active	la ressource était elle active au moment du crawl
	client_lang	langue associée à la ressource
	referer_url	url précédemment visitée par le crawler
Données	**Champ**	**Description**
	id	clé SHA-256 unique
	content	contenu (HTML, CSS, etc.) de la ressource

TABLE 2.1: Ensemble des champs disponibles dans les fichiers de méta-données et de données DAFF

Outre le stockage sous formats WARC et DAFF, J. Masanès (Masanès, 2006b, p.64) rappelle qu'il existe des méthodes alternatives de sauvegarde des archives. Associées aux stratégies de collecte situées côté serveur on trouvera les formes dites d**archivage sur fichiers servis en local** qui consistent à transformer un site Web archivé en une copie locale de l'ensemble de ses ressources. Ainsi les URIs absolues,

permettant (sur le Web) de naviguer d'une page à l'autre, seront transformées en URIs relatives à l'intérieur du fac-similé. Très coûteuse, cette méthode nécessite de transformer en profondeur la nature des pages archivées et devient vite ingérable à mesure qu'augmente le nombre de collectes.

Enfin, il reste toujours possible de copier un site, page après page, sous format PDF ou image (capture d'écran ou vidéo). Bien que facile à mettre en place (techniquement parlant) cette stratégie ne passera pas non plus à l'échelle [51] et aura pour conséquence d'arracher les sites et pages Web archivés à leur environnement hypertexte d'origine.

51. En 2013, K. Goldsmith imprime littéralement plusieurs centaines de milliers de pages Web en soutient à A. Schwartz, remplissant l'équivalent d'une pièce de 1,100 m^2

Pour terminer, les corpus d'archives Web répartis dans le monde se comptent par centaines. Alors que le Web vivant continue son expansion, le volume du Web archivé ne cesse de croitre. En 2017, la BNF avait archivé 18 milliards de pages Web (soit environ 370 TB) tandis que l'INA plafonnait à 43 milliards de pages pour un total avoisinant les 420 TB. Et depuis sa création, l'Internet Archive a collecté à elle seule pas moins de 650 milliards de ressources Web (pages, images, vidéos...) soit 40 PB de données. Ainsi, face à ces corpus qui s'amassent et à la nécessité de les exploiter, les archivistes du Web ont dû développer des outils dédiés à leur exploration.

Consultation

Par **consultation**, nous désignons les stratégies d'interrogation et les systèmes d'exploration des archives Web. Ainsi, pour permettre aux chercheurs d'analyser le résultat des collectes, les archivistes déploient des dispositifs techniques *au-dessus* des corpus existants.

Comme nous l'évoquions en Section 2.1, la fouille est tributaire des modalités d'accès aux données qui, pour 50% des initiatives (Costa *et al.*, 2013), passent par la mise en place d'un portail en ligne. Mais cela ne signifie pas pour autant que les archives sont entièrement accessibles. Su ce point, 38% des initiatives restreignent la consultation de leurs corpus : soit l'analyse doit se faire localement (INA, BNF, etc.), soit les archives ne sont pas intégralement mises à disposition du public (The Library of Congress, Australia's Web Archive, etc.). Contrairement à Internet Archive et aux Portuguese Web Archives qui proposent un plein accès, en ligne, à leurs collectages.

Les dispositifs de consultation, déployés par dessus les archives, reprennent l'architecture générale de la plupart des système de moteurs de recherche [52] (Grainger *et al.*, 2014; McCandless *et al.*, 2010). Les archives sont ainsi indexées puis mises à disposition d'un serveur de *search* qui les rend interrogeables [53]. L'**indexation** définie l'étape de transformation d'un document texte en une liste de mots ou d'ensembles de mots, cette étape est nécessaire à toute construction d'un

52. L'anglicisme *search* désigne tout autant les systèmes de moteurs de recherche eux même, que les stratégies déployées par ces derniers.

53. Nous développerons le processus d'indexation des archives Web plus en détail dans la Section 3.3

moteur de recherche. On appelle **index** la structure de données obtenue après indexation.

Côté utilisateur, la recherche se traduit par une interface Web, dans laquelle il est possible de rentrer une requête (un texte, un ensemble de mots clé, des filtres, etc.), puis de consulter les résultats correspondants sous la forme d'une liste ou d'un histogramme (Figure 2.12). Tout l'enjeu pour ces moteurs d'exploration d'archives est de proposer la meilleure technique de recherche possible pour fouiller efficacement un corpus d'archives Web. Ou comment, partant de la requête d'un chercheur, proposer avec justesse un ensemble de page archivées qui satisfasse ses interrogations ?

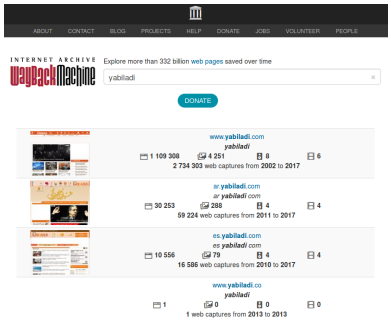


FIGURE 2.12: Interface de search de la WayBack Machine (`https://web.archive.org/web/*/yabiladi`)

Dans ce domaine, la recherche dite **plein texte** (*full-text*) est ce vers quoi tendent toutes les initiatives d'archive du Web (Costa et Silva, 2011, 2012). En recherche d'information, le full-text revient à faire correspondre les mots d'un document ou d'un ensemble de documents avec les critères fournis par un utilisateur (des mots clés, une phrase, etc.). Popularisée sur le Web par AltaVista, c'est la recherche telle que nous l'expérimentons quotidiennement en interrogeant Google. Mais, encore aujourd'hui, son application aux moteurs d'exploration d'archives Web reste limitée. Si l'INA, les portugais d'Arquivo.pt ou encore The Web Archive of Catalonia [54] proposent d'étendre cette fonctionnalité à l'ensemble de leurs corpus (Stack, 2006), à la BNF, au contraire, le full-text n'est applicable que sur les seules URL des pages archivées. c'est-à-dire qu'à une requête utilisateur donnée, le moteur de la BNF ne pourra faire correspondre qu'une recherche stricte par URL, sans regard pour le contenu même des pages. En 2016, Internet Archive ajoute à la Wayback Machine un système full-text basé sur le titre des pages Web collectées [55], avant cette date seule une recherche stricte par URL était proposée. Selon M. Costa (Costa *et al.*, 2013), 67% des initiatives d'archivage du Web proposaient en 2013 une recherche full-text complète, limitée ou dégradée.

54. `https://www.padicat.cat/en`

55. Si et seulement si le titre est présent dans la balise *title* du HTML de la page (`https://frama.link/K4CMmzHU`, `https://www.w3schools.com/html/html_head.asp`).

La taille importante des corpus ou la difficulté des archivistes à définir quels doivent être les éléments d'une page Web à indexer peuvent expliquer que le full-text soit si compliqué à mettre en place. Mais des alternatives existent : on peut ainsi envisager une recherche par catégories (Holzmann et Anand, 2016), par entités nommées (Spaniol et Weikum, 2012) ou en se basant sur des tendances issues des réseaux sociaux (Risse *et al.*, 2014). Une solution originale consiste à ne pas construire de moteur d'exploration mais à s'adosser à des systèmes existants. Le système Memento [56], par exemple, permet d'interroger les corpus d'Internet Archive directement depuis un navigateur Web, à la manière d'une extension.

56. `http://mementoweb.org/about/`

Créée en 2001, la Wayback Machine est le plus emblématique des systèmes de consultation d'archives Web (Tofel, 2007). L'ensemble des

ressources archivées y sont indexées page par page (bloc WARC par bloc WARC). Les index sont ensuite répartis sur les quelques 2500 serveurs de stockage du centre de stockage (*data center*) principal d'Internet Archive. Un ordonnanceur central envoie les requêtes des utilisateurs à l'ensemble des serveurs avant d'agréger les résultats. Il existe un système de cache destiné à améliorer les temps de réponse de la Wayback Machine, ainsi l'attente variera en fonction de la popularité de chaque requête. Le moteur des Portuguese Web Archives propose, lui, d'indexer les archives d'abord par date de téléchargement puis page par page (Costa *et al.*, 2013). Si l'utilisateur associe à sa requête une année ou un intervalle de temps précis, le résultat de ses recherches lui sera plus rapidement retourné. Les moteurs d'exploration peuvent aussi être entièrement décentralisés. A. Anand décrit Everlast (Anand *et al.*, 2009) comme un système de consultation pair-à-pair (*peer-to-peer*) où chaque élément du réseau est à la fois serveur et client, plus scalable donc.

Bénéficiant d'une distribution open source, la Wayback Machine est aujourd'hui réutilisée ou sert de base [57] à l'architecture de 62% des initiatives d'archivage du Web (Costa *et al.*, 2013). Mais force est de constater que les systèmes existants n'encouragent pas particulièrement à l'exploration des archives, à la découverte des corpus, ou plus simplement à leur exploitation à grande échelle (en terme de quantité de pages ou de temporalité). Il est de plus difficile de s'évader du cadre strict imposé par des interfaces invariablement semblables, proposant une expérience des archives Web toujours identique.

57. Dans son intégralité ou élément par élément : Heritrix pour la partie crawl, NutchWAX pour la partie recherche full-text.

La Wayback Machine est redoutablement efficace lorsqu'il s'agit de rechercher une version précise d'une page déjà connue. Notons aussi qu'elle propose une API [58] pour accélérer les traitements. Mais sans la possibilité de filtrer à priori le contenu des pages, tout le dispositif de consultation sera à développer du côté de l'explorateur qui bien souvent n'a pas les compétences ou les moyens techniques pour y arriver. Et si l'INA offre de meilleures fonctionnalités de recherche (plein text complet, n-grammes, etc.), le fait que l'on ne puisse accéder aux corpus que depuis des lieux dédiés reste un frein majeur à toute analyse. Ainsi, une asymétrie se dessine rapidement lorsque l'on parcourt la littérature basée sur les archives Web. Beaucoup de travaux portent sur la constitution en amont de corpus d'archives (sélection, collecte, etc.), très peu en revanche se lancent dans l'exploitation ou l'interrogation de corpus existants. Ces derniers, bien que précieux à juste titre, se *limitent* [59] soit à l'analyse de versions passées de sites identifiés à priori (Schafer et Thierry, 2016; Gebeil, 2016a), soit à l'extraction d'éléments singuliers d'un contenu archivé : des images (Ben-David *et al.*, 2018) ou des liens hypertextes (Weltevrede et Helmond, 2012).

58. `https://archive.org/help/wayback_api.php`

59. Pas forcément en conscience, mais nous pensons que les outils d'exploration jouent un rôle quand il s'agit de définir de la portée de ces travaux.

2.3 Les archives Web de l'Atlas e-Diasporas

Parallèlement au travail de cartographie présenté en Section 1.3, les chercheurs pilotant la construction de l'Atlas e-Diasporas prennent la décision d'archiver l'ensemble des sites Web déjà répertoriés. Tout autant pour les préserver des assauts du temps (Khouzaimi, 2015) que pour permettre la tenue de recherches futures : se donner la possibilité d'un retour arrière, analyser les évolutions et transformations subies par ces réseaux. Déjà associée à la collecte des sites, l'INA se voit confier la charge de l'archivage. Toutes les e-Diasporas seront concernées par cette campagne de sauvegarde, mais nous nous attarderons ici sur la seule description de la section marocaine de l'Atlas.

L'archivage du corpus marocain débute en Mars 2010 et se termine en Septembre 2014 après une collecte patiente et continue. Le collectage couvre l'ensemble des 156 sites de l'e-Diasporas marocaine. La fréquence de collecte associée à chaque site est définie en amont par les chercheurs. Celle-ci est au final soit hebdomadaire (pour 56% des sites), soit mensuelle (pour les 44% restants). La majorité des sites archivés à la semaine sont les plus fréquemment mis à jour : des blogs, des portails communautaires ou des médias. Les archives sont stockées suivant le format DAFF, vu comme l'union d'un fichier de méta-données (*metadata-r-00006.daff* : 13 GB) et d'un fichier de données (*data-r-00006.daff* : 151 GB). Ces fichiers représentent un total de 17 043 833 ressources collectées, parmi lesquelles nous comptons 16 897 787 pages Web (99%), 145 301 images, 700 vidéos et 44 enregistrements audio. Dans le Chapitre 5 nous explorerons les sites *yabiladi.com* et *larbi.org* dont une présentation détaillée est donnée dans la Table 2.2. Cette table introduit un premier élément de comparaison entre les archives e-Diasporas et leurs équivalents chez Internet Archive.

Si la fréquence d'archivage (telle que mise en place par l'INA) semble plus élevée du côté d'e-Diasporas, la durée de collecte est importante chez Internet Archive. L'idée ici n'est pas de prouver qu'un corpus est mieux qu'un autre, mais de saisir les particularités de chacun. Un corpus d'archives Web n'est jamais parfait, bien au contraire, et nous émettons ici l'hypothèse que c'est par une approche mixte, en conjuguant diverses sources de données que nous maximiserons la précision scientifique des explorations à venir. Ainsi, sur la période 2010-2014 et dans les cas précis de *yabiladi.com* et *larbi.org*, la capture réalisée pour e-Diasporas semble plus fidèle. Il faudra en revanche l'associer à Internet Archive lorsque nous chercherons à remonter au delà de 2010.

Table 2.2: Décompte des archives Web des sites *yabiladi.com* et *larbi.org*

	larbi.org	yabiladi.com
Nbr de pages archivées (e-Diasporas)	78 311	2 683 928
Nbr de pages archivées (Internet Archive)	24 537	887 981
Début de l'archivage (e-Diasporas)	Mars 2010	Mars 2010
Début de l'archivage (Internet Archive)	Oct. 2002	Fév. 2001
Fin de l'archivage (e-Diasporas)	Sept. 2014	Sept. 2014
Fin de l'archivage (Internet Archive)	Sept. 2018	Sept. 2018

Mais essayons maintenant d'étendre cette comparaison à l'ensemble des sites du corpus marocain. Voyons comment ces sites ont été archivés par différentes initiatives. Comme beaucoup de ces observations devront être faites à la main (notamment à la BNF), nous nous limitons tout d'abord aux seules front pages (pages racines) de chaque site Web de l'e-Diasporas marocaine. Pour chacune de ces pages, nous consultons successivement les archives e-Diasporas (produites par l'INA), les archives de la BNF et les archives d'Internet Archive, puis nous notons leurs dates de premier et dernier crawl afin de se donner une idée de l'étendue des collectes. Les résultats sont présentés et agrégés par les Figures 2.13 (pour l'INA), 2.14 (pour la BNF) et 2.15 (pour Internet Archive). Le nom de domaine des sites est inscrit en ordonnée, le temps en abscisse. Chaque ligne est divisée en années puis en mois (1 mois = un tiret). Si un tiret est colorié c'est qu'il se trouve entre les dates de premier et de dernier crawl du site correspondant.

L'intuition précédente est confirmée par ces trois figures : les corpus vus depuis Internet Archive et (dans une moindre mesure) depuis la BNF couvrent naturellement une plus grande étendue temporelle que la collecte de l'INA, limitée aux seules années 2010-2014. En revanche, leurs collectages sont incomplets, les sites marocains ne sont pas tous archivés. C'est assez naturel au regard du périmètre d'archivage de la BNF notamment, qui ne doit théoriquement couvrir que les sites du domaine français, ici la BNF aura archivé par effet de bord des sites marocains en .com ou .org ce qui nous amène à relativiser la notion de domaine Web national telle que présentée plus tôt (Section 2.1). Ce qui frappe, enfin, est la cohérence générale de notre corpus tel qu'il est présenté par l'INA. L'ensemble des sites sont archivés, collectés au moins une fois entre 2010 et 2014 et forment un ensemble thématiquement homogène.

Mais ces figures ne doivent pas non plus nous induire en erreur, nous ne voyons pas le détail des collectes : ni ce qu'il s'est passé entre les dates de premier crawl et de dernier crawl, ni le comportement du crawler vis-à-vis de pages éloignées de la racine des sites. Or, nous

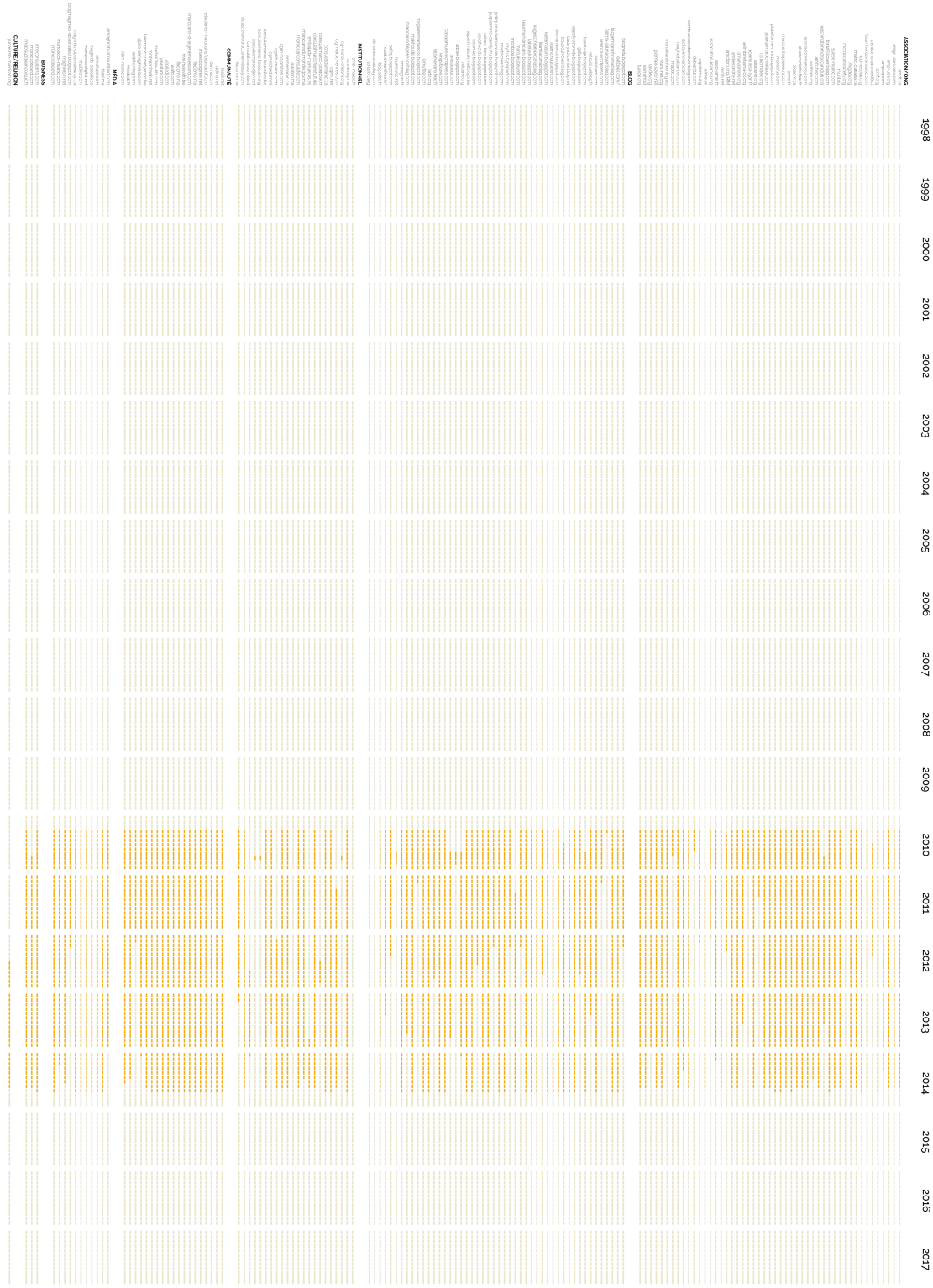


FIGURE 2.13: Préservation dans le temps des sites de l'e-Diaspora marocaine par l'INA (tirets jaunes)

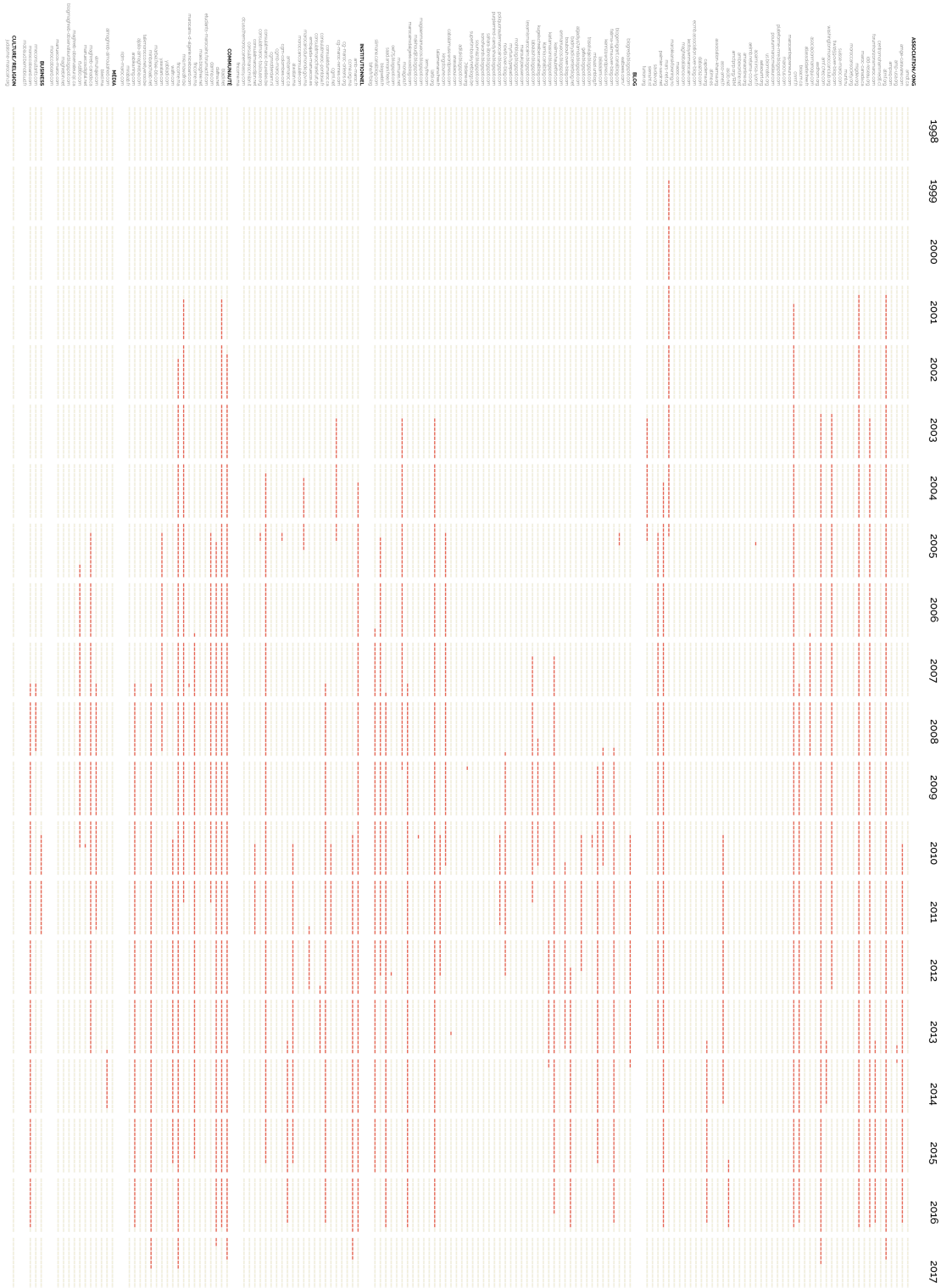


Figure 2.14: Préservation dans le temps des sites de l'e-Diaspora marocaine par la BNF (tirets rouges)

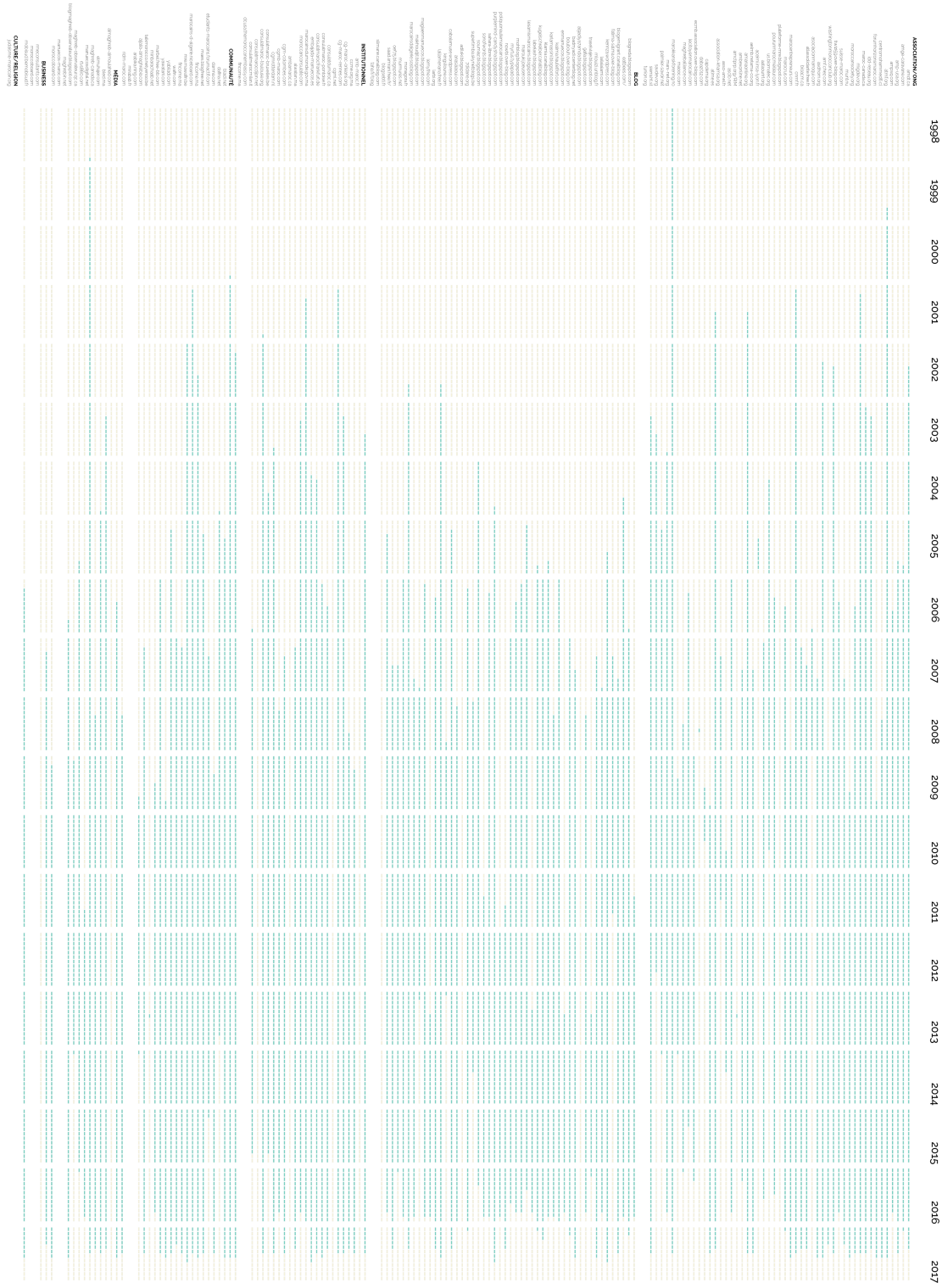


FIGURE 2.15: Préservation dans le temps des sites de l'e-Diaspora marocaine par Internet Archive (tirets bleus)

le découvrirons dans le chapitre suivant, archiver est avant tout une question de choix et de sélections, ce qui posera nombre de problèmes à l'explorateur d'archives Web.

***

Les archives Web ont été construites pour inscrire la mémoire du Web sur un support durable et préserver notre héritage numérique. Offrir ainsi la possibilité aux chercheurs de demain d'interroger, de questionner et de critiquer le Web qui nous est contemporain. Mais plus on archive et plus la taille de ce Web passé grandit, laissant parfois les chercheurs seuls face à des corpus trop larges et trop vastes pour être explorés sans méthode ni stratégie clairement définies. Les archives Web doivent rester une matière vivante. Prenons garde à ce qu'elles ne deviennent pas des capsules temporelles [60] que l'on enterre dans l'espoir, qu'un jour, peut être, quelqu'un se décide à les rouvrir.

60. https://en.wikipedia.org/wiki/Westinghouse_Time_Capsules

Il existe, selon nous, un espace et un intérêt scientifique indéniable pour des recherches portants sur de larges corpus d'archives Web. Mais cet espace appelle la création de méthodes d'exploration capables de changer d'échelle analytique au cas par cas. Aussi, nous faisons l'hypothèse qu'une stratégie quantitative ne pourra se concevoir sans une approche qualitative complémentaire (Chapitre 5) et que l'automatisation des traitements ne pourra se faire sans une part de travail manuel (Chapitre 4).

Au cours de ce chapitre, nous nous sommes attachés à décrire la genèse de l'archivage du Web comme technique de préservation d'un nouvel héritage numérique. L'idée étant de comprendre la nature des corpus que nous manipulerons dans la suite de cette thèse. Au tournant des années 2000, de nombreuses initiatives privées et publiques se sont emparées du sujet, déployant en un temps record (à l'échelle du Web) des moyens humains et techniques. Néanmoins cette dynamique semble aujourd'hui s'essouffler et force est de constater que, même si les corpus grandissent toujours plus, peu de chercheurs se sont déjà aventurés dans les archives. Le Web passé reste un terrain en partie inexploré.

Pour se saisir du Web, il aura fallu, aux pionniers de l'archivage, inventer et déployer de nouvelles méthodes de collecte et de stockage. Ces choix techniques façonnent et régissent les archives Web telles que nous les découvrons aujourd'hui. Détachée du Web vivant, l'archive Web se consulte dans des lieux sanctuarisés, souvent à la main et à travers des outils qui, malgré eux, réduisent les archives Web à de simples documents. La sensation du Web comme environnement n'est pas restituée dans les archives. L'exploration y est forcément ciblée, réduite à une URL ou un mot-clé.

Cependant, comme pour le Web vivant, l'unité d'exploration des

archives reste la page Web. WARC et DAFF sont deux formats construits au-dessus des pages qu'ils capturent et qu'ils dotent, par là même, d'une nouvelle temporalité. Une fois archivée chaque version d'une même page se voit associée une date de téléchargement. Cette date devient dès lors le seul marqueur temporel par lequel nous pouvons explorer les archives Web. Dans le chapitre suivant, nous exposerons les divers implications et biais d'analyse que peut causer cette datation.

Chapitre 3

Traces discrétisées et temporalité figée

Au cours de ce chapitre, nous amorcerons un changement de point de vue, glissant du regard de l'archiviste vers celui de l'explorateur d'archives Web.

Un **explorateur d'archive** est une personne ayant l'intention de découvrir ou d'étudier un corpus d'archives Web donné (fini ou toujours en construction). Son geste pourra tout autant être motivé par une question de recherche précise que par sa seule curiosité. Ce faisant, l'explorateur devra démêler les traces d'un Web passé pour faire émerger une information ou un savoir de cette masse de données.

Dans un premier temps, nous déconstruirons la structure des archives Web pour en saisir les règles et la grammaire interne. En effet, selon N. Brügger (Brügger, 2009), le Web archivé n'est déjà plus le Web. C'est un autre espace, un autre environnement. Lorsqu'un site du Web vivant est sélectionné, stocké et collecté, il subit une série de transformations qui forcent les archivistes à recréer, en totalité ou en partie, le système d'information de la toile. Pour explorer les archives, il faut se détacher des automatismes acquis en parcourant le Web vivant. Ce faisant, nous présenterons ici certaines propriétés et certains biais inhérents au Web archivé. Biais qu'il nous faudra prendre en considération avant toute analyse.

Pour l'explorateur, les archives Web se présentent d'abord comme des traces discrétisées du Web vivant, arrachées à un flux d'information en continu. En effet, la discrétisation du Web par les archives est le fruit d'une sélection, mais surtout d'un ensemble de destructions, comme le souligne J. Derrida (Derrida, 2014, p.60) : archiver c'est avant tout détruire ce que l'on ne peut conserver.

Par ailleurs, la collecte propulse les ressources archivées dans une nouvelle temporalité. Les pages du Web passé n'appartiennent plus au temps du Web vivant mais au temps des archives : une temporalité faite d'instantanés figés et sans possibilité d'extension. Il n'y a pas de continuité absolue entre deux versions d'une même page archivée. D'un crawl à l'autre tout peut changer (Section 2.2). Ainsi, il nous

1. *crawl blindness* en anglais

faudra discuter des phénomènes de leurres et de cécité des collectes [1], de la notion de cohérence entre pages et de la présences de contenus sur-archivés qui peuvent être source de nombreux biais d'analyse.

Passées ces mises en gardes, nous décrirons le développement de notre propre moteur d'exploration d'archives Web. Un moteur adapté au format DAFF et suffisamment flexible pour être le support de nos expérimentations à venir. Nous détaillerons notre chaine d'extraction et d'enrichissement des archives, ainsi que la pièce maitresse de tout moteur de recherche : le schéma d'indexation et ses implications.

Enfin, nous constaterons que les archives Web ne sont pas des traces directes du Web vivant, mais plutôt les traces directes des crawlers. Nous donnerons ainsi des exemples d'artéfacts de crawl, présents dans les archives de l'Atlas e-Diasporas et qui, à nos yeux, sont des freins majeurs à toute exploration large des corpus. Ce sera l'occasion de porter un regard critique sur les archives telles que nous les connaissons et d'ouvrir la voie à une exploration fragmentée du Web passé.

3.1 Détruire pour mieux archiver

J.L. Borges ouvre la seconde partie de son recueil de nouvelles *Fictions* (Borges, 1974) par un court texte intitulé *Funes ou la mémoire*. Il y fait le compte rendu concis de la rencontre entre son narrateur et le mystérieux Irénée Funes, personnage ayant la capacité de ne rien oublier, jamais. Funes a une mémoire prodigieuse :

« *En effet, non seulement Funes se rappelait chaque feuille de chaque arbre de chaque bois, mais chacune des fois qu'il l'avait vue ou imaginée. Il décida de réduire chacune de ses journées passées à quelque soixante-dix mille souvenirs, qu'il définirait ensuite par des chiffres. Il en fut dissuadé par deux considérations : la conscience que la besogne était interminable, la conscience qu'elle était inutile. Il pensa qu'à l'heure de sa mort il n'aurait pas fini de classer tous ses souvenirs d'enfance.* » — (Borges, 1974, p.116-117)

L'esprit de Funes est engorgé de souvenirs d'une infinie précision, enregistrés en continus. Mais la mémoire pour fonctionner, nous dit Borges, a besoin d'oublier, de sélectionner et de généraliser. C'est en substance la thèse soutenue par J. Derrida qui décrit le geste de l'archiviste comme un geste de pouvoir : le pouvoir de choisir ce qui doit être préservé ou non. L'archivage est le résultat d'une sélection féroce qui doit détruire avant de sauver : « *Il n'y a pas d'archives sans destruction, on choisit, on ne peut pas tout garder.* » (Derrida, 2014, p.60). C'est ainsi que l'organisation légitime de l'héritage collectif revient aux seuls archivistes qui définissent au présent la mémoire de demain, en classifiant et hiérarchisant dans les bibliothèques les traces de nos expériences passées. Ce faisant, pour Derrida « *l'archive commence là où la trace s'organise, se sélectionne* » (Derrida, 2014, p.61), car toute expérience finit tôt ou tard par s'effacer,

il en va de sa nature même. Ainsi, pour maintenir le lien qui nous renvoie à ce qui n'est plus là, il faut archiver nos traces avant qu'elles ne disparaissent.

Le Web vivant est tout autant un flux continu d'information qu'un territoire en perpétuelle expansion (Section 1.1). Pour en archiver les traces, il faut procéder par **discrétisation**, c'est-à-dire : diviser une forme continue en une ou plusieurs valeurs individuelles. Les systèmes de stockage WARC et DAFF (Section 2.2) réduisent le Web en un ensemble discret de pages archivées. Or, depuis le lancement d'AltaVista en 1995 [2], la page Web est considérée comme valeur élémentaire d'indexation, de fouille et d'exploration de la toile. Il en va de même pour les archives Web pour qui la page demeure l'unité de base de toute collecte. Ainsi, dans la suite de ce manuscrit, nous schématiserons une campagne de crawl comme telle :

2. AltaVista fut le plus important moteur de recherche pré-Google, capable d'indexer une grande partie des pages du Web et de les rendre accessibles via des requêtes plein-texte (`https://en.wikipedia.org/wiki/AltaVista`).

Un site Web archivé consiste en n pages Web numérotées $\{p_1,...,p_n\}$. Un corpus d'archives Web est le résultat d'un ou plusieurs crawls successifs $\{c_1,...,c_l\}$. Nous appelons crawl c_i le processus de collecte des pages Web $\{p_1,...,p_n\}$ d'un site Web donné. Le temps nécessaire au téléchargement des pages est supposé négligeable. Nous appelons $t_i(p_j)$ la date de téléchargement de la page p_j au cours du crawl c_i. La première date de téléchargement d'une page p_j est, enfin, notée $\min_i t_i(p_j)$. La Figure 3.1 illustre cette mécanique pour les pages p_1, p_2, p_3, l'ordre de crawl entre deux collectes et la fréquence des visites pouvant, par ailleurs, varier.

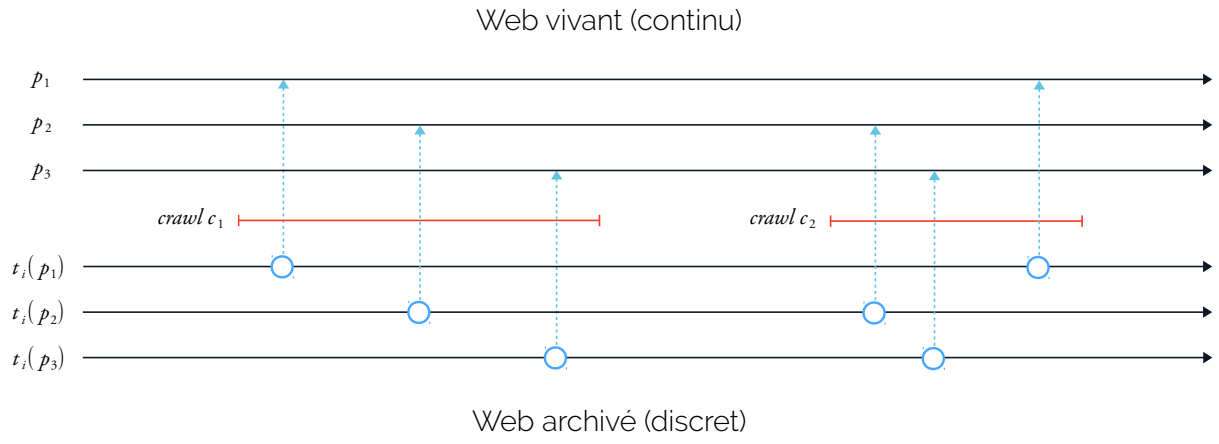


FIGURE 3.1: Archivage (cercles bleus) des pages p_1, p_2, p_3 au cours de deux crawls c_1, c_2

3.2 Un temps sans durée

Au cours d'une collecte, les pages archivées sont propulsées dans une nouvelle temporalité. Elles n'appartiennent plus au temps présent du Web vivant, mais au temps figé des archives passées. Comment dès lors capturer le temps présent ? Cette question nous ramène à Saint Augustin, dont l'expression du présent à partir de l'instant influence encore aujourd'hui la pensée occidentale. Pour saint Augustin, le présent est une suite infinie de points élémentaires, des instantanés sans étendue :

« *[...] Et cette même heure se compose elle-même de parcelles fugitives. Tout ce qui s'en détache, s'envole dans le passé ; ce qui en reste est avenir. Que si l'on conçoit un point dans le temps sans division possible de moment, c'est ce point-là seul qu'on peut nommer présent. Et ce point vole, rapide, de l'avenir au passé, durée sans étendue ; car s'il est étendu, il se divise en passé et avenir. Ainsi, le présent est sans étendue.* » — (Augustin, 421, livre XI, chap. XV, 20)

Le présent se déploie sous nos yeux comme un temps insaisissable qui, à peine éprouvé, cesse déjà d'exister pour se diluer dans le passé. La seule manière de le capturer reste donc de le diviser et de le réduire à ses plus petits éléments. Ainsi en va-t-il des archives Web qui sont, par construction, des instantanées du Web vivant : une suite de blocs DAFF régulièrement collectés et associés à des dates de téléchargement.

Mais dans le temps des archives il n'y pas de durée. Toute page collectée n'a d'étendue temporelle que sa seule date de téléchargement. Sur ce point, l'un des enjeux des explorations à venir sera justement de réinstaller de la durée dans les corpus archivés. Les phénomènes que nous souhaitons observer et étudier ont besoin d'être rapportés à une durée. Que l'on parle de l'évolution lente d'une communauté de bloggeurs ou de l'éruption soudaine d'un événement dans un forum de discussion, il faudra à chaque fois pouvoir en éprouver l'étendue dans le temps.

Pour réintégrer de la durée dans les archives, nous nous proposons de discuter de la notion de **persistance**. Une page archivée sera dite persistante si d'une version à l'autre, son contenu reste inchangé. Dans la formalisme DAFF, les données des pages sont identifiées par des clés SHA-256 (Section 2.2). Ces clés sont des signatures uniques construites à partir du contenu même des pages archivées. Ainsi, en comparant les deux clés SHA-256 de deux versions successivement crawlées d'une même page, il est possible de savoir si cette page a évolué ou non. Par ce procédé, nous pouvons identifier des chaînes de persistance entre différents collectages.

Chaque chaîne de persistance s'ouvre sur une date de **dernière modification**. Nous appelons ainsi $\mu_i(p_j)$ la date de dernière modification d'une page p_j au cours d'un crawl c_i, avec $\mu_i(p_j) \leq t_i(p_j)$. Par définition, au sein d'un même crawl, la date de dernière modification d'une

page précédera toujours (ou sera égale à) sa date de téléchargement. La Figure 3.2 donne à voir des chaînes de persistance entre les multiples captures de la page p_1. Notons tout de même, que nous faisons ici l'hypothèse qu'il n'existe pas de changements cachés[3] entre deux versions identiques de la même page.

3. Par exemple, un article est publié sur une page puis supprimé dans la foulée, et ce, entre deux crawls successifs.

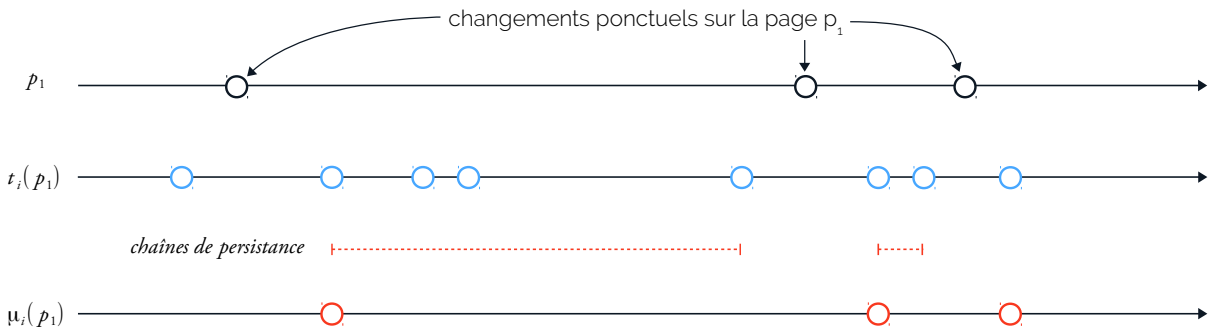


FIGURE 3.2: Chaînes de persistance entre captures (bleu) et dates de dernière modification (rouge) pour la page p_1

Intuitivement, il devient alors possible de dire d'une page archivée qu'elle n'a pas évolué depuis telle ou telle collecte, qu'elle a duré dans le temps. De plus, grâce aux chaînes de persistance, la datation des corpus d'archives Web s'affine, se fait plus précise. Une page ne sera plus maintenant seulement rapportée à sa seule date de téléchargement mais également à sa date de dernière modification, potentiellement bien antérieure. La Table 3.1 propose ainsi une échelle de datation, utile pour évaluer la précision historique d'un élément du Web passé dont l'unité d'analyse reste, pour le moment, la page Web :

TABLE 3.1: Échelle de datation d'une page Web archivée

Unité	Nature de la date
page	lancement du crawl
page	téléchargement
page	dernière modification

↓ précision historique

En nous appuyant sur cette grammaire, nous souhaitons maintenant discuter de trois biais majeurs dont il faut prendre connaissance avant de débuter toute exploration.

Cécité de crawl

À ce que nous appelons **cécité de crawl**[4] correspond l'ensemble des changements subis par une page Web mais non captés par le crawler ou,

4. *Crawl blindness* en anglais

tout au moins, mal daté par ce dernier. C'est une notion assez intuitive, dont nous donnons une illustration avec la Figure 3.3. Dans cet exemple, une page p_1 subit quatre évolutions successives e_1, e_2, e_3, e_4 correspondant respectivement à : la publication d'une image accompagné d'un texte (e_1 puis e_2), la publication d'une seconde image directement suivie par sa suppression (e_3 puis e_4). Aux yeux de l'explorateur seul le résultat de e_2 restera gravé dans les archives (points bleus). Jamais il n'aura connaissance de l'état e_1 ni même de l'existence de e_3.

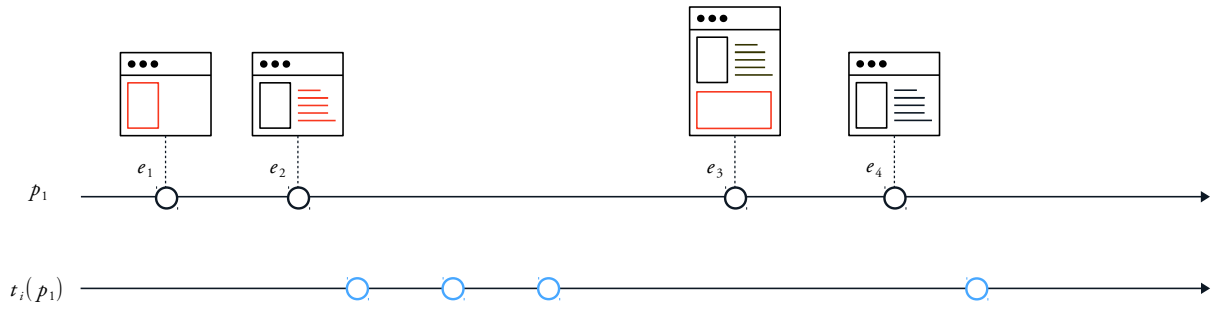


FIGURE 3.3: Cécité de crawl pour une page p_1

Ces loupés sont essentiellement dus à la difficulté de calibrer un crawler vis-à-vis de la fréquence de mise à jour d'un site (Section 2.2).

Cohérence entre pages

Dans les archives discrétisées du Web, deux pages collectées ne sont pas forcément cohérentes entre elles. Prenons l'exemple de deux pages du Web vivant connectées par un lien de citation hypertexte. L'une citant l'autre. Depuis la présentation de l'Atlas e-Diasporas (Section 1.3), nous savons à quel point la nature de ces liens est importante aux yeux des sociologues et des historiens.

Mais qu'en est-il, si dans les archives la capture de ces sites est espacée de plusieurs mois ou de plusieurs années ? Ce lien a-t-il encore du sens ? Peut on dire que ces pages sont toujours cohérentes entre elles ? Sur ce point, M. Spaniol (Spaniol *et al.*, 2009) propose une définition générale de la **cohérence** entre deux pages archivées, ainsi :

1. *Une page est toujours cohérente avec elle même*
2. *L'intervalle d'invariance* $[\mu_i(p_j), \mu_i^*(p_j)]$ *de la page* p_j *est borné par la date de dernière modification* $\mu_i(p_j)$ *par rapport à* $t_i(p_j)$ *et le prochain changement* $\mu_i^*(p_j)$ *subit par* p_j *directement après* $t_i(p_j)$
3. *Deux pages ou plus sont cohérentes si il existe un seul point dans le temps*

(ou un intervalle) $t_{\text{coherence}}$ tel que l'on puisse trouver une intersection non vide des intervalles d'invariance de toutes ces pages :

$$\forall p_j, \exists t_{\text{coherence}} : t_{\text{coherence}} \in \bigcap_j [\mu_i(p_j), \mu_i^*(p_j)] \neq \emptyset$$

La cohérence, telle qu'énoncée ici, est une cohérence absolue et théorique. Or, en pratique il est impossible d'anticiper la date de prochain changement $\mu_i^*(p_j)$, celle-ci étant inconnue au moment du crawl[5]. Ce faisant, une autre notion doit être introduire (Spaniol *et al.*, 2009). Nommée **cohérence par observation**[6], elle est définie comme suit :

5. Le crawler ne connait que les changements antérieurs à sa propre campagne de collecte.
6. *Observable coherence* en anglais

Deux pages ou plus sont cohérentes par observation, si il existe un seul point dans le temps $t_{\text{coherence}}$ tel que l'on puisse trouver une intersection non vide d'intervalles couvrant respectivement la date de téléchargement $t_i(p_j)$ et la date de dernière modification correspondante $\mu_i(p_j)$ (avec $\mu_i(p_j) \leq t_i(p_j)$) :

$$\forall p_j, \exists t_{\text{coherence}} : t_{\text{coherence}} \in \bigcap_j [\mu_i(p_j), t_i(p_j)] \neq \emptyset$$

La Figure 3.4 illustre, la notion de cohérence par observation entre deux pages crawlées p_1 et p_2. Les intervalles d'invariance respectifs de chacune de ces pages se chevauchent et forme un **intervalle de cohérence** non vide, p_1 et p_2 sont donc cohérentes entre elles.

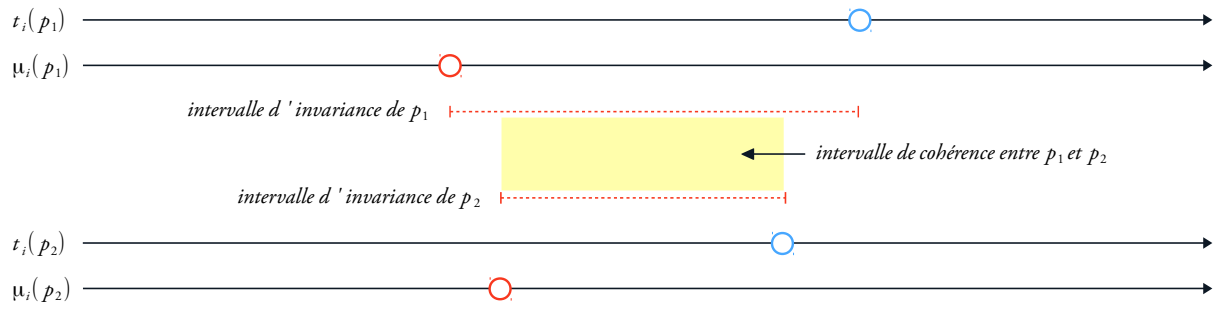


FIGURE 3.4: Cohérence par observation entre les pages p_1 et p_2 et intervalle de cohérence (jaune)

Rappelons que les définitions de cohérence et de cohérence par observation sont proposées dans le cadre d'un crawl unique et en cours de collecte, il pourrait être intéressant d'étendre leurs énoncés dans le cas de plusieurs crawls successifs et terminés.

Contenus dupliqués

L'une des particularités du formalisme DAFF est justement de ne pas dupliquer dans les archives des ressources Web qui n'auraient pas évolué (Section 2.2). Seules les pages ayant subi une transformation sont ainsi re-collectées. Néanmoins, d'un crawl à l'autre, il est possible qu'une partie du contenu de la page soit similaire à la version précédemment capturée et ce malgré les divers changements qu'elle aurait pu subir. On pense notamment aux pages d'accueil des sites d'actualités ou des forums qui présentent des informations publiées sous la forme de listes où chaque nouvel élément est inséré en en-tête. Mécaniquement certains éléments peuvent se retrouver, à plusieurs reprises, enregistrés dans les archives comme le présente la Figure 3.5.

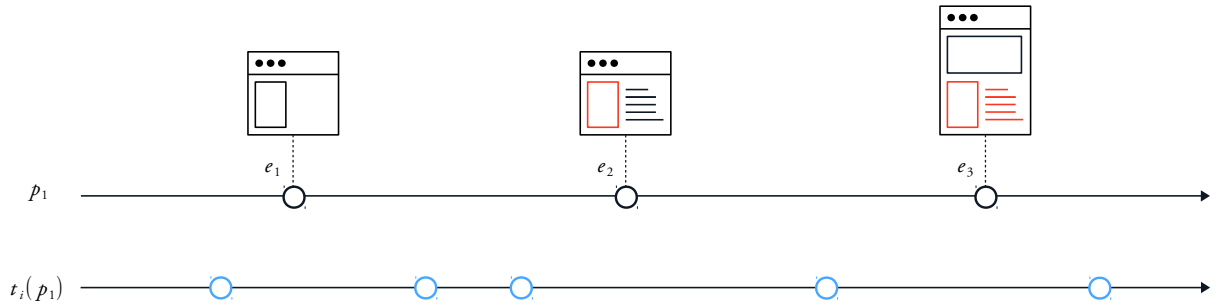


Figure 3.5: Contenu d'une page (en rouge) collecté plusieurs fois

Cela peut poser de lourds biais d'analyse si l'on cherche par exemple à connaitre la distribution d'un mot-clé extrait des pages archivées. Ce dernier pourra, du fait de la structure même du corpus, être artificiellement sur-représenté dans les résultats.

3.3 Construire un moteur d'exploration d'archives Web

Ces différents biais maintenant présentés, nous pouvons nous tourner vers la description de l'architecture de notre moteur d'exploration d'archives Web. Nos corpus étant en DAFF, il n'a pas forcément été possible de réutiliser des éléments open-source issus d'autres moteurs (quasiment tous conçus pour accueillir du WARC). De fait cette section est intéressante pour qui souhaite mettre en place ou comprendre dans le détail la mécanique d'une tel système. Notre architecture suit néanmoins la structure classique d'une chaîne d'extraction, d'analyse et de visualisation de données à grande échelle (Marz et Warren, 2015).

La Figure 3.6 donne à voir une illustration de son fonctionnement. Il s'agira donc ici d'une description plutôt orientée ingénierie dont la majeure partie a fait l'objet d'une publication démonstration[7].

7. Lobbé, Q. (2018), *Revealing Historical Events out of Web Archives*, TPDL 2018

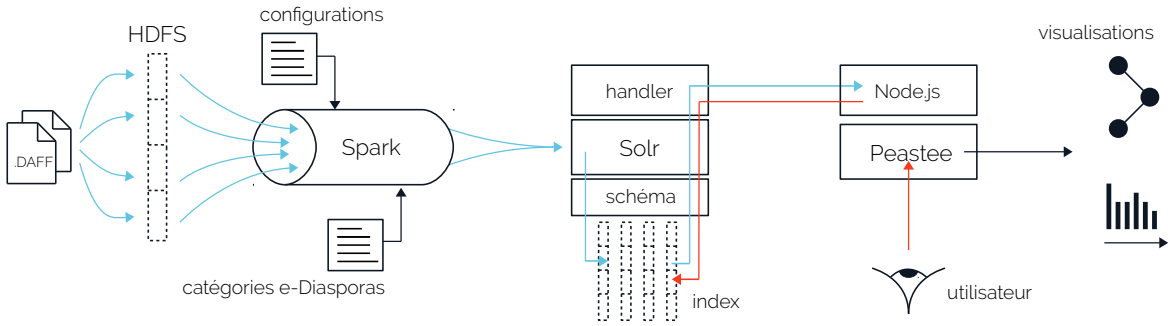


FIGURE 3.6: Architecture de notre moteur d'exploration d'archives Web

Extraction et enrichissement

Nous suivons ici l'hypothèse selon laquelle notre moteur doit rendre accessible les archives Web d'une seule e-Diaspora à la fois. Nous prendrons comme exemple les archives du corpus marocain.

La première étape consiste à extraire les informations contenues dans les fichiers DAFF. Rappelons que chaque corpus est divisé en deux fichiers DAFF : les données d'une part (*data*) et les méta-données (*metadata*) d'autre part. Pour ce faire, nous commençons par adapter une librairie Java (fournie par les équipes de l'INA, *dlweb-commons*) qui cherche à transférer les archives des fichiers DAFF vers le système de stockage d'Hadoop[8], le Hadoop Distributed File System (HDFS). Le HDFS est un système de fichiers qui permet de manipuler de larges volumes de données, de manière distribuée (c.-à-d., réparti entre plusieurs machines) et passant relativement bien à l'échelle (pouvant supporter une forte montée en charge). Le format DAFF a ceci de limitant, qu'il reste pensé pour le stockage et non pour la manipulation des données. Filtrer un fichier DAFF par URL ou date de téléchargement n'est, par exemple, pas trivial.

8. `https://hadoop.apache.org/`, `https://fr.wikipedia.org/wiki/Hadoop`

Une fois chargées dans le HDFS, les données et méta-données DAFF sont envoyées dans un système de traitement nommé Spark[9]. Spark permet de travailler par batchs (c.-à-d., par petits lots de données) dans un environnement distribué : les données et méta-données sont segmentées en sous-ensembles plus facilement manipulables, puis répartis sur plusieurs machines où elles subissent toutes les mêmes traitements en parallèle (filtres, jointure, groupement, etc.). Spark est un outil flexible dans lequel nous pouvons définir une suite d'instructions (ou *pipeline*) ayant pour finalité la fusion des données et méta-données DAFF en une

9. `https://spark.apache.org/`

seule et même source de données d'archives. La Figure 3.7 décrit la manière dont s'enchainent ces diverses transformations. Les méta-données sont traitées en premier et peuvent suivant la configuration du système être filtrées par date de téléchargement ou nom de domaine (par site Web). Puis, en nous rappelant que les méta-données possèdent chacune un pointeur vers le bloc de données DAFF dont elles sont l'extension (champ *content* en DAFF, Section 2.2), nous remplaçons l'identifiant des méta-données par l'identifiant du bloc de données correspondant. Cette manipulation nous permet ensuite de grouper les méta-données par identifiants communs, c'est-à-dire, toutes les méta-données d'une seule et même chaine de persistance (Section 3.2).

C'est à cette étape que nous identifions notamment les dates de dernière modification. De là, nous opérons une jointure entre les méta-données et données DAFF afin de rassembler enfin les informations issues du crawler et le contenu même des pages archivées. Pour terminer, nous préparons ces nouvelles données à être envoyées dans le moteur de recherche, sans oublier de les enrichir avec des informations tirées de l'Atlas e-Diasporas (type de sites, langue, etc.).

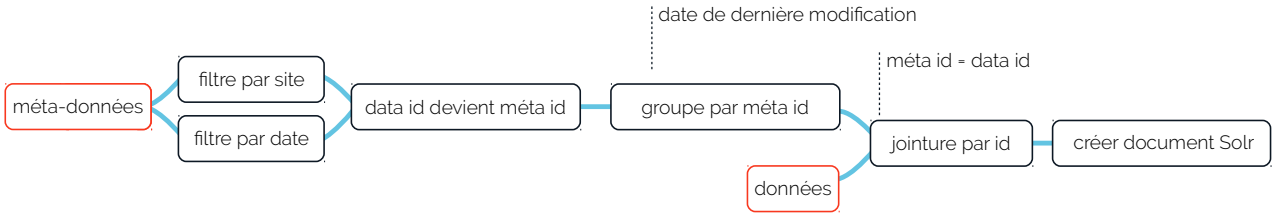


Figure 3.7: Pipeline de transformation des données et méta-données DAFF

Deux configurations différentes ont été testées pour Spark, l'une distribuée entre plusieurs machines d'un même *cluster* (c.-à-d., un groupe de machines), l'autre distribuée sur l'ensemble des cœurs d'une seule machine puissante. Si la première configuration s'est révélée la plus rapide (traitement de l'ensemble du corpus marocain en 3 jours), il a fallut néanmoins s'en détourner. En effet, Spark étant particulièrement dur à piloter sur un réseau souvent instable, il était fréquent de voir le traitement des données s'arrêter après avoir perdu connexion avec une ou plusieurs machines. De fait, nous avons préféré nous contenter de la seconde configuration, plus lente (une dizaine de jours) mais garantissant la totalité du traitement [10].

10. Notre système est téléchargeable ici-même `https://github.com/lobbeque/archive-miner`

Adapter un moteur de recherche

Au cours de la Section 2.2, nous avons présenté l'ensemble des méthodes aujourd'hui utilisées pour consulter des corpus d'archives Web.

La plupart d'entre elles, s'appuient sur l'utilisation de moteurs de recherches qui offrent la possibilité d'interroger une base de documents [11] en plein texte. De notre côté, nous avons choisi d'adapter, à la nature particulière de nos archives Web, une solution open-source existante nommée Solr/Lucene.

Solr [12] est un serveur de *search*, c'est-à-dire qu'il permet de faire le lien entre une requête utilisateur (un mot-clé, une dimension, etc.) et un ensemble de documents préalablement indexés. Solr est construit au-dessus de la librairie d'indexation Lucene [13] dont le principe de base est de stocker un texte dans un **index inversé**.

Le fonctionnement simple d'un index inversé est illustré par la Figure 3.8. Un premier tableau de données (a) renferme le contenu de deux documents *doc*1 et *doc*2 dont les identifiants (*doc id*) sont respectivement 0 et 1. Un second tableau de données (b) contient ce que l'on appelle un dictionnaire de termes (*term dict*). Dans ce dictionnaire, sont répartis l'ensemble des mots uniques (numérotés de 0 à 4) contenus dans les documents 0 et 1. Ces mots sont identifiés par des *term ids* et sont, de plus, associés à une *posting list* qui fait correspondre à chaque mot la liste des documents où il est présent.

Ainsi, pour réaliser une recherche plein texte, il suffira de parcourir l'ensemble du dictionnaire jusqu'à trouver les mots correspondant à la requête de l'utilisateur et, par extension, les documents associés. Ces deux tableaux de données (a) et (b) forment ce que l'on appelle un **segment**, soit le bloc de base de tout index inversé. On nomme **indexation** l'action de stocker un texte dans un index inversé. Différentes stratégies peuvent être mises en place pour accélérer la recherche dans un index, en optimiser la taille, etc. [14].

Un moteur de recherche se doit d'ordonner ses résultats avant de les retourner à l'utilisateur. On nomme cette étape le **classement** (*ranking*). Pour cela, il fait appel à une **fonction de similarité** qui trie les documents résultants en leur attribuant à chacun un score. Parmi les nombreux critères de ranking, les systèmes de search favoriseront souvent les documents où les mots clés recherchés sont les plus fréquents. Cette mesure sera pondérée par l'ajout d'une prime aux termes rares, c'est-à-dire : peu présents dans l'ensemble des index. C'est tout le sens du fameux *tf-idf* [15] dont nous réutilisons ici une version légèrement modifiée : la *defaultSimilarity* de Lucene [16].

Nous n'avons pas eu l'occasion de tester une fonction de similarité propre aux archives Web, la nôtre est finalement très générique. C'est pourtant une question intéressante, puisque comme le suggère (Weikum *et al.*, 2011) les moteurs d'exploration d'archives pourraient prendre en compte l'aspect temporel des documents collectés pour procéder au classement des résultats.

11. Les données manipulées par des moteurs de recherche sont de manière générale appelées **documents**.

12. Voir (Grainger *et al.*, 2014) et `http://lucene.apache.org/solr/`

13. Voir (McCandless *et al.*, 2010) et `http://lucene.apache.org/index.html`

(a)

doc id	document
0	funes ou la mémoire
1	le livre de sable

(b)

term id	term dict	posting list
0	infini	0,1
1	livre	0,1
2	funes	0
3	mémoire	0
4	sable	1

FIGURE 3.8: Principe de base d'un index inversé

14. Pour de plus amples détails sur l'indexation via Lucene/Solr, voir mon cours sur le fonctionnement interne des moteurs de recherche (Lobbé, Q. 2016, Voyage au cœur d'un index Lucene, `http://qlobbe.net/ressources/search.pdf`).

15. *Term frequency-inverse document frequency*, fonction de similarité qui évalue l'importance d'un terme dans un document au regard d'un corpus donné (`https://fr.wikipedia.org/wiki/TF-IDF`).

16. En plus du simple tf-df, cette fonction de similarité prend en compte la taille du document et la taille des champs vérifiant la requête (`https://frama.link/RsYNvx-6`).

Le schéma d'indexation

Tout document destiné à l'indexation doit d'abord passer au crible du **schéma** qui est considéré comme la pierre angulaire de tout moteur de recherche.

Le schéma est un fichier décrivant dans le détail la façon dont tout document sera indexé, il forme l'ossature de l'indexation. En effet, un document n'est pas indexé d'un seul tenant. Pour maximiser les chances de le voir satisfaire une requête, il peut être nécessaire de le découper en plusieurs **champs** (*fields*), ayant chacun des attributs particuliers. Par exemple, un article issu d'un site de news pourra être segmenté suivant son titre, sa date, l'auteur et finalement le cœur du texte. Ce dernier sera indexé de manière classique en vue d'une recherche plein texte, l'auteur en revanche pourra être destiné à une recherche par *facet*, c'est-à-dire par **dimension** (par exemple : Quels sont les textes de tel ou tel auteur ?). Si tel est le cas, son indexation se fera via des *docValues* [17]. Dans notre schéma, la plupart des informations issues du fichier de méta-données DAFF sont destinées à une recherche par dimension, telles que :

17. https://lucene.apache.org/solr/guide/6_6/docvalues.html

```
<!-- archive fields -->
<field name="id"                type="string"  indexed="true"    multiValued="false" required="true" />
<field name="archive_active"    type="boolean" indexed="true"    multiValued="false"/>
<field name="archive_corpus"    type="string"  indexed="true"    docValues="true" multiValued="false"/>
<field name="archive_ip"        type="string"  indexed="true"    docValues="true" multiValued="false"/>
<field name="archive_length"    type="double"  indexed="true"    docValues="true" multiValued="false"/>
<field name="archive_level"     type="int"     indexed="true"    docValues="true" multiValued="false"/>
<field name="archive_referer"   type="string"  indexed="true"    docValues="true" multiValued="false"/>
<field name="archive_mime"      type="string"  indexed="true"    docValues="true" multiValued="false"/>
<field name="is_page"           type="boolean" indexed="true"    multiValued="false" default="false"/>

<!-- client fields -->
<field name="client_country"    type="string"  indexed="true"    docValues="true" multiValued="false"/>
<field name="client_ip"         type="string"  indexed="true"    docValues="true" multiValued="false"/>
<field name="client_lang"       type="string"  indexed="true"    docValues="true" multiValued="true" />

<!-- crawl fields -->
<field name="crawl_id"          type="string"  indexed="true"    docValues="true" multiValued="true" />
<field name="crawl_id_f"        type="string"  indexed="true"    docValues="true" multiValued="false"/>
<field name="crawl_id_l"        type="string"  indexed="true"    docValues="true" multiValued="false"/>
<field name="crawl_date"        type="date"    indexed="true"    docValues="true" multiValued="true" />
<field name="crawl_date_f"      type="date"    indexed="true"    docValues="true" multiValued="fasle"/>
<field name="crawl_date_l"      type="date"    indexed="true"    docValues="true" multiValued="true" />
```

```
<!-- download fields -->
<field name="download_date"     type="date"    indexed="true"    docValues="true" multiValued="true" />
<field name="download_date_f"   type="date"    indexed="true"    docValues="true" multiValued="false"/>
<field name="download_date_l"   type="date"    indexed="true"    docValues="true" multiValued="false"/>

<!-- page fields -->
<field name="page_site"         type="string"  indexed="true"    docValues="true" multiValued="false"/>
<field name="page_url"          type="string"  indexed="true"    docValues="true" multiValued="false"/>
<field name="page_url_id"       type="string"  indexed="true"    docValues="true" multiValued="false"/>

<!-- extracted page fields -->
<field name="page_link"         type="string"  indexed="true"    docValues="true"  multiValued="true"/>
<field name="page_meta_title"   type="string"  indexed="true"    docValues="false" multiValued="false"/>
<field name="page_meta_desc"    type="text"    indexed="true"    docValues="false" multiValued="false"/>
<field name="page_meta_img"     type="string"  indexed="true"    docValues="false" multiValued="false"/>
<field name="page_meta_date"    type="date"    indexed="true"    docValues="true"  multiValued="false"/>
<field name="page_meta_author"  type="string"  indexed="true"    docValues="true"  multiValued="false"/>
<field name="page_title"        type="text"    indexed="true"    docValues="false" multiValued="false"/>

<!-- searchable page fields -->
<field name="page_text"         type="text"    indexed="true"  stored="false" multiValued="true"/>
<field name="page_text_shingle" type="shingle" indexed="true"  stored="false" multiValued="true"/>
```

FIGURE 3.9: Schéma d'indexation de notre moteur d'exploration d'archives Web

Pour chaque champ, nous définissons un nom et un type : une date, un nombre (int), du texte (string et text)... Tous les champs sont indexés (`indexed="true"`). Ceux destinés à une recherche par dimension sont associés à une docValue. Certain champs, comme les dates de téléchargement, sont multivalués (une page peut avoir été crawlée plusieurs fois à l'identique). Le champ de recherche plein text par défaut est le champ page_text qui couvre l'ensemble du texte d'une page archivée [18]. Dans Spark, lors de la transformation des DAFF en document Solr, nous extrayons les liens de citation hypertextes (`page_link`) et les informations contenues dans l'en tête des pages [19] (`page_meta_img`, `page_meta_date`, etc.). Les diverses dates associées à une page sont également présentes : allant de la date de lancement du crawl (`crawl_date_f`), à la date de téléchargement (`download_date`) et en passant par la date de dernière modification (`download_date_f`) [20].

Mais les documents à indexer ne sont pas les seuls à devoir passer par le schéma. En effet, un moteur de recherche est un système à double entrée (Figure 3.6), conjuguant des documents et des requêtes utilisateurs grâce à une fonction de similarité. Pour ce faire, documents et requêtes doivent parler la *même langue*, c'est-à-dire qu'une requête utilisateur devra être traitée de la même manière que les documents à indexer, subir les mêmes transformations et interroger les bons champs. Cette suite de transformations est appelée **analyzer** [21]. Sur ce point, documents et requêtes suivent les traitements suivants : tout d'abord, les majuscules deviennent minuscules (Figure 3.10, a), puis le texte

18. Ce champ subit un traitement particulier (un découpage en bi-gramme : page_text_shingle) dont nous reparlerons en Section 4.5

19. Tout ce qui est contenu dans les balises HTML <*meta*> (`https://www.w3schools.com/Tags/tag_meta.asp`)

20. Nous avons, un temps, intégré à notre schéma les informations de l'en-tête HTTP des pages archivées. Mais étant trop peu fiable, nous avons décidé de ne plus les prendre en compte.

21. Pour plus d'informations sur les analyzers et tokenizers dans Solr : `https://wiki.apache.org/solr/LanguageAnalysis`

22. Mots qui ne sont généralement pas intéressants pour l'analyse : et, il, elle... (voir `https://fr.wikipedia.org/wiki/Mot_vide`)

est découpé en termes distincts (Figure 3.10, b) et les *stopwords*[22] sont écartés (Figure 3.10, c), s'en suit une phase appelée **stemming** dans laquelle on ne garde finalement que la racine des termes restants (Figure 3.10, d) avant indexation (Figure 3.10, e).

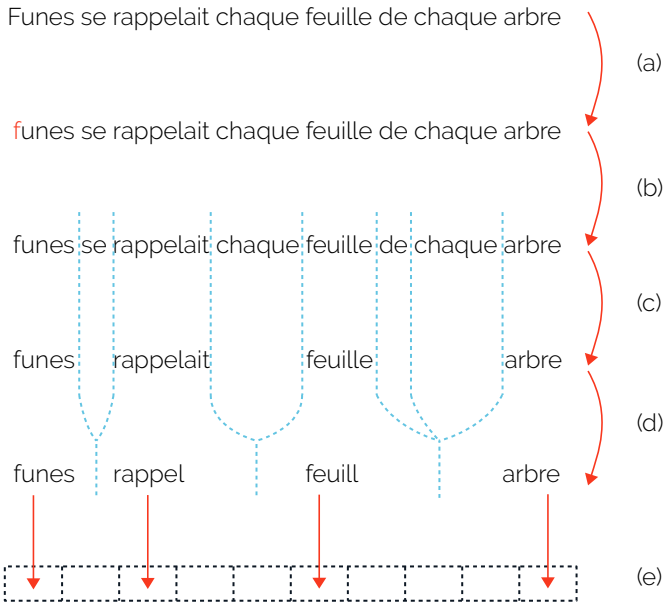


FIGURE 3.10: Cycle de transformation d'un texte dans notre moteur de recherche

Dans le cas particulier des archives Web, l'exploration de pages collectées peut être focalisée autour d'un instant précis. Si l'utilisateur en fait la demande, notre moteur lui proposera différentes stratégies de recherche pour retrouver les pages les plus proches d'une date donnée : soit en amont (Figure 3.11, a) , soit en aval (Figure 3.11, c) ou soit autour de cette dernière (Figure 3.11, b). Internet Archive par exemple utilise la première option dans la WayBack Machine.

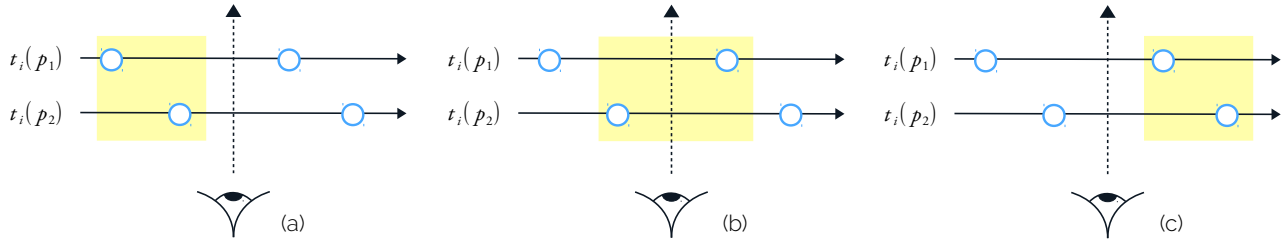


FIGURE 3.11: Stratégies de choix d'un ensemble de pages par rapport à une date précise

Notre moteur de recherche est finalement déployé sur les serveurs. Les index sont distribués entre plusieurs machines pour améliorer les temps

de réponse et d'indexation du système. On appelle **sharding** l'action de scinder un index en plusieurs sous-index avant de les distribuer. Nous suivons une configuration classique *maître–esclave* [23] où l'instance maître de notre moteur de recherche centralise les requêtes utilisateur avant de les dispatcher entre ses diverses instances esclaves qui, elles seules, sont habilitées à retourner des résultats [24].

23. `https://lucene.apache.org/solr/guide/6_6/solrcloud.html`

24. Notre moteur de recherche peut être téléchargé ici `https://github.com/lobbeque/archive-search`

Interface de visualisation

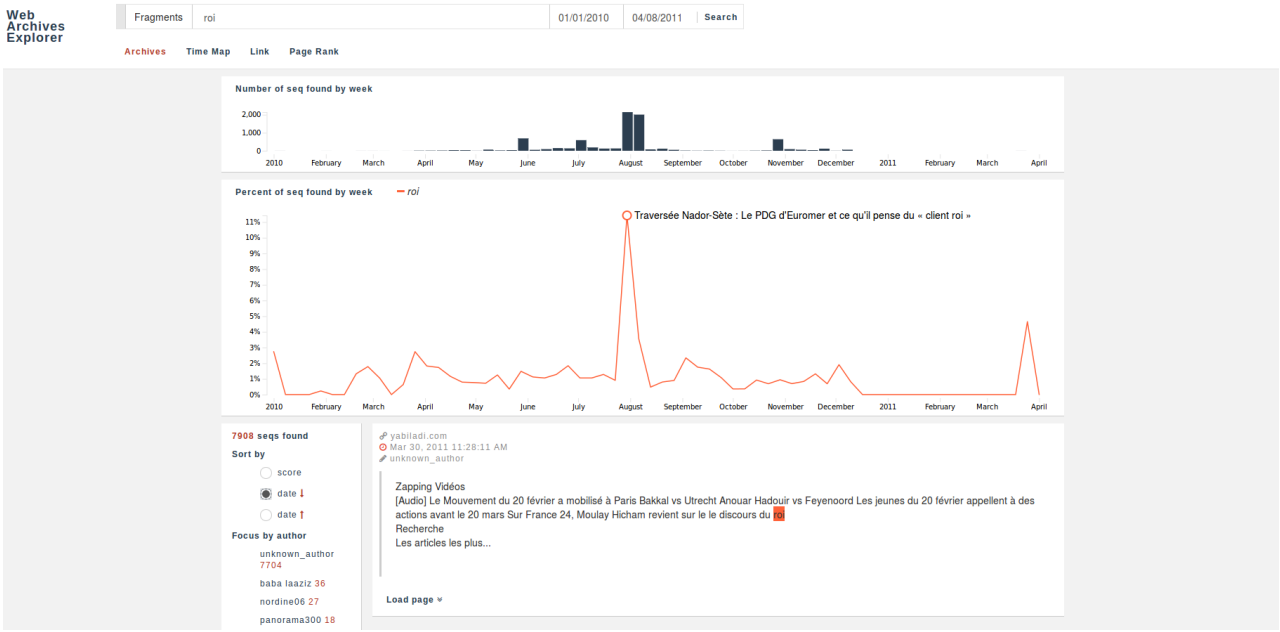


FIGURE 3.12: Capture d'écran de notre interface de visualisation

Un système de visualisation et d'interrogation des archives est développé [25] au-dessus de notre moteur. Il s'agit en fait d'une *application Web* permettant à tout utilisateur d'écrire une requête et de se voir présenter les résultats sous diverses formes. Une application Web est composée de deux briques logicielles distinctes : un serveur en *node.js* [26] se charge tout d'abord de faire la liaison avec le moteur de recherche, une interface Web permet ensuite de visualiser les documents archivés. Les éléments de visualisation sont développés en *d3.js* et l'architecture de l'interface en tant que telle se base sur *angularjs* [27]. Notre interface suit un modèle *en liste* très classique : les résultats sont présentés les uns à la suite des autres et des facets, à la marge, permettent de les filtrer ou de les trier après coup. Divers histogrammes offrent à voir une répartition dans le temps des pages archivées ayant correspondu à la requête de l'utilisateur. La Figure 3.12 présente une capture d'écran de cette interface, une démonstration en vidéo permet de se faire une

25. Nommé *Peastee* en référence au narrateur de la nouvelle de H.P. Lovecraft *The Shadow Out of Time* (1935), ce système est téléchargeable ici `https://github.com/lobbeque/peastee`

26. `https://nodejs.org/en/about/`

27. D3 est une librairie Javascript de visualisation de données `https://d3js.org/`, Angular est un framework Javascript pour les applications Web `https://angularjs.org/`

idée plus précise de son fonctionnement[28]

28. https://youtu.be/snW4O-usyTM

Notre application Web accueille les nombreux prototypes que nous avons pu expérimenter tout au long de ces trois années de travail. Nous ne reviendrons pas ici en détail sur leurs développements respectifs, mais, bien qu'incomplets ou inaboutis, ces prototypes restent des jalons qui nous ont permis de cheminer vers les résultats présentés au Chapitre 5. Nous retiendrons, premièrement, une visualisation *en oursin* de l'arborescence des URL d'un site archivé mois après mois (Figure 3.13). La racine d'un site (sa page d'accueil) est située au centre de l'oursin, plus on s'en éloigne et plus on est censé découvrir les diverses pages constitutives de ce site.

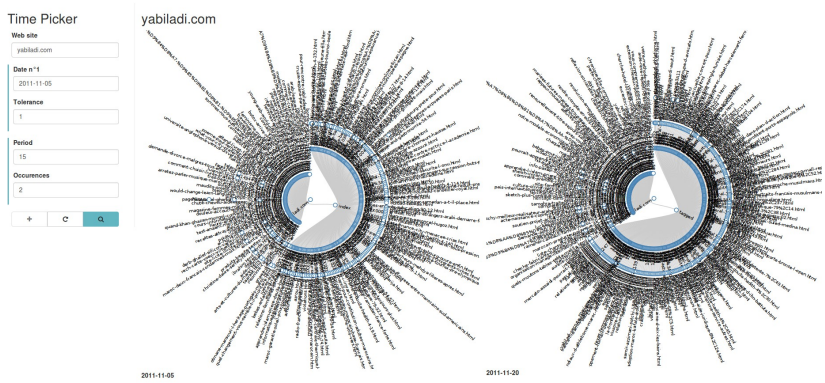


Figure 3.13: Prototypes de visualisation de la structure des URL d'un site archivé

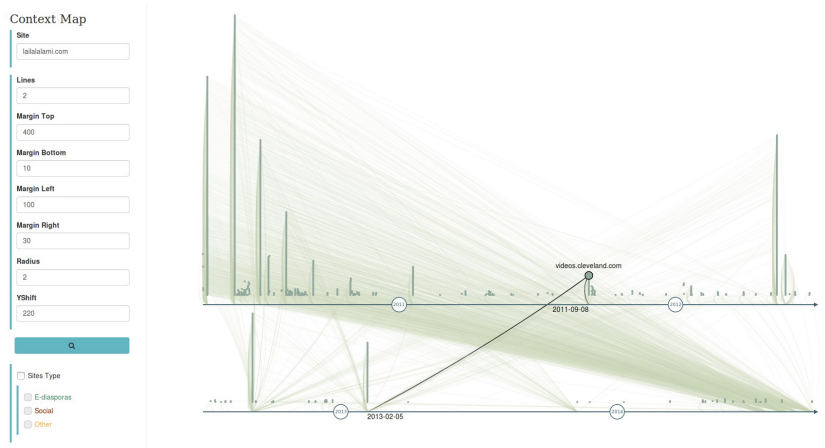


Figure 3.14: Prototypes de visualisation des liens hypertextes produits par un site archivé

Pour terminer, un autre prototype propose de voir la distribution

temporelle (en abscisse) des liens hypertextes (en ordonnée) sortant d'un site archivé (Figure 3.14, différenciés par catégories : réseaux sociaux, sites migrants e-Diasporas et reste du Web), la production de liens de citation pouvant être un bon indicateur d'activité dans le temps.

FIGURE 3.15: *yabiladi.com* (rouge) dans l'e-Diaspora marocaine (Section 1.3)

3.4 Les archives ne sont pas des traces directes du Web

Notre moteur d'exploration maintenant présenté, nous voilà enfin en mesure d'interroger les archives de l'Atlas e-Diasporas. Depuis le Chapitre 1, le site *yabiladi.com* et plus particulièrement son forum de discussion attire notre attention. De part la place qu'il occupe dans le corpus marocain depuis le début des années 2000, le site a su jouer un rôle clé pour l'ensemble de la diaspora en ligne. Notre première requête cherche donc à nous faire découvrir la répartition temporelle de *yabiladi.com* dans les archives. Nous voulons saisir et comprendre la dynamique des publications postées sur sa section forum afin d'identifier des moments-clés de l'histoire du site.

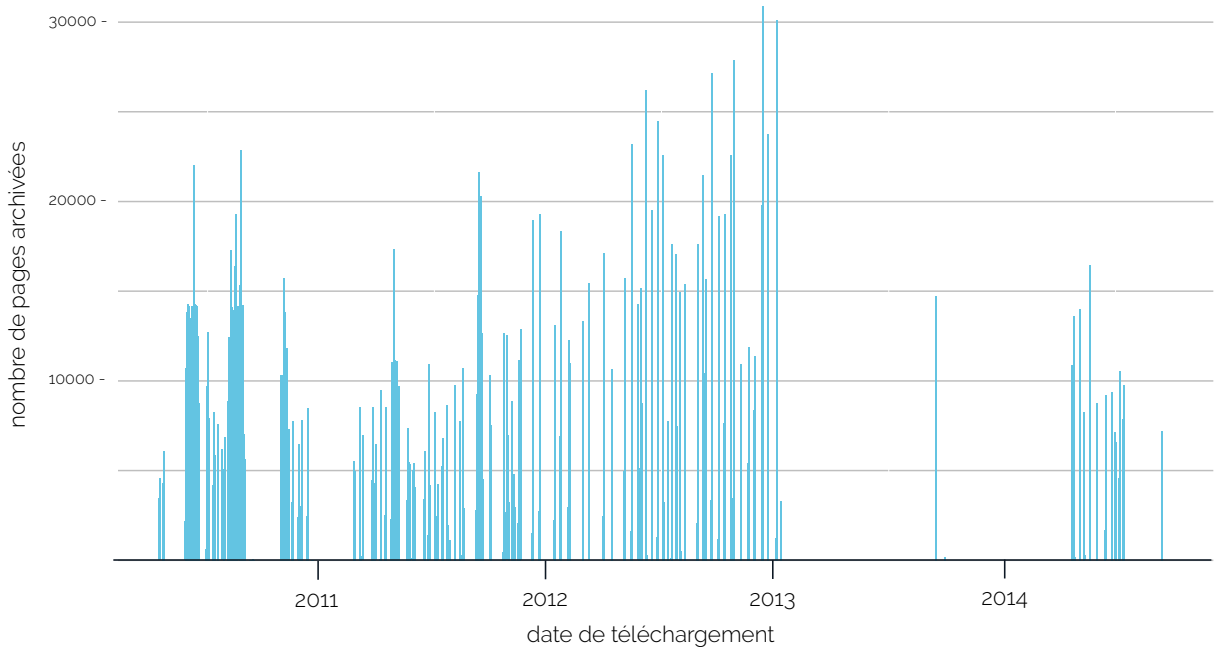


FIGURE 3.16: Distribution du nombre de pages archivées par jour pour *yabiladi.com*

Mais les résultats que nous retourne notre moteur d'exploration semblent très curieux, la Figure 3.16 présente ainsi la répartition par jour, du nombre de pages forum collectées pour *yabiladi.com*, de Mars 2010 à Septembre 2014. Ainsi, il semble que le site ait littéralement cessé de

produire du contenu de Janvier 2013 à début 2014. Or, si l'on passe maintenant par la Wayback Machine [29], on se rend rapidement compte que chez, Internet Archive, des ressources Web ont bel et bien été capturées à ces même dates. Où se situe donc notre erreur ?

29. Voir `https://frama.link/Jv7WpLCJ`

Nous interprétons nos résultats de la mauvaise manière. Depuis le début de ce chapitre, nous cherchons à mettre en évidence les nombreux biais d'analyse possibles pour qui souhaite mener à bien une exploration d'archives Web. Nous avons notamment évoqué les cécités de crawl (Section 3.2), mais sans nous attendre à trouver dans nos propres corpus une telle défaillance. Après enquête auprès des équipes de l'INA, il s'avère que le crawler de l'institution a stoppé sa collecte de *yabiladi.com* pendant toute une année, avant qu'archivistes et chercheurs ne s'en rendent compte et relancent le robot. Ce que la Figure 3.16 donne à voir est un **artéfact de crawl**, c'est-à-dire un effet mécanique du crawler qui aura influencé la forme même du collectage, au delà du simple décalage ou de l'imprécision inévitable pour ce type de campagne.

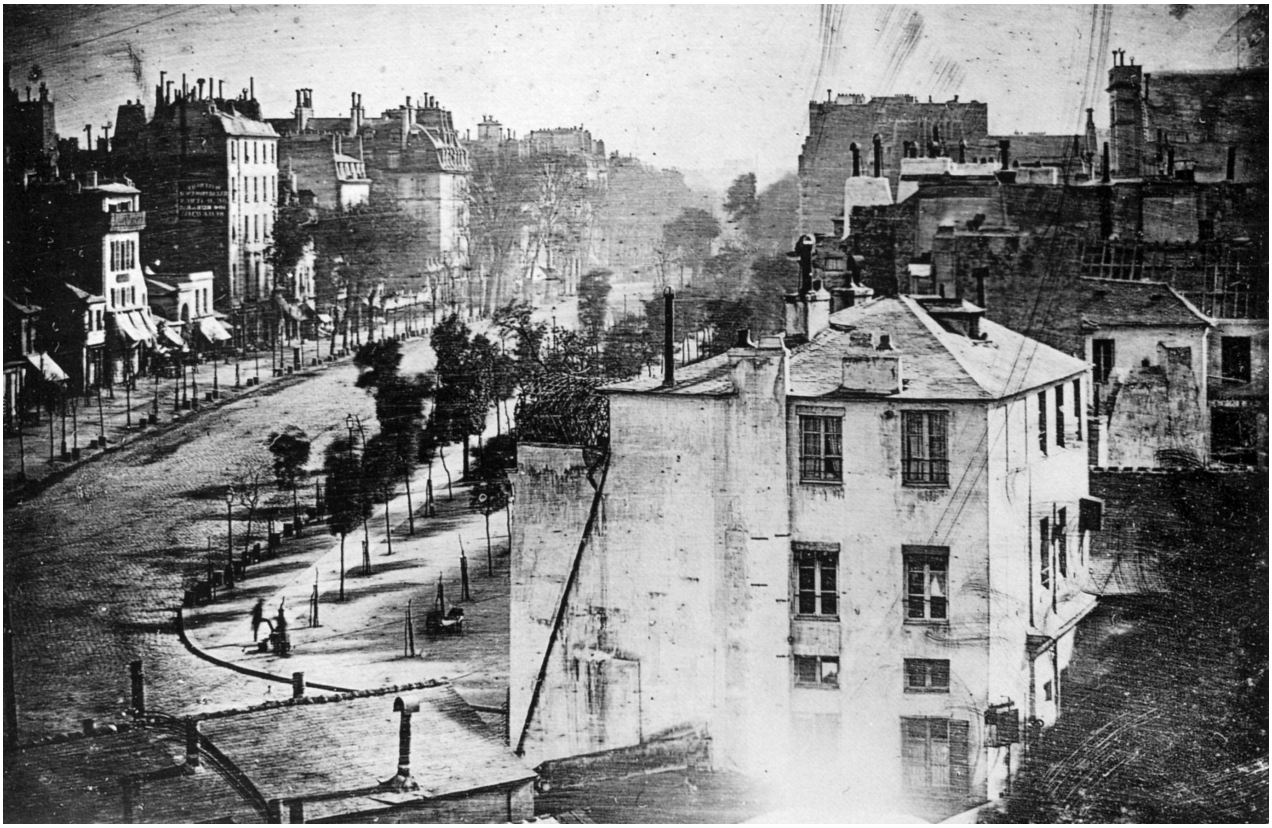


Figure 3.17: *Boulevard du Temple*, Louis Daguerre, 1838

Les archives Web ne sont pas les traces directes du Web, elle sont les

traces directes des crawlers. Les outils de collecte façonnent l'image de ce qu'ils sont supposés préserver. Ils en modifient la forme, potentiellement le fond, et par là même, la nature des interprétations que nous faisons du corpus si l'on n'y prend pas garde. Ce que l'on voit dans les archives reste avant tout le geste de l'archiviste et de ses dispositifs de capture.

***

En 1838, L. Daguerre réalise un daguerréotype [30], le *Boulevard du Temple*, qui est aujourd'hui reconnu comme l'une des première photographie figurant un être humain : deux hommes seuls dans une rue vide (Figure 3.17). En réalité, le boulevard était ce jour là bondé. Demandant un temps d'exposition particulièrement long, les seuls sujets (en plus des bâtiments et des arbres) que le dispositif a su capturer sont ceux qui étaient restés pratiquement immobiles : un cireur de chaussures et son client assis devant lui. Cet exemple illustre parfaitement ce que nous rencontrons dans les archives Web. Associées à des dates de téléchargement qui les arrachent à leur temporalité, les archives Web sont des objets discrétisés et figés. Sans lien direct avec la réalité dont elles sont pourtant censées être le reflet.

Afin d'améliorer la pertinence historique de nos analyses à venir et pour nous affranchir des crawlers et de leurs artéfacts, nous proposerons, dans la suite de cette thèse, de descendre sous le niveau des pages Web capturées et de nous appuyer sur une nouvelle unité d'exploration des corpus archivés : le **fragment Web**.

30. Procédé photographique basé sur l'exposition à la lumière d'une surface d'argent pure (`https://fr.wikipedia.org/wiki/Daguerréotype`).

Chapitre 4

Fragmenter les archives Web

Les artéfacts de crawl sont indissociables des archives Web telles que nous les connaissons (Section 3.4). Ils sont liés organiquement à la structure même des ressources collectées (Section 2.2), issus de l'association d'une page Web et d'une date de téléchargement. Ces artéfacts et leurs effets induisent nombres de biais pour qui souhaite explorer le Web passé : collectages non réguliers, sur-représentation de certaines parties d'un site, incohérences entre les contenus préservés, etc.

Nos travaux portant sur l'exploration de corpus d'archives Web déjà existants ou constitués de longue date, nous ne proposerons pas ici d'alternative aux formats WARC et DAFF. Nous chercherons plutôt à définir, partant d'une collecte terminée, une stratégie d'analyse capable de s'affranchir de l'héritage pesant des crawlers ou, tout au moins, d'en atténuer les effets. Par ailleurs, nous souhaitons mener une exploration large (en terme de pages à visiter) et profonde (en terme de durée à balayer) de nos corpus d'archives Web. Mais, nous voulons aussi garder la possibilité de nous appuyer, au besoin, sur une analyse plus fine de certains éléments. Cela implique le développement d'une méthodologie hybride capable de débrayer du quantitatif vers le qualitatif. Cette approche s'articulera autour d'une nouvelle entité qui pourra faire cohabiter traitements algorithmiques à grande échelle et campagnes de validation humaine.

Aussi, nous proposerons dans ce chapitre de changer d'unité d'exploration en introduisant les **fragments Web**. Nous pensons, en effet, qu'il peut être bénéfique de mener une analyse au-dessous du niveau des pages Web archivées. Pour valider cette intuition, le fragment Web offrira aux explorateurs une plus grande souplesse et de nouveaux outils pour interroger les archives. Il se voudra également objet d'étude à part entière. À travers les fragments Web, nous questionnerons directement le geste des auteurs et des lecteurs des sites collectés, en suivant les indices de leur passage sur la toile. Revenir à l'humain dissimulé sous les archives. Pour ce faire, nous porterons notre réflexion sur la question de la datation des archives Web en associant à chaque fragment une date d'édition. Ainsi, nous nous approcherons, au mieux, du

Web tel qu'il a été de son vivant. Enfin, nous reviendrons en miroir sur les modalités techniques et théoriques d'un moteur d'exploration basé, cette fois-ci, sur le fragment Web comme unité principale d'indexation. Un cas simple de détection d'événements dans les archives Web nous permettra d'en faire la démonstration.

4.1 En dessous des pages Web

Comme le résume (Weikum *et al.*, 2011), les archives Web sont de véritables mines d'or pour qui souhaite étudier l'histoire du Web passé [1]. Mais tout trésor est difficile d'accès et nous avons déjà évoqué, au regard de l'état de l'art (Section 2.1), à quel point les corpus archivés restaient pour nous des territoires inexplorés, repliés et fortifiés.

1. « *These archives host a wealth of information, providing a gold mine for sociological, political, business, and media analysts.* » (Weikum *et al.*, 2011)

L'archive est pourtant une matière qui ne doit pas rester fermée (Ketelaar, 2006), mais toujours prête à être questionnée. C'est au travers des lectures, discussions et interprétations successives des archives que s'écrit l'histoire. Pour plonger au cœur des archives Web, essayons d'ouvrir une brèche dans nos corpus afin d'y extraire une nouvelle entité. Ce fragment Web, comme nous le nommons, est issu de la segmentation des pages Web collectées. Sa construction s'appuie sur plusieurs éléments, plusieurs inspirations. Tout d'abord, il s'agira pour nous d'adopter une attitude plus souple vis-à-vis des archives en cherchant à les décomposer pour mieux les explorer. Ensuite, nous inscrirons les fragments dans la droite lignée des strates du Web au sens où les décrit N. Brügger (Brügger, 2009). Nous nous attarderons alors sur la question de la datation des ressources collectées en introduisant les dates d'édition à notre grammaire. Nous nous servirons de ces dates pour finalement descendre vers une plus grande précision historique et ramener les archives vers la temporalité du Web passé.

Découper, déplacer, monter

Funes vit dans l'indexation d'un présent perpétuel (Section 3.1). Condamné à ne plus jamais rien oublier, il lui devient impossible de penser, de raisonner et de s'inventer :

« *[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. [...] Penser c'est oublier des différences, c'est généraliser, abstraire. Dans le monde surchargé de Funes il n'y avait que des détails, presque immédiats.* » — (Borges, 1974, p.117-118)

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 fil de nos histoires et de nos vies. Nous

décrivions en Section 2.2, comment les conditions d'accès aux archives Web rendaient difficile leur exploration par les chercheurs [2]. Chevillées aux niveaux des seules pages Web les outils d'analyse existants (la Wayback Machine tout autant que notre propre moteur d'exploration, Section 3.3) ne nous permettent pas de manipuler les résultats de nos requêtes. Les archives sont consultables, certes, mais restent enfermées dans des *interfaces-vitrines* plutôt que de nous être restituées sur des tables de montage.

2. Notons néanmoins l'existante du projet `https://archivesunleashed.org/` et des outils développés par l'Omilab `https://github.com/omilab/`.

FIGURE 4.1: C. Marker, 1977, Le Fond de l'Air est Rouge, (`https://youtu.be/d01E4GYjF1s`)

En achevant *Le Fond de l'Air est Rouge* en 1977, le cinéaste C. Marker revient amer sur l'avènement des mouvements contestataires et révolutionnaires dans années 60, événements dont il a été le témoin direct. Il remonte ainsi et assemble 15 années de ses propres archives filmiques qu'il aborde sous un angle inédit : « *on ne sait jamais ce que l'on filme, on ne sait jamais ce qu'il y a derrière une image* » (*Le Fond de l'Air est Rouge*, Partie II, 14mn 22s) nous dit-il en voix off. Détachées de lui et faisant désormais partie de l'histoire, ses archives peuvent enfin être confrontées et ré-interrogées. En cela 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.

FIGURE 4.2: J.L. Godard, 1993, Je Vous Salue, Sarajevo, (`https://youtu.be/WKbfu8rRrho`)

Dans son court métrage *Je Vous Salue, Sarajevo*, réalisé en 1993 pendant la Guerre de Bosnie-Herzégovine, J-L. Godard déconstruit une photographie du reporteur de guerre R. Haviv. Il fragmente cette image pour faire se correspondre des inserts éclatés à la manière d'un collage-poème ou d'un cinétract [3]. Par le collage, les fondus et les découpes Godard rompt la continuité de l'archive qu'il utilise comme source première. Il peut ainsi rendre compte, image après image, de la cruauté qui frappe les rues Sarajevo. Le film finit par dévoiler entière, l'image dans toute son horreur, **décomposer** pour mieux **recomposer**.

3. Mini-films non signés à caractère militant, réalisés en mai et juin 1968 (`https://fr.wikipedia.org/wiki/Cinétract`) et influencés notamment par C. Marker.

Il y a dans les travaux de Godard et de Marker une souplesse d'action vis-à-vis des archives que nous pourrions appliquer à nos propres corpus. 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 carcan de leurs pages Web respectives ? Peut-on ralentir ou accélérer le cours de nos archives ?

Les strates 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** [4].

4. En anglais : *analytical Web strata.*

Brügger suggère de construire un système d'analyse dynamique pour réajuster, au besoin, 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.) [5]. Cette approche, notons-le, n'est pas confinée au Web archivé, elle peut très bien s'adapter au Web vivant. Brügger définit ainsi 5 niveaux d'analyses, allant du plus englobant au plus élémentaire, comme l'illustre la Figure 4.3.

5. « *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* », (Brügger, 2009, p.10)

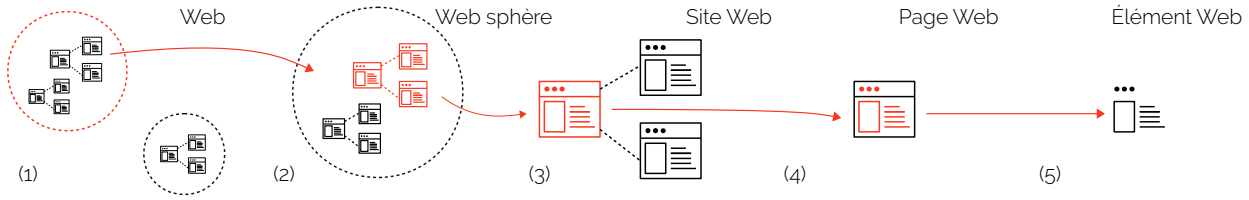


Figure 4.3: Les 5 strates analytiques du Web, d'après (Brügger, 2009)

Le premier niveau englobe l'entièreté des sites du Web vivant (Figure 4.3, (1)). 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 réseaux, supports numériques, etc.). Une sphère Web désigne un ensemble de sites Web sélectionnés par un chercheur (Figure 4.3, (2)). C'est une construction ad hoc motivée par une question de recherche donnée, une thématique précise [6]. 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 1.3) peuvent être considérés comme des sphères Web. Sites et pages Web (Figure 4.3, (3-4)) sont ensuite définis de manière identique à ce que nous proposons en Section 1.1. L'élément Web, quant à lui, est considéré comme l'élément textuel minimal d'une page Web (Figure 4.3, (5)). La définition de Brügger reste néanmoins assez floue sur ce dernier point, la nature d'un élément Web n'étant pas énoncée de manière non ambiguë : ce peut être un ensemble de caractères écrits sur une page, des images fixes ou mobiles, ainsi que des sons. Mais nous pouvons aussi penser que l'élément Web minimal puisse être une balise HTML vide (Section 4.3). Brügger en revanche écarte de cette liste les menus, barres d'informations et autres éléments de navigations.

6. La notion de sphère Web est inspirée des travaux de K. Foot sur le volet numérique des campagnes électorales états-uniennes du début des années 2000 (Foot et Schneider, 2006).

De notre côté, nous voulons penser le futur fragment Web comme un

sous-ensemble cohérent et auto-suffisant d'une page Web. Il s'inscrira dans la continuité des strates du Web, en se situant quelque part entre l'élément Web et la page Web. Un fragment pourra, en fonction des cas, être un élément Web seul, un groupe de plusieurs éléments, voire la page Web dans son entièreté[7].

Dès à présent, pour tout site Web composé de n pages Web $\{p_1,...,p_n\}$, nous supposons que chacune de ses pages p_j consiste en m fragments Web numérotés $\{f_{j1}, ..., f_{jm}\}$ (Figure 4.4).

7. Nous reviendrons dans le détail sur la question de l'étendue du fragment Web dans la Section 4.3.

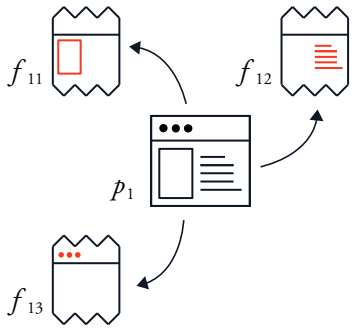


FIGURE 4.4: Une page p_1 et ses fragments Web f_{11}, f_{12}, f_{13}

Dater une page archivée

La datation des pages archivées peut être réévaluée à l'aune du fragment Web. Depuis la fin du précédent chapitre une question demeure : Comment tendre vers une plus grande précision historique ? Comment s'affranchir des seules dates de téléchargement ? Comment *bien* dater une page Web et son contenu ?

Les archives Web sont les traces directes des crawlers (Section 3.1). En DAFF ou en WARC, une page archivée sera toujours adressée par sa seule et unique date de téléchargement. Dans la plupart des moteurs d'exploration (par exemple la WayBack Machine, Figure 4.5), cette date est l'unique dimension temporelle interrogeable. Il est néanmoins possible d'établir une échelle de datation plus complète en introduisant la notion de date de dernière modification (Section 3.2 et Table 3.1). Échelle dont nous pensons maintenant pouvoir à nouveau améliorer la précision, en associant aux futurs fragments Web une **date d'édition**.

En effet, une page Web évolue (Section 2.2) dès que son contenu est édité par un tiers : humain ou robot. Par édition, nous entendons ici 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

FIGURE 4.5: Répartition des archives de *yabiladi.com* dans la WayBack Machine (`https://web.archive.org/web/*/www.yabiladi.com`)

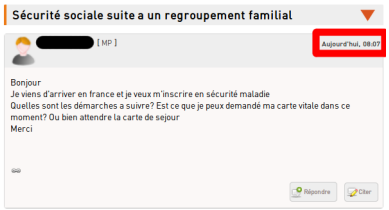


FIGURE 4.6: Date d'édition (rouge) d'un post de forum sur *yabiladi.com*

d'exploration.

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 4.6), reste alors à les interpréter et à les formater avant indexation (De Jong *et al.*, 2005; Kanhabua et Nørvaag, 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 URI ont été par ailleurs collectées [8] (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 et Kitsuregawa, 2006; Spitz *et al.*, 2018). Quoi qu'il en soit, l'identification et l'extraction d'une telle date d'édition reste possible et nous nous y emploierons en Section 4.3.

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

Mais, dès à présent, faisons par avance l'hypothèse d'être en capacité de doter chaque fragment Web d'une date d'édition. Ainsi, à tout fragment $\{f_{j1}, ..., f_{jm}\}$ d'une page p_j nous associons maintenant une date d'édition $\phi(f_{j1}), ..., \phi(f_{jm})$. De plus, nous nommons **date de création** de tout la page p_j la plus ancienne date d'édition de l'ensemble de ses fragments soit $\min_k \phi(f_{jk})$. La Figure 4.7 décrit l'imbrication de ces nouvelles datations.

Figure 4.7: Dates d'édition des fragments Web $\{f_{11}, f_{12}\}$ (rouges) et date de création de la page p_1 (noir)

Par ailleurs, nous faisons ici l'hypothèse implicite que toute datation extraite du contenu d'une page Web n'est ni erronée ni falsifiée. Dans le cas contraire, il existerait une possibilité non nulle d'avoir une date de création antérieure à la date de dernière modification. Ce faisant, nous pouvons maintenant mettre à jour notre échelle de datation en y ajoutant dates de création et d'édition, telles que :

Niveau	Nature de la date
page	lancement du crawl
page	téléchargement
page	dernière modification
page	création
fragment	édition

précision historique ↓

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

En pratique, tout fragment Web devra être associé à une date d'édition. Mais si cela est impossible, sa datation sera rapportée à la date de création de la page Web à laquelle il appartient. 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 [9]. Serait-il alors possible, comme le suggère l'historien J. Morsel, d'écrire une histoire *symptomale* [10] (Morsel, 2016) à partir de nos corpus d'archives Web ? Cela reviendrait à considérer que certains fragments Web se trouvent chargés de la présence latente d'un auteur, dissimulée sous la surface des pages archivées et prête à être questionnée.

Cette nouvelle perspective d'exploration nous mènera 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 5. Avec le fragment Web, une nouvelle dimension d'analyse des archives s'offre donc à nous : l'exploration par acteur (auteur, contributeur, commentateur, etc.), plutôt que la simple exploration par page ou par site.

9. 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).

10. Alors que la trace, telle que nous la décrivions jusqu'ici (Section 3.1), 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).

Désagréger pour changer de temporalité

Le Web est un flux grandissant d'information tout autant qu'un territoire en perpétuelle expansion. Par l'action des crawlers, les archives Web sont arrachées à la temporalité continue du Web vivant pour rejoindre celle figée et discrétisée des corpus collectés (Section 3.1). Les archives, malheureusement, ne peuvent revenir au temps du Web vivant, mais, grâce au fragment Web, nous pouvons les faire basculer dans la temporalité du Web tel qu'il a été. Temporalité que nous allons essayer de caractériser ci-dessous.

Commençons par une expérience. Au cours de la Section 3.4, nous

avions visualisé, dans le temps, la répartition des pages archivées de la section forum du site *yabiladi.com*, et ce, par rapport à leurs seules dates de téléchargement (Figure 3.16). Maintenant, essayons plutôt de nous focaliser sur les dates d'édition des fragments Web de chacune de ces pages et tentons une comparaison.

Faute de ne pas avoir encore défini clairement la nature d'un fragment Web, nous nous contenterons ici de l'approximation suivante : chaque *post* (un message individuel) publié sur le forum de *yabiladi.com* sera considéré comme fragment de la page dont il dépend. Un post est écrit par un unique auteur (identifié comme tel) et associé à une date d'édition (Figure 4.6).

11. Cette extraction de données n'est donc pas générique, voir la Section 4.3 pour une approche plus générale.

D'un point de vue pratique, nous procédons à une extraction focalisée [11] dans nos archives, afin de ne conserver que les dates d'édition des posts collectés. Ces dernières peuvent être facilement identifiées dans le code des pages Web par un nœud HTML unique, par exemple :

```
<div class="com-date">17 Novembre 2009</div>
```

Aussi, nous modifions notre moteur d'exploration au niveau de la jointure entre données et méta-données DAFF (Figure 3.7) et ajoutons un traitement permettant d'extraire les dates d'édition.

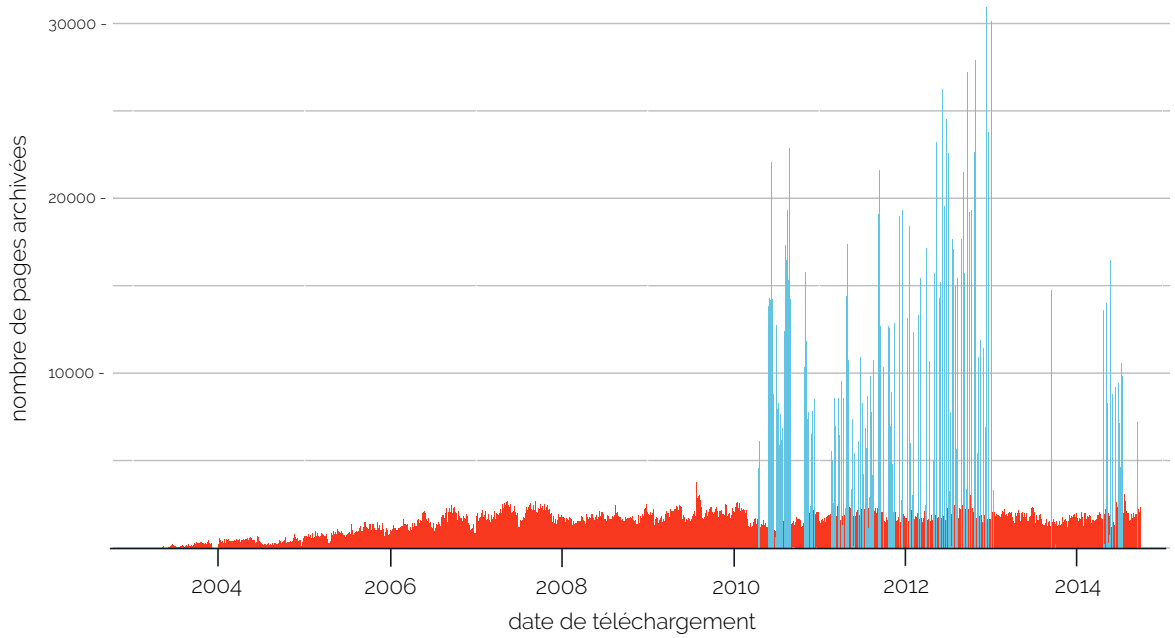


Figure 4.8: Distribution, pour *yabiladi.com*, du nombre de pages et de fragments archivés par jours et suivant leurs dates de téléchargement (bleu) et d'édition (rouge) respectives

Nous construisons ensuite un index dédié dans Solr. Ne reste plus alors qu'à visualiser les deux distributions côte à côte : d'une part la

répartition des pages par date de téléchargement (Figure 4.8, bleu) et d'autre part la répartition des fragments correspondants par date d'édition (Figure 4.8, rouge). Cette dernière (rouge) semble *gommer* l'artéfact de crawl précédemment observé sur l'année 2013 (bleu). La distribution des fragments est homogène et ne souffre d'aucune cécité remarquable. Détachée de l'influence du crawler, elle n'en subit plus les effets.

Par ailleurs, nos archives Web semblent chargées d'une mémoire plus étendue que celle initialement prévue. Ainsi, partant d'une collecte débutée en Mars 2010, nous voilà maintenant capables de considérer et d'analyser des fragments Web édités 7 années plus tôt, jusqu'en 2003 pour les plus anciens. Les pages archivées contiennent, en elles-mêmes, les traces sédimentaires de publications antérieures.

Par l'extraction et l'étude des fragments Web, nous nous donnons les moyens d'un saut dans le passé considérable. Ces fragments sont potentiellement porteurs d'une mémoire préexistante à chaque collecte et dont nous pouvons dater avec précision l'apparition. En désagrégeant les archives Web, en les fragmentant, nous changeons une nouvelle fois de temporalité pour entrer dans **le temps du Web tel qu'il a été**. Là ou le temps des archives était figé et fait de séries de captures discrètes d'une même page, le temps du Web tel qu'il a été est un temps fragmenté, c'est-à-dire un temps éclaté, où chaque fragment se voit défini relativement par rapport à lui même.

Nos expérimentations pratiques ne portent que sur les seules dates d'édition. Mais, dans le temps du Web tel qu'il a été, chaque fragment Web suit sa propre temporalité, détachée de celle des autres. Une ligne allant de son apparition sur le Web (date d'édition) jusqu'à sa possible disparition de la toile. Isolées les unes des autres, c'est à l'explorateur d'archives que revient le rôle de naviguer entre ces lignes de temps éclatées. L'explorateur sélectionne, découpe et assemble des fragments pour construire ce que l'anthropologue T. Ingold nomme un **trajet**, support de nos explorations à venir :

« *Dans le cas du trajet, en revanche, on s'engage dans une voie qu'on a déjà explorée avec d'autres, ou qui a été explorée par d'autres, en reconstruisant l'itinéraire au fur et à mesure de sa progression.* » — (Ingold et Renaut, 2013, p.26)

Un trajet est fait de détours, de contours et de bifurcations. À mesure qu'il se conçoit, le trajet se développe et s'inscrit dans le temps. Suivant le cours de son analyse, c'est par le montage que le chercheur chemine d'un fragment Web à l'autre, dans le sens et l'ordre qu'il juge pertinent. En conjuguant les multiples lignes de temps il s'affranchit ainsi des formes classiques et linéaires d'accès aux archives Web. Ouvrant la voie à de nouveaux degrés de liberté, les fragments pourront être

associés sur la base d'un lien hypertexte partagé, d'une présence sur la même page à un instant donné, d'une filiation commune, etc. Ces trajets entre fragments deviennent sous la plume de J. Bashet des **lignes processuelles** (Baschet, 2018, p.227). L'historien cherche, ce faisant, à rompre avec une vision linéaire de l'histoire dont il faudrait faire éclater la continuité :

« *En effet, il ne s'agit en aucun cas de penser l'Histoire tout entière comme un seul processus unifié, mais de saisir, dans l'histoire, un entrelacement complexe de multiples processus.* » — (Baschet, 2018, p.227)

En suivant le devenir historique de multiples lignes processuelles, l'écriture de l'histoire revient à raisonner autour de **moments** singuliers où convergent et se croisent temporalités et processus hétérogènes :

« *Et on proposera plutôt d'explorer diverses manières de penser l'événement - le surgissement, le nouveau, la rupture mais aussi l'imprévu, l'imprévisible, l'improbable - à partir d'une pensée des processus. Ainsi, outre qu'elle peut naître ou disparaître, une ligne processuelle connaît par elle-même des variations de rythme et des moments singuliers de concentration ou d'expansion des forces à l'œuvre : l'événement tient alors à une étape particulière de maturation ou correspond, peut-être à un seuil d'ébullition ou de cristallisation.* » — (Baschet, 2018, p.227-228)

Avec le passage de la page au fragment, nous basculons d'une unité d'exploration à l'autre. Le fragment Web nous invite à un changement d'échelle temporelle et spatiale dans le rapport que nous entretenons aux archives Web. Situé entre la page et l'élément Web, le fragment peut contenir en lui la trace du Web tel qu'il a été : une mémoire jusqu'ici retenue dans les fichiers archivés. Le chercheur associe alors un à un les fragments qu'il juge pertinents et conduit, chemin faisant, son exploration pour saisir l'histoire du Web et ses cristallisations autour de moments singuliers.

Dans notre méthodologie, la place du chercheur est donc centrale. C'est lui qui, par ses choix de montage (basés sur sa propre expertise ou sur des indices qu'il aura recueilli en amont) définit les fragments Web à explorer et la manière de les parcourir. Il peut, dans cette tâche, se faire aider de scripts informatiques pour automatiser certains traitements. Le fragment Web doit ainsi être interprétable par une machine : un programme pourra l'analyser, le manipuler, le stocker, etc. Mais le fragment doit aussi rester compréhensible, en lui même, afin d'être étudié par un chercheur (sociologue, historien...). Nous discuterons, en Section 4.3, de l'implication ou non du chercheur dans le choix même de la forme des fragments Web. Nous donnerons, enfin, dans le Chapitre 5, deux exemples d'explorations désagrégées de nos corpus et basées sur le fragment Web.

4.2 Le fragment Web : définition

La définition suivante est intentionnellement générique. Nous souhaitons par là que d'autres chercheurs puissent se saisir après nous du fragment Web [12]. Par ailleurs, la nature des fragments dépendant beaucoup du contexte de l'analyse et de la sensibilité propre à chaque chercheur, soit qu'il voudra une fragmentation plus ou moins englobante, soit qu'il se satisfera d'éléments abstrait, nous ne donnerons pas ici de définition technique précise du fragment. Nous proposerons, dans la Section 4.3, notre propre système d'extraction des fragments Web depuis une page archivée, d'autres approches et stratégies peuvent naturellement exister.

12. Notons que le fragment Web ne doit pas être confondu avec la notion de fragments d'URL (`https://en.wikipedia.org/wiki/Fragment_identifier`).

Considérant la page web comme unité de consultation de base du World Wide Web, bâtie sur des modalités d'écriture propre au support numérique et constatant que du point de vue de la perception humaine (Bernard, 2003; Michailidou et al.*, 2008) une page web est le résultat de l'agencement logique d'éléments sémantiques distincts, nous nommons* ***fragment Web*** *un sous-ensemble sémantique et syntaxique d'une page Web donnée et vérifiant les propriétés suivantes :*

1. *Il y a une relation d'échelle entre une page Web et ses fragments Web. Ceux-ci peuvent couvrir l'entièreté de la page ou n'être qu'un élément unitaire de cette dernière.*
2. *Un fragment Web est un assemblage cohérent d'éléments textuels, visuels, sonores ou logiciels extraits d'une page Web. Le fragment Web doit ainsi être compréhensible par lui même.*
3. *Au sein d'une même page Web, deux fragments Web ne peuvent pas se superposer, même partiellement.*
4. *Certains éléments d'un fragment Web peuvent faire l'objet d'une catégorisation lors de l'extraction. Un fragment Web peut ainsi être associé à un titre, à un auteur, à une date d'édition, etc.*
5. *Le fragment Web capture l'ensemble des dispositifs d'écriture (nœuds HTML, widgets de CMS, éditeurs de texte...) et de partage (liens hypertextes, liens de syndication, liens de publications...) utilisés pour publier son contenu sur le Web.*

4.3 Scraping et méthodologie d'extraction

Extraire de l'information issue d'une page Web

On appelle **scraping** l'ensemble des techniques et méthodes employées pour extraire de l'information depuis une page Web. Un **scraper** est, de

fait, un robot chargé de faire du scraping [13]. En pratique, un scraper ne travaille jamais au hasard, il est focalisé, c'est-à-dire qu'il est paramétré pour ne conserver que certaines parties ou éléments distincts d'une page : des noms, des adresses, des images, des mots clés, etc. La recherche de nos fragments Web passera obligatoirement par une étape d'extraction, il faudra alors scraper l'ensemble des pages archivées.

13. Par abus de langage et faute d'une meilleure traduction, on dira d'un scraper qu'il **scrape** une page Web.

De prime abord, les scrapers ne considèrent pas les pages Web telles que nous les voyons depuis nos écrans, interprétées par les navigateurs Web. Lors d'une extraction, les scrapers parcourent, d'abord et avant tout, des portions de code issues des fichiers HTML et CSS qui forment l'ossature de ces pages.

Le **HTML** [14] est un langage de programmation décrivant, à la fois, la structure et le contenu d'une page Web. Le HTML est la grammaire de l'hypertexte. Écrites en HTML, les pages sont alors *rendues* par le navigateur Web qui transcrit visuellement les instructions codées.

14. *HyperText Markup Language*

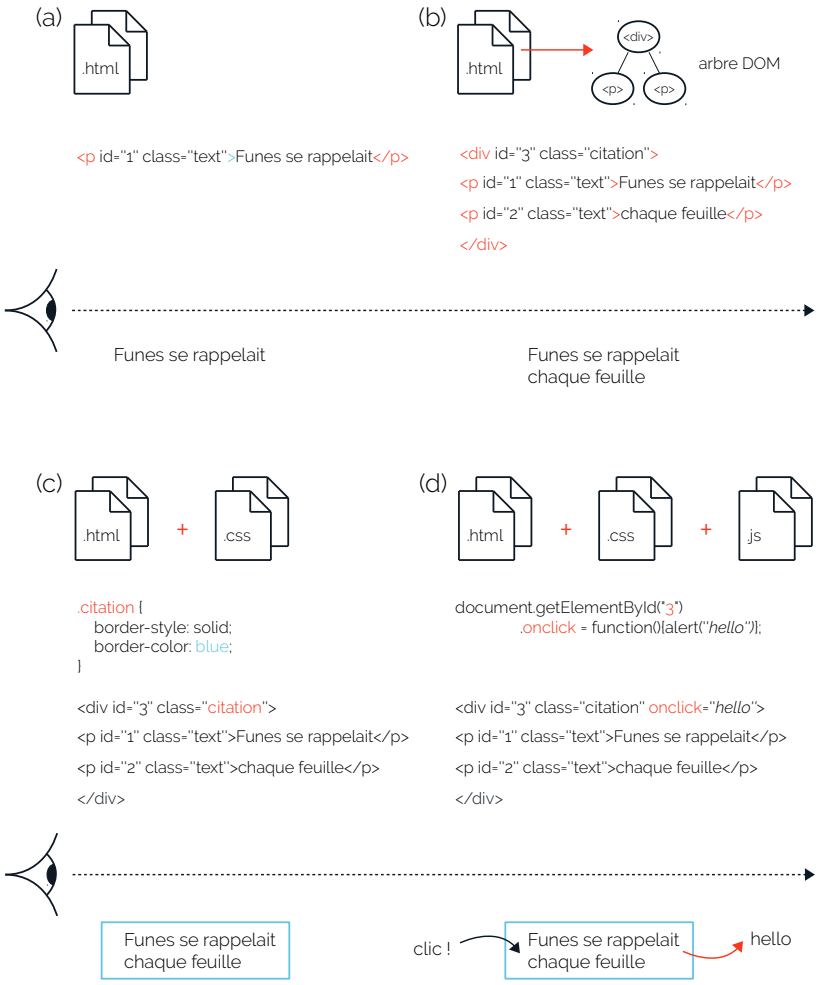


FIGURE 4.9: Ajout successif d'éléments HTML, CSS et JavaScript à une page Web et transcription sur l'écran d'un internaute

Le HTML est un langage à **balise**, c'est-à-dire que chaque élément d'une page Web est délimité par une balise ouvrante à une extrémité (ex : `<p>`) et fermante à l'autre bout (ex : `</p>`). La nature de ces balises, leur syntaxe et leur agencement permettent d'enrichir le contenu textuel de l'élément qu'elles définissent (Figure 4.9, (a)). Une balise HTML est identifiée par un nom (*tag*) qui donne une indication sur le type de l'élément caractérisé (du texte `<p>`, un lien `<link>`...). Elle peut être complétée (entre autres) par un *id* (un identifiant unique) et une ou plusieurs *class* (attribut qualifiant un ou plusieurs éléments). Classes et identifiants peuvent servir à associer un comportement spécifique à un ou plusieurs éléments cibles.

Une page Web est un document structuré (Figure 4.9, (b)). Les éléments HTML s'agencent entre eux suivant la forme particulière d'un arbre : l'**arbre DOM** [15]. Par convention, on appelle **nœud** HTML tout élément d'un arbre DOM. Il existe ainsi un seul et unique nœud racine, plusieurs nœuds parents, enfants, etc. Un scraper peut accéder à un nœud HTML donné soit en l'adressant directement via son id ou sa classe, soit en parcourant l'arbre DOM.

15. *Document Object Model tree*, en anglais. En informatique un arbre est une structure de données où chaque élément (nœud) est codé hiérarchiquement par rapport aux autres (`https://fr.wikipedia.org/wiki/Arbre_binaire`)

Le **CSS** [16] est un langage pensé pour décrire l'aspect visuel (rendu et animation) d'un nœud HTML à l'écran. Le CSS transcrit ainsi des règles de style directement depuis le HTML de la page Web ou dans un fichier dédié. Une correspondance est alors faite entre l'identifiant (et/ou la classe) d'un nœud et la règle à lui appliquer (Figure 4.9, (c)). On jouera ainsi sur la couleur, la police d'écriture, la marge, les bordures, etc., de chaque élément.

16. *Cascading Style Sheets*, en anglais

Les éléments dynamiques d'une page Web sont générés, majoritairement, par du code **JavaScript** (Figure 4.9, (d)). Le JavaScript peut s'écrire dans un ou plusieurs fichiers séparés ou être directement ajouté au HTML, au sein de balises dédiées. Il est alors possible de manipuler des nœuds HTML et de leur attribuer à chacun comportement. En réaction au geste d'un internaute, par exemple : un clic, un défilement, etc. D'autres technologies comme Adobe Flash [17] ont, par le passé, également su gérer ces aspects dynamiques. Au final, HTML, CSS et JavaScript forment le triptyque le plus courant face auquel doivent se débattre les scrapers.

17. `https://fr.wikipedia.org/wiki/Adobe_Flash`

En effet, un scraper doit savoir s'adapter à la complexité des pages qu'il parcourt. Par exemple, une page contenant du JavaScript devra parfois être préalablement interprétée par le scraper (et non juste parcourue), afin de prendre en compte les éléments dynamiques qui ne se chargeraient qu'à l'affichage écran [18]. Une campagne de scraping se prépare donc en amont, il s'agira alors de définir une stratégie adaptée au contexte de notre extraction. Le scraper peut, ainsi, s'attarder sur les éléments visuels d'une page Web (Cai *et al.*, 2003), ou se contenter de

18. Voir les bibliothèques Phantomjs (`http://phantomjs.org/`) et After-load (`https://www.npmjs.com/package/after-load`).

la seule information sémantique présente dans le HTML (Jatowt *et al.*, 2007). Là où la première solution sera en quête d'une vision humaine des pages Web, la seconde privilégiera la recherche d'un temps de réponse acceptable.

La vitesse des traitements est, de fait, une composante essentielle de tout scraping, notamment dans ses applications industrielles. Peut-on se permettre d'utiliser des méthodes coûteuses pour rendre certaines images ou interpréter du code ? Ce faisant, les travaux portants sur l'extraction d'information depuis une page Web sont souvent pensés pour s'intégrer, d'abord et avant tout, à de vastes campagnes d'analyses (Weninger et Hsu, 2008; Adar *et al.*, 2009; Oita et Senellart, 2015). La validation humaine n'est pas la finalité de ces recherches. De notre côté, nous préférons nous rapprocher de méthodes plus qualitatives, conçues sciemment pour être utilisées, en premier lieu, par des êtres humains.

L'extension Readability [19] du navigateur Web Firefox propose, par exemple, aux internautes d'expérimenter une forme de lecture *zen* sur la toile. Le système cherche ainsi à identifier le contenu principal de chaque page (le corps d'un article), à l'extraire et à le présenter dépouillé des publicités, menus et autres suggestions qui pourraient nuire à la lecture. Bien que Readability se limite seulement à certains types de sites (les sites de news notamment), une partie de son fonctionnement pourra être adaptée à notre propre moteur.

19. Voir `https://github.com/mozilla/readability`, Mercury Reader pour Chrome ou encore l'application Pocket.

Des pages bruitées à nettoyer

Notre tâche consiste ici à fragmenter une page Web donnée, c'est-à-dire (suivant les termes nouvellement introduits), à extraire de nos fichiers archivés, des sous-ensembles cohérents de nœuds HTML. Pour ce faire nous prendrons chaque page une par une, nous la nettoierons, puis nous définirons une mesure censée traduire la cohérence entre des nœuds HTML deux à deux, avant de les grouper en un ensemble de fragments Web distincts. Soit une page Web p composée de m nœuds HTML $\{n_1, ..., n_m\}$, organisés en un arbre DOM t et associés à des règles CSS.

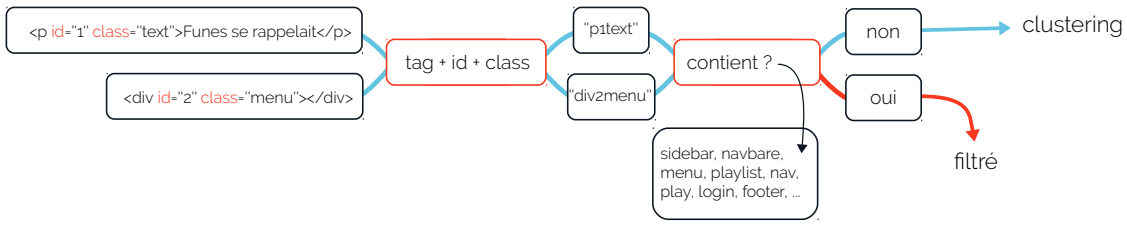


Figure 4.10: Processus de nettoyage, nœud par nœud, d'une page Web

Nous commençons par *nettoyer*[20] p de tous les nœuds qui ne nous semblent pas pertinents : publicités, menus de navigation, scripts... D'expérience, nous faisons l'hypothèse que le Web est une source de données très bruitée, tout l'enjeu est ici de trouver un juste milieu entre nettoyage et conservation des données pertinentes. Des méthodes très efficaces, par apprentissage supervisé, existent (Kohlschütter *et al.*, 2010), cherchant à définir des masques de nettoyage pour chaque site. Or, ce genre de technique demande la création en amont d'un ensemble d'apprentissage annoté. Opération qui pourrait s'avérer rapidement fastidieuse si nous l'appliquions aux différentes versions de pages collectées par nos archives.

20. Voir la section suivante pour une discussion sur le nettoyage des pages

comme nous ne connaissons pas, à priori, les évolutions structurelles et stylistiques subies par un site Web au cours de son histoire, la définition d'un ensemble d'apprentissage nous semble ici compromise.

De notre côté, nous préférons favoriser une approche par heuristiques (Jatowt *et al.*, 2007). Ainsi, nous définissons pour chaque nœud de p un label, résultat de la concaténation de son tag, de son id et de sa classe :

$$label = tag + id + classe$$

L'idée est ici de vérifier la valeur informative d'un nœud via un ensemble d'expressions régulières appliquées à son label (Figure 4.10). Pour ce faire, nous parcourons l'arbre DOM t et si l'un des labels contient les termes *menu* ou *navbar*, par exemple[21], alors celui-ci ne sera pas conservé dans la suite du processus d'extraction. À la fin de cette étape nous ne retenons, à titre illustratif, que 30% des nœuds des pages forums de *yabiladi.com*, ce taux descend autour de 15% pour les pages *actualités*, plus bruitées donc.

21. Voir la liste complète dans le code source (`https://github.com/lobbeque/rivelaine/blob/master/scala/src/main/scala/qlobbe/Rivelaine.scala`, l. 72-73).

Classification de nœuds HTML

Transposé au champ d'action de la recherche d'information, un fragment Web peut être vu comme un groupe cohérent de nœuds HTML. Toute la difficulté revient à traduire informatique cette notion de cohérence entre nœuds. Sur ce point, la librairie open-source Fathom[22] a été conçue pour identifier des *clusters* (groupes) de nœuds HTML au sein d'une page Web donnée. L'idée de Fathom est simple : deux nœuds proches l'un de l'autre sont placés dans un même groupe, un nœud proche d'un groupe déjà formé le rejoindra alors, et ainsi de suite. L'algorithme réalise un clustering par agglomérations successives de l'ensemble des nœuds de l'arbre DOM t.

22. `https://github.com/mozilla/fathom`

Dans la Figure 4.11, la page p contient cinq nœuds HTML identifiés par les tags de leurs balises : A, B, C, D, E. Deux nœuds n_1, n_2 sont considérés proches l'un de l'autre si la distance $d(n_1, n_2)$ les séparant est

inférieure strictement à une variable *minDist* : la distance minimale autorisant le rapprochement de deux nœuds (ici valant arbitrairement 4).

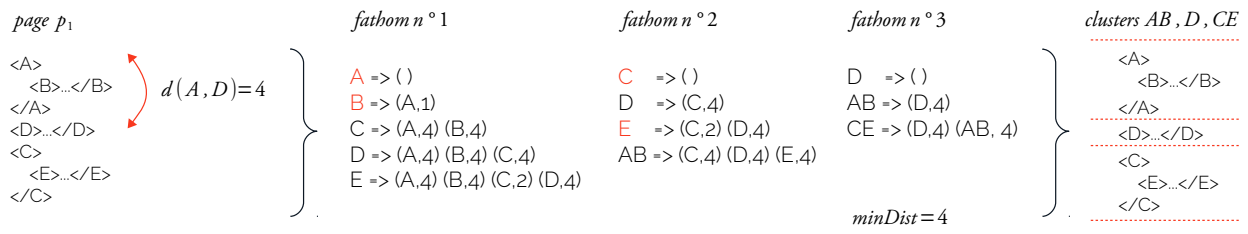


FIGURE 4.11: Processus de classification des nœuds d'une page Web

Dans cet exemple [23], la distance $d(A, D)$ vaut 4, A et D ne sont donc pas considérés comme proches. L'ensemble des m nœuds de p_1 sont codés, dans Fathom, sous la forme d'une matrice d'adjacence $m \times m$ nommée M (ici de taille 5). À chaque ligne (*row*) de M correspond un nœud et sa distance respective vis-à-vis des autres nœuds de p. À chaque itération de Fathom, une fonction *closestRows* détermine les deux nœuds ou groupes de nœuds les plus proches. Dans notre exemple, au premier passage ce sont A et B qui sont considérés les plus proches avec $d(A, B) = 1$, puis vient le tour de C et E situés à une distance 2 l'un de l'autre. À la troisième itération plus aucun groupe ne peut être aggloméré, Fathom a donc identifié trois clusters distincts de nœuds. Le pseudo-code, ci dessous, décrit avec plus de détail le fonctionnement de Fathom [24] :

23. Les distances sont ici données à titre illustratif, voir la section suivante pour une discussion sur les fonctions de distance

24. Implémentation originale disponible ici : `https://github.com/mozilla/fathom/blob/master/clusters.mjs#L161`

```
while rows(M) > 1 and dist(closestRows(M)) < minDist do
    {r_i, r_j} = closestRows(M)
    newRow = {}
    for r ∈ rows(M) do
        if r ≠ r_i and r ≠ r_j then
            newRow[r] = min(d(r_i, r), d(r_j, r))
        end
    end
    remove(M[r_i])
    remove(M[r_j])
    remove(M[*][r_i])
    remove(M[*][r_j])
    append(M, newRow)
end
```

Fathom a besoin d'être paramétré pour fonctionner correctement. Il faut ainsi établir en amont la valeur de *minDist* et définir une fonction de

distance qui fasse sens. Dans notre cas nous voulons que cette dernière traduise la cohérence sémantique, syntaxique et visuelle d'un fragment Web.

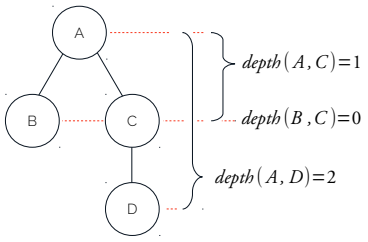


FIGURE 4.12: Différence de profondeur entre deux nœuds HTML au sein d'un même arbre DOM. $d(n_1, n_2) = 0$ s'ils sont frères. $d(n_1, n_2) = 1$ si l'un est le parent de l'autre

Définir une fonction de distance

Une **fonction de distance** est un objet mathématique qui traduit et calcule l'éloignement entre deux entités, données en variable. Dans notre cas, deux nœuds HTML n_1, n_2 seront considérés comme *proches* si le résultat de la fonction de distance $d(n_1, n_2)$ est strictement inférieure à la valeur de $minDist$. Ces nœuds pourront alors être groupés par Fathom et formeront un seul et même fragment Web.

Par défaut, Fathom réutilise la fonction de distance implémentée dans Readability et designée pour grouper ensemble les nœuds appartenant au contenu principal d'une page Web. Cette fonction s'appuie, d'une part, sur la différence de profondeur entre deux nœuds de l'arbre t (Figure 4.12, $depth(n_i, n_j)$) et, d'autre part, sur le nombre d'éléments les séparant une fois l'arbre t mis à plat ($length(n_i, n_j)$). Un malus est ensuite ajouté si les nœuds ne possèdent pas les mêmes balises HTML ($tag(n_i) \neq tag(n_j)$). Le pseudo code suivant décrit l'intégration de cette fonction de distance [25] à Fathom, lors de la création de la matrice d'adjacence M :

25. Voir l'implémentation originale : `https://github.com/mozilla/readability/blob/master/Readability.js#L760`

```
for n_i ∈ nodes do
    for n_j ∈ nodes do
        d(n_i, n_j) = 0
        d(n_i, n_j) = d(n_i, n_j) + depth(n_i, n_j)
        d(n_i, n_j) = d(n_i, n_j) + length(n_i, n_j)
        if tag(n_i) ≠ tag(n_j) then
            d(n_i, n_j) = d(n_i, n_j) + malus
        end
        return d(n_i, n_j)
    end
end
```

En l'état, cette fonction ne traduit qu'une forme de cohérence structurelle (*depth*, *length*), voire sémantique si l'on considère la différence de balises telle quelle. Notre définition du fragment Web se veut plus complète. Nous allons donc enrichir cette fonction de distance en partant d'un principe simple : par défaut deux nœuds sont considérés comme proches, tout motif d'éloignement sera sanctionné d'un malus.

Tout d'abord, la cohérence visuelle jouant un rôle important dans la segmentation d'une page Web donnée [26], nous éloignerons les nœuds dont la couleur d'arrière plan ($color(n_i) \neq color(n_j)$, issue du CSS) n'est

26. « *[...] constatant que du point de vue de la perception humaine (Bernard, 2003; Michailidou* et al.*, 2008) une page web est le résultat de l'agencement logique d'éléments sémantiques distincts* », Section 4.2

pas la même.

Puis, nous ajouterons un malus aux éléments séparés par des *breaking lines* ($vips(n_i, n_j)$). Dans l'algorithme de segmentation Vips (**?**), les breaking lines sont des nœuds HTML identifiés comme potentiels séparateurs de contenu sur le Web. Ce sont des sauts visuels entre deux portions de texte d'une même page. Les balises *<hr>*, *
*, etc., sont des breaking lines.

Enfin, nous savons grâce aux travaux de J. Goody (Goody *et al.*, 1979), que la liste, comme forme d'organisation graphique de l'écriture, est fortement utilisée par l'Homme depuis qu'il a ressenti le besoin de stocker et d'organiser ses données [27]. La liste abstrait par un jeu de discontinuités (retour à la ligne) et de continuités (récurrences) les données ainsi présentées, permettant un classement de ces dernières suivant de multiples critères.

27. Des listes de comptes des tablettes sumériennes aux listes de résultats de Google.

Les pages Web, dans leur organisation à l'écran, n'échappent pas à cette règle : listes d'articles, de commentaires, de suggestions, de liens... Une manière de traduire la cohérence d'un fragment Web serait d'identifier chaque fragment à un élément d'une de ces listes, quelque soit sa taille. Pour ce faire, nous définissons, par l'observation, des **masques de continuité** entre nœuds HTML. La Figure 4.13 illustre cette intuition.

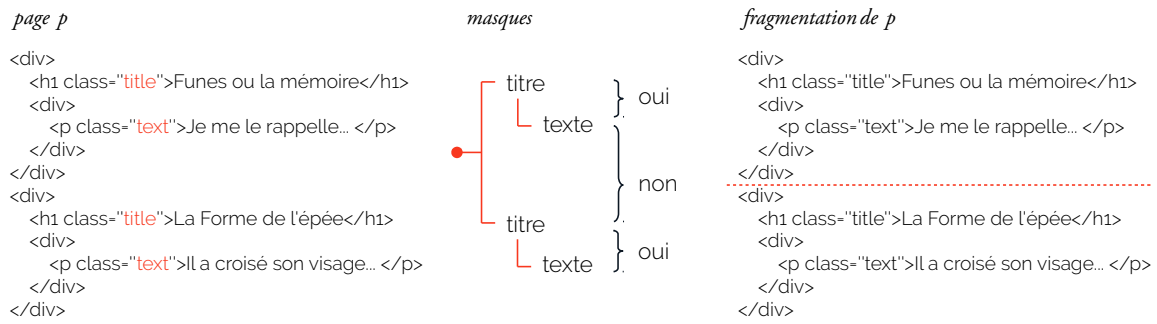


Figure 4.13: Ségmentation d'une page Web suivant des masques de continuité

La page p peut se voir comme une liste de nouvelles écrites par Borges, chacune présentée par un titre et un phrase faisant office de texte. Nos masques de continuité prennent en compte, d'une part, la nature des nœuds HTML (par une catégorisation de ces derniers [28]) et, d'autre part, une relation de hiérarchie entre ces mêmes nœuds : Le second est il plus profond que le premier ? Sont-ils au même niveau ? Nous faisons l'hypothèse, que dans l'arbre DOM t un nœud de profondeur moindre (le titre) subordonnera la réception d'un nœud plus éloigné de la racine (le texte). Ainsi, dans cette exemple nous définissons comme masque l'énoncé suivant : un nœud titre suivi d'un nœud texte plus

28. Voir la section suivante pour une discussion sur la catégorisation des nœuds HTML

profond. Si deux nœuds ne valident pas ce masque (ici un texte suivi d'un titre moins profond), un malus leur sera attribué dans la fonction de distance. Cela nous permet, au final, de fragmenter la page p comme une liste de deux nouvelles, de deux fragments Web.

Nous définissons ainsi plusieurs masques (Figure 4.14), supposés englober la majorité des cas de figure rencontrés sur le Web. Une condition ($incoherent(n_i, n_j)$) est ajoutée à notre fonction de distance, qui traduit maintenant une forme de cohérence structurelle, visuelle et syntactique entre deux nœuds HTML. Fathom est maintenant capable de fragmenter les pages Web archivées. Le pseudo-code [29] suivant présente l'agencement de l'ensemble de nos malus au sein de la fonction d :

29. Voir l'implémentation détaillée : `https://github.com/lobbeque/rivelaine/blob/master/nodejs/cluster.js#L41`

```
for n_i ∈ nodes do
    for n_j ∈ nodes do
        d(n_i, n_j) = 0
        d(n_i, n_j) = d(n_i, n_j) + depth(n_i, n_j)
        d(n_i, n_j) = d(n_i, n_j) + length(n_i, n_j)
        if tag(n_i) ≠ tag(n_j) then
            d(n_i, n_j) = d(n_i, n_j) + malus
        end
        if color(n_i) ≠ color(n_j) then
            d(n_i, n_j) = d(n_i, n_j) + malus
        end
        if vips(n_i, n_j) then
            d(n_i, n_j) = d(n_i, n_j) + malus
        end
        if incoherent(n_i, n_j) then
            d(n_i, n_j) = d(n_i, n_j) + malus
        end
        return d(n_i, n_j)
    end
end
```

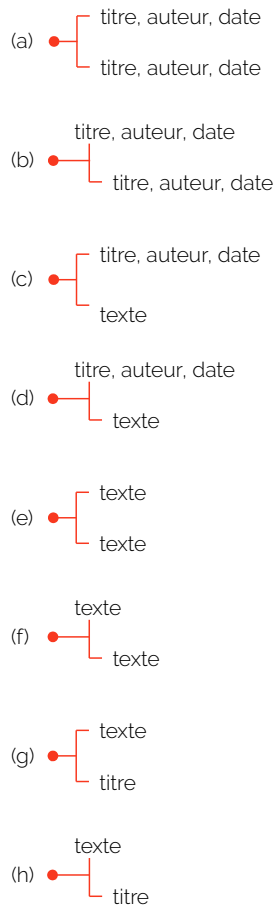


FIGURE 4.14: Masques de continuité entre deux types de nœuds HTML de profondeur variable

Catégorisation

Pour définir nos masques de cohérence, il nous faut catégoriser certains nœuds HTML par rapport à la nature de leur contenu. La catégorisation se fait ici au niveau des nœuds et non à l'intérieur du texte de l'élément HTML observé. Ces catégories interviennent, donc, au moment de la segmentation d'une page Web en fragments, mais elles deviendront aussi, à terme, dimension d'interrogation à part entière des archives Web. Pour trouver l'ensemble des fragments Web écrits par un seul et même auteur, il faut, en amont, avoir déterminé les nœuds HTML

susceptibles de nous renseigner sur l'identité de cette personne. Nous définissons ainsi six catégories de nœuds :

1. Les nœuds titres : titre d'une page, d'un article...
2. Les nœuds auteurs : auteur d'un post de blog, de forum...
3. Les nœuds dates : toute information temporelle
4. Les nœuds textes : tout élément textuel qui ne soit ni un auteur, ni une date, ni un titre
5. Les nœuds d'expression : nœuds permettant de partager, promouvoir, éditer du contenu en ligne (liens hypertextes, éditeurs intégrés, bouton partager...)
6. Les nœuds autres : ceux qui n'ont pas pu être catégorisés

La catégorisation en elle même se fait sur la base d'expressions régulières et d'heuristiques ad hoc. Nous cherchons, ainsi, à faire correspondre le label de chaque nœud à un ensemble de termes (balises, classe, etc.) caractéristiques d'une catégorie donnée. Par exemple, trouver le terme *title* associé à une balise `<h1>`[30] dans le label d'un nœud, nous permettra de le placer dans la famille des titres. Chaque label passe alors au crible de nos expressions régulières[31], allant de la plus discriminante à la plus englobante, comme l'illustre la Figure 4.15 :

30. Balise HTML définissant un titre

31. Voir le détail des patterns recherchés `https://github.com/lobbeque/rivelaine/blob/master/nodejs/utils.js`

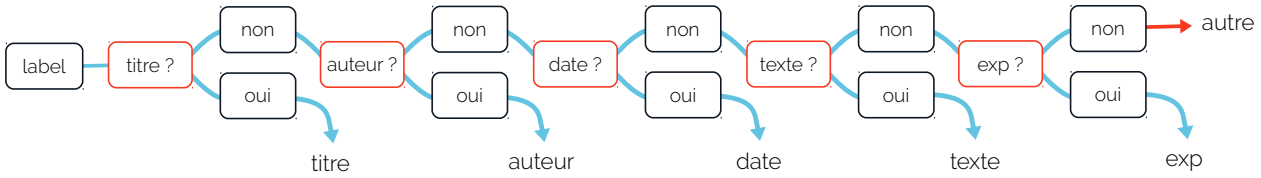


Figure 4.15: Chaine d'expressions régulières permettant la catégorisation d'un nœud HTML sur la base de son label

Au sortir de cette étape, nous souhaitons normaliser les nœuds dates. En effet, ces nœuds renferment les dates d'édition que nous cherchons à extraire depuis le début de ce manuscrit. Mais il y a ici une différence importante à faire entre une date encapsulée dans un nœud date et une date trouvée dans le corps d'un nœud texte ou titre.

La première peut être vue comme une information directement relative à la page et à son contenu. Elle caractérise par exemple une date de publication, d'écriture d'un commentaire ou de réponse à un article[32]. La seconde, en revanche, peut faire référence à n'importe quel aspect du contenu textuel des pages analysées : la date d'un événement narré dans un article, la date d'une référence bibliographique, etc. La structure même du HTML ne nous renseigne pas sur le contexte de ces dates. Pour nous, elles ne sont pas directement liées à un geste

32. Ces nœuds dates sont souvent crées automatiquement par les CMS et autres éditeurs de sites

d'édition en ligne. Ces dates pourront, par contre, faire l'objet d'une analyse ultérieure.

Ainsi, nos nœuds dates passent finalement au travers d'une fonction de normalisation, qui, sur la base d'un dernier jeu d'expressions régulières [33], transforme une date écrite en langage naturel en une date formatée [34] et compréhensible par notre moteur de recherche Solr. Les dates générées sans horaires sont, par défaut, remmenées à minuit du jour indiqué. Les dates peut précises du type *hier, 08 :30* ou *il y a 15 minutes*, ne sont pas prises en compte [35].

33. Voir le détail des patterns : `https://github.com/lobbeque/rivelaine/blob/master/scala/src/main/scala/qlobbe/Patterns.scala`
34. Suivant le standard ISO 8601
35. Une amélioration serait ici de se référer à la date de création ou à la date de téléchargement de la page.

Discussions

Depuis le début de cette section, nous suivons le parcours d'une page Web p d'une étape à l'autre de son processus de segmentation. La page est d'abord nettoyée, pour ne conserver que les nœuds à haute valeur informative. Les éléments HTML restants sont ensuite catégorisés suivant leur nature propre. Grâce à l'algorithme de classification Fathom, associé à une fonction de distance enrichie, nous groupons les nœuds entre eux et segmentons alors p en une suite de fragments Web distincts.

Nous appelons **Rivelaine** [36] notre librairie d'extraction des fragments Web. Une première implémentation est développée en Scala et retourne les fragments suivant le format [37] JSON. À titre d'exemple, nous demandons à rivelaine de fragmenter une page issue du forum de *yabidali.com*. Parmi l'ensemble des fragments identifiés, nous nous attardons sur ceux semblant se rapprocher de messages postés par des membres du site, tels que :

36. Open-source et téléchargeable ici-même : `https://github.com/lobbeque/rivelaine`
37. *JavaScript Object Notation*, format de données textuelles permettant de représenter des informations structurées (`https://fr.wikipedia.org/wiki/JavaScript_Object_Notation`)

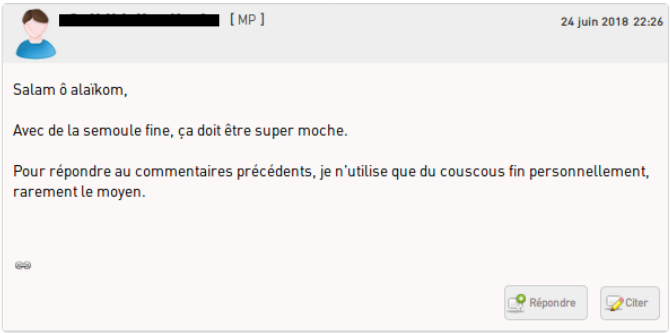


FIGURE 4.16: Exemple de message posté sur `https://www.yabiladi.com/forum/semoule-fine-pour-couscous-54-9290786.html`

```
{
    "category": [
        "author","date","text","exp","exp"
    ],
    "author": [
        [redacted]
    ],
    "date": [
        "24 juin 2018 22:26"
    ],
    "dateClean": [
        "2018-06-24T22:26:00Z"
    ],
    "exp": [
        "Répondre","Citer"
    ],
    "href": [
        "/profil/751194/[redacted].html","/mp/send/9291087/message","//www.yabiladi.com/forum/semoule-fine-pour-
        couscous-54-9290786-9291087.html#msg-9291087","javascript:///","//www.yabiladi.com/forum/semoule-fine-pour-
        couscous-54-9290786-9291087-quote=1.html#REPLY"
    ],
    "ratio": 0.008675232394023458,
    "node": [
        "<div class="com-auteur-1"> <a href="/profil/751194/[redacted].html"><strong>[redacted]</strong></a>
        <small>[ <a rel="nofollow" href="/mp/send/9291087/message">MP</a> ]</small><div class="com-date  "><strong>24 juin
        2018 22:26</strong></div></div>","<div class="com-content-1 fc"> Salam ô alaïkom, <br> <br> Avec de la semoule
        fine, ça doit être super moche. <br> <br> Pour répondre au commentaires précédents, je n'utilise que du couscous
        fin personnellement, rarement le   moyen.<div class="x2"><div class="message-moderation"></div><div class="com-
        footer-1"> <span id="permalink"> <a rel="nofollow" href="//www.yabiladi.com/forum/semoule-fine-pour-couscous-54-
        9290786-9291087.html#msg-9291087"> <img src="/images_new/1490800596_link.png"> </a> </span><div
        class="msg_actions"> <a class="iconn repondre1" rel="nofollow" href="javascript:// " onclick="$j('html,
        body').animate({ scrollTop: $j('#REPLY').offset().top }, 800);"><span>Répondre</span></a> <a class="iconn citer-1"
        rel="nofollow" href="//www.yabiladi.com/forum/semoule-fine-pour-couscous-54-9290786-9291087-
        quote=1.html#REPLY"><span>Citer</span></a></div></div></div></div>"
    ],
    "label": [
        "DIV  com-auteur-1","DIV  com-content-1 fc"
    ],
    "offset": 4,
    "text": " [redacted] [ MP ]24 juin 2018 22:26 Salam ô alaïkom, Avec de la semoule
        fine, ça doit être super moche.   Pour répondre au commentaires précédents, je n'utilise
        que du couscous fin personnellement, rarement le moyen.      Répondre Citer"
}
```


FIGURE 4.17: Fragment Web de la page `https://www.yabiladi.com/forum/semoule-fine-pour-couscous-54-9290786.html`

Tel que retourné par Rivelaine, le fragment Web correspondant se présente à nous comme un ensemble de champs JSON interrogeables (Figure 4.17). Ainsi, le champ `node` renferme l'ensemble des nœuds HTML du fragment. Son contenu textuel complet est, quant à lui, présenté dans le champ `text`. Dans le détail, le fragment est constitué d'un auteur[38], d'une date, d'un texte et de deux nœuds d'expression (champ `category`), identifiés grâce à leurs labels (champ `label`). Quand plusieurs nœuds sont imbriqués les uns dans les autres, seul le label du nœud de plus haut niveau est retourné. Plusieurs liens hypertextes ont été extraits (champ `href`) et nous savons que ce fragment est le 4e a avoir été identifié sur cette page (champ `offset`). Mais il ne représente que 0.008% de l'ensemble du HTML de la page (champ `ratio`). Enfin, notons que l'extraction n'est malheureusement pas parfaite. Certains liens hypertextes pourraient ne pas être conservés (par

38. Le pseudonyme de l'auteur est masqué à dessein.

exemple : `javascript:///`) et nous n'avons pas su identifier le bouton `[MP]` comme un nœud d'expression permettant de répondre directement à l'auteur du message.

À mesure que se construisait Rivelaine, un changement de stratégie s'est opéré. Il s'agissait au début de définir une méthode de fragmentation universelle, capable de s'adapter à tous types de pages et à tous les âges du Web. Mais force est de constater que les heuristiques, non génériques, ont pris une place grandissante dans l'architecture de Rivelaine. L'automatisme, premièrement visé, s'est effacé au profit d'un agencement compliqué de *bricolages* par heuristiques et d'ajustements manuels. En effet, Fathom (pour ne citer que cet élément) est une méthode infiniment paramétrable. La forme d'un fragment Web, dépend ainsi de la valeur de *minDist* ou du rapport de grandeur entre les divers malus de la fonction de distance. Doit-on favoriser la cohérence visuelle ? Structurelle ? Sémantique ? Au détriment d'une autre ?

Le Web est ainsi fait qu'il reste, pour tout chercheur, pour tout explorateur d'archives, un terrain d'étude non trivial. Le Web est vaste et sauvage. Les sites ne se ressembleront jamais, certains seront *mieux* construits que d'autres et aucune méthode de fragmentation ne pourra satisfaire l'ensemble des chercheurs. Ainsi, lors d'une présentation de Rivelaine [39], l'historien Y. Scioldo-Zürcher suggéra d'ajouter plus de contexte aux fragments Web, d'élargir la segmentation des pages. La forme d'un fragment est, au final, dépendante de la question de recherche qui nous motive et de ce que l'on aimerait trouver dans les archives Web.

39. Lobbé Q., 2017, *Workshop : Introducing Web Fragments, Computational tools for the social study of Web archives*, Open University Of Israel, Tel Aviv

Sur les conseils de T. Drugeon, il a été décidé de replacer les chercheurs au cœur de la définition des fragments Web. Pour ce faire, une seconde implémentation de Rivelaine a été développée, cette fois en nodeJs [40], afin de la rendre autonome et interrogeable en tant que Web service. Nous construisons, par dessus cette application, un *addon* [41] (c.-à-d., une extension) au navigateur Firefox permettant de tester, sur le Web vivant, la fragmentation d'une page affichée à l'écran. Il est ainsi possible de jouer directement sur les réglages de Fathom et de sa fonction de distance pour apprécier la forme des futurs fragments Web. Une fois le chercheur satisfait, le paramétrage de Rivelaine est ajouté à la configuration du moteur d'exploration des archives et l'extraction peut débuter.

40. Voir `https://github.com/lobbeque/rivelaine/tree/master/nodejs`

41. Voir `https://github.com/lobbeque/rivelaine/tree/master/addon`, les addon ne sont malheureusement plus compatibles avec les versions récentes de Firefox

La fragmentation des archives Web, nous invite à une forme de souplesse et d'agilité vis-à-vis des données qui ne doit pas être le seul fait des outils que nous développons. Toute notre méthodologie d'exploration doit pouvoir débrayer, au besoin, d'un traitement large et automatisé des archives vers une analyse plus focalisée où une grande part du travail se fera à la main. Au cas par cas. Il en va ainsi de toute

étude sur le Web, soit une alliance à dimensions variables, entre le chercheur, l'algorithme, l'heuristique, le moteur...

Ainsi, Rivelaine et la stratégie d'extraction qu'elle renferme, ne doivent pas être vues comme la seule et unique manière de construire des fragments Web. Rivelaine n'est qu'une possibilité, parmi d'autres qui, nous l'espérons, arriveront bientôt. Ce que nous défendons, dans ce manuscrit, est la fragmentation des archives Web comme principe d'exploration et non la forme particulière de tel ou tel fragment. Il nous faudra, en revanche, être très clair, dans le Chapitre 5, sur la définition de nos espaces d'explorations à venir.

4.4 Penser une exploration désagrégée

Voyons, maintenant, comment les fragments Web peuvent nous aider à améliorer certains biais d'analyse identifiés en Section 3.2. Nous nous plaçons dans l'hypothèse d'une exploration désagrégée d'un corpus d'archives Web. Nous rappelons qu'un site Web archivé se compose de n pages Web numérotées $\{p_1,...,p_n\}$. Mais, désormais, une page p_j consiste elle même en m fragments Web numérotés $\{f_{j1}, ..., f_{jm}\}$. Nous supposons connaitre et avoir identifié la date d'édition de chacun de ces fragments $\phi(f_{j1}), ..., \phi(f_{jm})$.

Atténuer les cécités de crawl

En désagrégeant les archives Web, nous faisons l'hypothèse qu'une date d'édition sera toujours antérieure ou égale à une date de téléchargement. C'est une hypothèse plus ou moins forte et qui tient essentiellement à la nature des sites explorés. Une date d'édition peut, en effet, être falsifiée si elle est directement écrite par un humain. Mais, pour un site comme *yabiladi.com*, les dates d'édition, extraites de la partie forum, sont à la base automatiquement générées par le CMS Phorum. Sur l'ensemble de l'e-Diasporas marocaine, nous décomptons plus de 55% de sites associés à un CMS. Nous pouvons donc avoir raisonnablement confiance en la véracité des dates d'éditions manipulées, ainsi :

$$\forall p_j, f_{jk} \exists \phi(f_{jk}) : \phi(f_{jk}) \leq \mu_i(p_j) \leq t_i(p_j)$$

$$\text{où } c_i \text{ est un crawl dans lequel } f_{jk} \text{ existe}$$

Dans la Section 4.1, nous révélions que, par la désagrégation des archives, il était possible d'accéder à une mémoire antérieure à toute collecte, comme le rappelle la Figure 4.18.

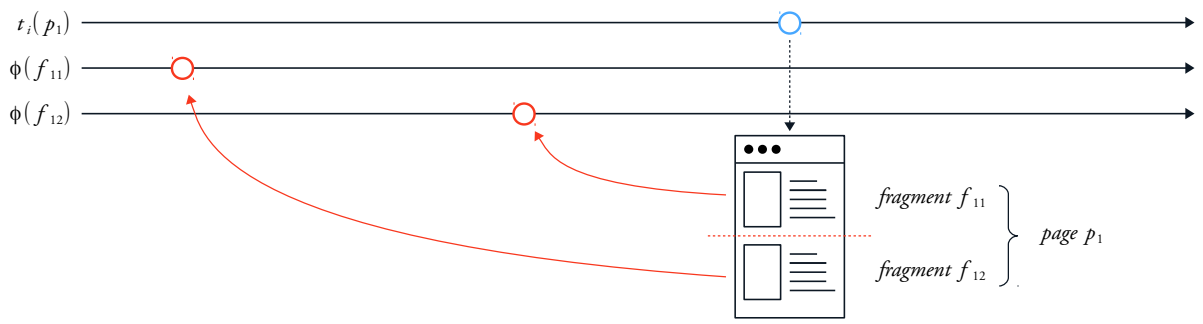


FIGURE 4.18: La fragmentation de la page p_1 permet d'accéder à des éléments antérieurs à la date de collecte $t_1(p_1)$

Mais essayons maintenant de quantifier ce potentiel gain de mémoire. Quelle peut être la différence, en jours, entre une date de téléchargement et une date d'édition ?

Pour ce faire, reprenons le cours de l'expérience, débutée en Section 4.1, où nous comparions la distribution (pour *yabiladi.com*) du nombre de pages et de fragments archivés suivant leurs dates de téléchargement et d'édition respectives. Nous sélectionnons les 109 534 pages archivées de la section forum de *yabiladi.com* que nous segmentons [42], via notre moteur, en 422 906 fragments Web associés à une date d'édition. Nous considérons, ensuite, la plus ancienne date d'édition de chaque page $\min_k \phi(f_{jk})$ pour s'approcher au plus près de leurs dates de création (Section 4.1).

42. Malus maximal sur la cohérence visuelle et les masques de continuité, pour qu'à tout post du forum corresponde un fragment Web

Nous calculons alors la différence $\min_i t_i(p_j) - \min_k \phi(f_{jk})$ entre dates de création et dates de première collecte. La Figure 4.19 donne à voir le gain en jours pour chaque page archivée.

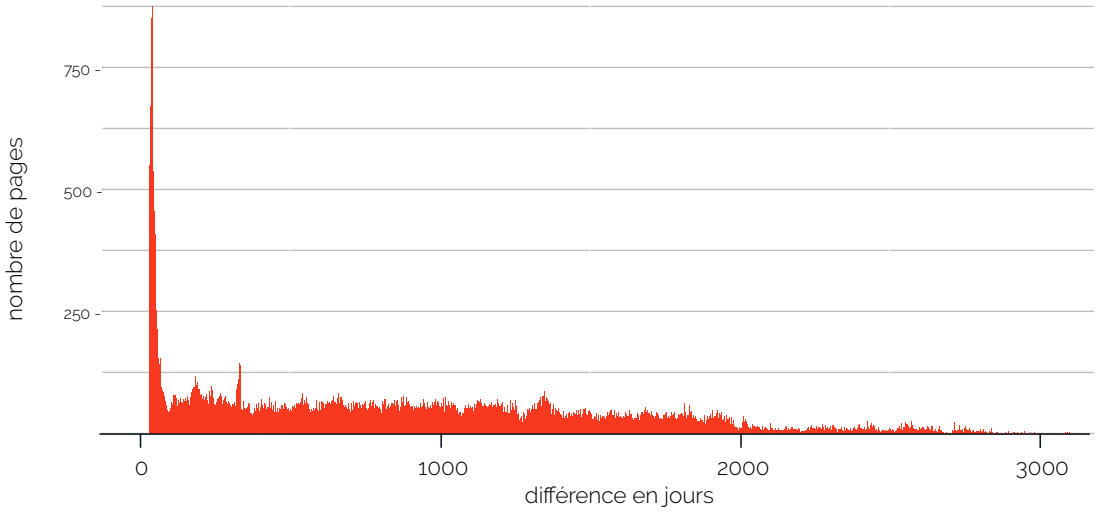


FIGURE 4.19: Distribution de la différence $\min_i t_i(p_j) - \min_k \phi(f_{jk})$ en jours

Les quartiles correspondants sont donnés par la Table 4.2. Pour 50% des pages archivées de *yabiladi.com* (Table 4.2, Q2) le gain est estimé à plus de deux années, la maximum étant de 3131 jours.

Table 4.2: Quartiles de la différence $\min_i t_i(p_j) - \min_k \phi(f_{jk})$ en jours

Quartiles	Différence en jours
Q1	256
Q2	777
Q3	1340
max	3131

Mais les bénéfices que nous constatons ici, peuvent être simplement liés à la date de commencement du crawl. En effet, la collecte menée par l'INA a débuté bien après la création de *yabiladi.com*. L'écart entre date de création et date de première collecte ne doit donc pas nous surprendre outre mesure.

Aussi, concentrons-nous plutôt sur des moments singuliers de cette campagne de crawl. D'une part, les six premiers mois de crawl qui correspondent à l'initialisation du corpus, où l'on cherche à capter des pages potentiellement antérieures de 7 ou 8 ans[43]. D'autre part, une année dite de routine (2012-2013), où le crawl n'a plus rien à rattraper et doit simplement se contenter d'archiver les pages nouvellement créées. Nous calculons alors, pour chaque moment, la même différence $\min_i t_i(p_j) - \min_k \phi(f_{jk})$, les résultats sont donnés par la Table 4.3.

43. *yabiladi.com* est crée fin 2002 et la campagne de crawl débute, elle, en 2010

Table 4.3: Quartiles de la différence $\min_i t_i(p_j) - \min_k \phi(f_{jk})$ en jours, pour les 6 premiers mois de crawl (cas n°1) et les années 2012-2013 (cas n°2)

Quartiles	Différence cas n°1	Différence cas n°2
Q1	428	39
Q2	875	49
Q3	1340	1229
max	2628	2389

Dans le premier cas, la différence de dates est, comme prévue, plus importante : celle-ci passant à 2 ans et 4 mois pour 50% des pages archivées (Table 4.3, cas n°1, Q2). Pour le second cas, les gains sont bien moins importants et oscillent majoritairement entre un et deux mois (Table 4.3, cas n°2, Q1 et Q2). En revanche, il est intéressant de noter que, même dans un contexte de crawls routiniers, des pages très anciennes continuent à être collectées (Table 4.3, cas n°2, Q3), et ce, alors que l'on se serait plutôt attendu à ne découvrir que des pages nouvellement mises en ligne.

Fragmenter les archives Web permet ainsi d'atténuer les diverses cécités de crawl. Mais ces bénéfices seront plus marqués dans le cadre d'une collecte déclenchée après la création du site ciblé (notre cas), que lors d'un crawl routinier (le cas d'Internet Archive).

Cohérence relative entre pages

Lorsque nous explorions les archives Web au niveau des seules pages (Section 3.2), nous avions défini la cohérence par observation comme l'existence d'un unique instant $t_{\text{coherence}}$ où se croisent les intervalles d'invariance respectifs des pages considérées (Spaniol *et al.*, 2009), et ce, dans le cadre d'un crawl unique.

Avec le fragment Web, nous pouvons dépasser cette définition et introduire la notion de **cohérence relative par observation**. Entre deux pages, la cohérence telle que nous la connaissons est absolue (Figure 3.4), l'entièreté des pages est considérée. Or, nous nous plaçons maintenant dans le cadre d'une analyse focalisée sur un élément particulier par page archivée.

Par exemple, si un chercheur souhaite vérifier la cohérence entre deux articles collectés, il pourrait vouloir connaitre la nature précise de l'intervalle d'invariance correspondant. À ses yeux, la cohérence serait hors sujet ou abusive, si le seul élément d'invariance entre les deux pages se révèlait être une barre de navigation plutôt que le corps des articles. Comment, dès lors, considérer la cohérence relativement à une question de recherche donnée ?

Sur ce point, nous définissons un sous-ensemble discret de fragments d'intérêt $\{f^*_{j1}, ..., f^*_{jl}\}$ (avec $l \leq m$). Le chercheur sélectionne, dans chaque page, un fragment Web qui lui semble pertinent pour son analyse. Puis il vérifie la cohérence $t^*_{\text{coherence}}$ entre toutes ces pages relativement aux fragments considérés, ainsi :

$$\forall p_j, \forall f^*_{jk} \in \{f_{j1}, ..., f_{jm}\}, \exists t^*_{\text{coherence}} :$$

$$t^*_{\text{coherence}} \in \bigcap_j [\phi(f^*_{jk}), t_i(p_j)] \neq \emptyset$$

En désagrégeant les archives Web, le chercheur peut reprendre la main sur les analyses qu'il entend mener au cœur des archives. Il peut focaliser, à souhait, son expérience d'exploration et déplacer son point d'observation relativement à son sujet. La Figure 4.20 décrit, de manière graphique, la différence entre cohérence par observation et cohérence relative par observation, pour deux pages p_1, p_2 et deux fragments Web choisis f^*_{11}, f^*_{21}.

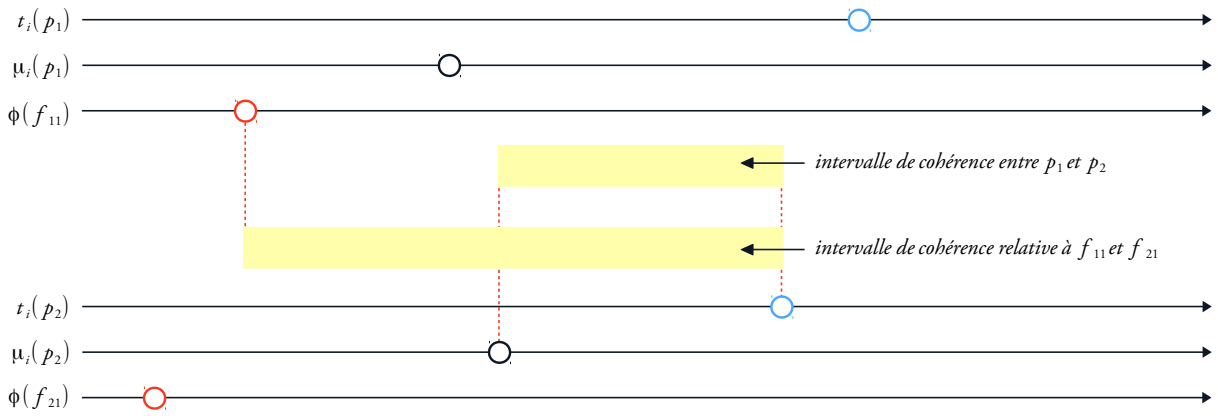


Figure 4.20: Cohérence par observation et cohérence relative par observation

Il serait intéressant d'étendre cette notion de cohérence relative par observation à la prise en compte de plusieurs fragments Web par pages archivées ou de plusieurs crawls. Enfin, nous n'avons pas eu ici l'occasion de proposer une expérimentation pratique de la cohérence relative par observation. Mais nous pensons qu'il est tout-à-fait faisable de l'intégrer à un système d'exploration d'archives Web et encourageons la mise en place future de travaux voulant traiter ce sujet.

Dédupliquer les corpus

En Section 3.2, nous nous sommes inquiétés de la possibilité de voir plusieurs fois le même contenu archivé dans un corpus. Cela peut provoquer des biais d'analyse, notamment si l'exploration est basée sur une recherche plein texte. Mais par la désagrégation des archives, il est possible de dédupliquer des éléments qui auraient été re-collectés d'un crawl à l'autre. En utilisant le fragment Web comme unité d'exploration, nous pouvons définir une **fonction d'identité** nommée *id*. Cette fonction compare l'invariance dans le temps, d'un fragment f_{jk} extrait d'une page p_j, au cours de deux crawls consécutifs c_1 et c_2 à $t_1(p_j)$ et $t_2(p_j)$ tel que :

$$id(c_1(f_{jk})) = c_2(f_{jk})$$

La Figure 4.21 illustre cette idée. D'un point de vue technique, la fonction d'identité ne peut être mise en place qu'au moment de retourner les résultats d'une requête depuis le moteur de recherche (Section 3.3). En effet, l'extraction des fichiers DAFF étant parallélisée, il serait trop couteux de maintenir un index dynamique des fragments Web déjà identifiés. Nous laissons donc, à dessein, des doublons dans les index. C'est lors de la restitution de ces résultats par Solr que nous les grouperons via la fonction d'identité.

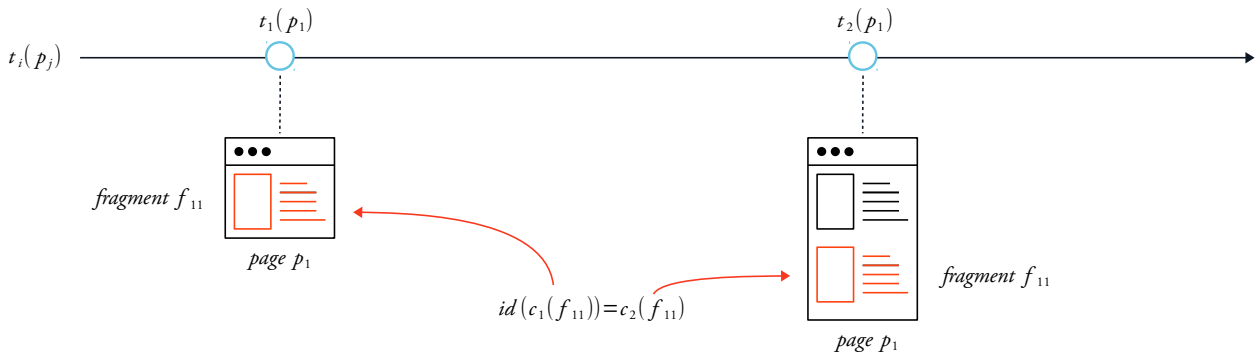


FIGURE 4.21: Dédupliquer les archives Web entre deux crawls c_1 et c_2 grâce à une fonction d'identité

Pour ce faire, nous associons à chaque fragment Web un champ unique appelé `frag_text_id`. Ce champ est le résultat du passage de l'ensemble du contenu textuel du fragment Web dans une fonction de hachage[44] SHA-256. Nous utilisons la fonction *group by*[45] de Solr pour grouper les fragments potentiellement dupliqués par `frag_text_id` unique.

44. Fonction déjà évoquée en Section 2.2 pour les identifiants des fichiers DAFF.
45. Voir `https://lucene.apache.org/solr/guide/6_6/result-grouping.html`

Dans notre modèle d'exploration désagrégé, une page Web archivée et observée à un instant t donné, ne sera plus que le résultat de l'assemblage de fragments Web précédemment publiés. La page Web, en tant que telle, disparait de notre modèle de données.

4.5 Intégration au moteur d'exploration

Notre moteur d'exploration d'archives Web, tel que nous l'avions décrit en Section 3.3, prend la page Web comme unité principale d'exploration. Mais, depuis, nous avons opéré un changement analytique de la page vers le fragment Web. Expliquons maintenant comment intégrer le fragment à notre moteur et proposons un premier cas d'usage, basé sur la détection d'événements dans les archives.

D'un schéma à l'autre

D'un point de vue pratique, l'extraction des fragments s'intègre à l'ensemble des traitements supervisés par Spark (Section 4.3). Une fois la jointure effectuée entre méta-données et données, notre moteur demande à Rivelaine de segmenter les pages archivées avant de les envoyer dans Solr pour indexation (Figure 4.22).

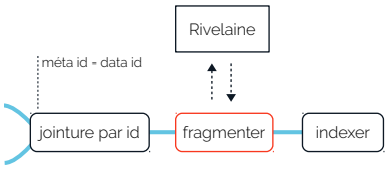


FIGURE 4.22: Intégration de la fragmentation au pipeline de transformation des archives (Voir Figure 3.7)

Un nouvel index doit alors être créé pour accueillir les fragments, le premier étant pensé pour les pages Web uniquement. Deux stratégies s'offrent ici à nous. Tout d'abord, conserver la page comme élément de référence, à laquelle nous subordonnons les fragments (Figure 4.23, (a)).

Ou éliminer la notion même de page et n'indexer que les fragments (Figure 4.23, (b)).

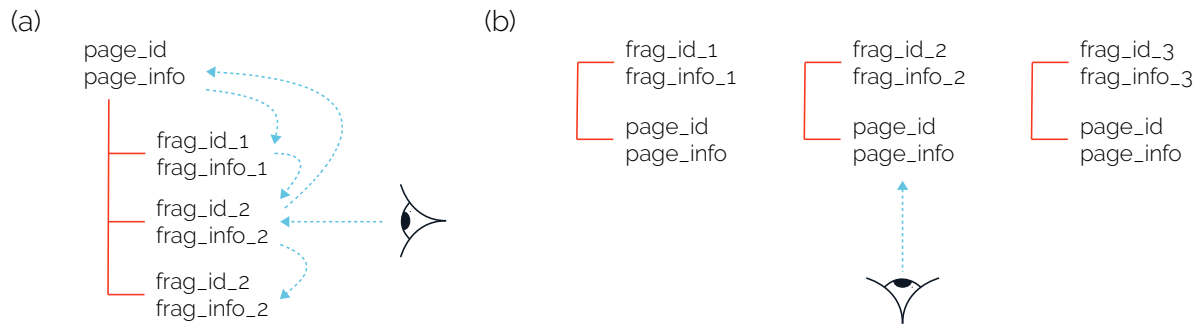


Figure 4.23: Différentes stratégies d'indexation du fragment Web dans un moteur de recherche et complexité de la recherche (bleu)

La première option, plus intuitive, conserve le lien ontologique entre la page et ses fragments Web. Elle nécessite la mise en place dans Solr, d'une structure dite de *nested documents* [46]. Dans un même index, deux types de documents cohabiteraient, un document page et des documents fragments. Malheureusement, cette stratégie est couteuse, notamment lorsqu'il s'agit d'identifier et de retourner des résultats. En effet, dans les moteurs de recherche, il est toujours conseillé (surtout lorsque le nombre de documents est important) de dupliquer au besoin les documents indexés. L'espace disque occupé par l'index sera plus important, mais les performances du *search*, en tant que tel, seront améliorées, la complexité d'une recherche par documents à plat étant moindre que celle par documents imbriqués pour laquelle toute la structure doit être retournée (Figure 4.23, tracés bleus).

46. Documents imbriqués, `http://yonik.com/solr-nested-objects/`

Nous nous orientons donc vers la seconde option, qui a pour effet de dupliquer les informations associées à une page à l'intérieur de chaque fragment. Si l'utilisateur veut, au besoin retrouver l'ensemble des fragments d'une même page, il pourra s'orienter vers une requête *group by* sur le champ `page_id` dans Solr.

Soit le schéma d'indexation des fragments Web présenté par la Figure 4.24. Dans ce schéma, l'`id` de chaque document correspond à l'identifiant unique d'un fragment. Les champs issus de Rivelaine (`frag_type`, `frag_offset`...) sont intégrés à l'index et la recherche plein texte est maintenant réalisée sur le seul champ `frag_text`. Pour retrouver l'ensemble des fragments d'une même page, on se servira du champ `page_url_id` et pour dédupliquer les fragments (Section 4.4), on s'appuiera sur la valeur de `frag_text_id` [47]. Les différents niveaux de notre échelle de datation (Table 4.1) sont indexés, le champ `page_date` correspondant à la date de création d'une page (Section 4.1). Le champ

47. Clé SHA-256 unique

frag_date est, lui, supposé contenir les dates d'édition de chaque fragment. Néanmoins, si nous sommes dans l'impossibilité d'associer une date d'édition à un fragment, le champ frag_date se verra attribuer la valeur de page_date, voire de la date de téléchargement download_date dans le pire des cas. Sur ce point, le type dateLvl nous renseigne sur le niveau de précision alloué au champ frag_date.

```
<field name="id"                type="string"  indexed="true"    multiValued="false" required="true" />

<!-- archive fields -->
<field name="archive_active"    type="boolean" indexed="true"    multiValued="false"/>
<field name="archive_corpus"    type="string"  indexed="true"    docValues="true" multiValued="false"/>
<field name="archive_country"   type="string"  indexed="true"    docValues="true" multiValued="false"/>
<field name="archive_lang"      type="double"  indexed="true"    docValues="true" multiValued="false"/>
<field name="archive_mime"      type="string"  indexed="true"    docValues="true" multiValued="false"/>

<!-- crawl fields -->
<field name="crawl_id"          type="string"  indexed="true"    docValues="true" multiValued="true" />
<field name="crawl_id_f"        type="string"  indexed="true"    docValues="true" multiValued="false"/>
<field name="crawl_id_l"        type="string"  indexed="true"    docValues="true" multiValued="false"/>
<field name="crawl_date"        type="date"    indexed="true"    docValues="true" multiValued="true" />
<field name="crawl_date_f"      type="date"    indexed="true"    docValues="true" multiValued="fasle"/>
<field name="crawl_date_l"      type="date"    indexed="true"    docValues="true" multiValued="true" />

<!-- download fields -->
<field name="download_date"     type="date"    indexed="true"    docValues="true" multiValued="true" />
<field name="download_date_f"   type="date"    indexed="true"    docValues="true" multiValued="false"/>
<field name="download_date_l"   type="date"    indexed="true"    docValues="true" multiValued="false"/>

<!-- page fields -->
<field name="page_domain"       type="string"  indexed="true"    docValues="true" multiValued="false"/>
<field name="page_url"          type="string"  indexed="true"    docValues="true" multiValued="false"/>
<field name="page_url_id"       type="string"  indexed="true"    docValues="true" multiValued="false"/>
<field name="page_link"         type="string"  indexed="true"    docValues="true"  multiValued="true"/>
<field name="page_title"        type="text"    indexed="true"    docValues="false" multiValued="false"/>
<field name="page_date"         type="date"    indexed="true"    docValues="true"  multiValued="true" />

<!-- fragment fields -->
<field name="frag_type"         type="string"  indexed="true"    docValues="true"  multiValued="true" />
<field name="frag_author"       type="string"  indexed="true"    docValues="false" multiValued="true" />
<field name="frag_date"         type="date"    indexed="true"    docValues="true"  multiValued="true" />
<field name="frag_date_level"   type="dateLvl" indexed="true"    docValues="true"  multiValued="false"/>
<field name="frag_href"         type="string"  indexed="true"    docValues="false" multiValued="true" />
<field name="frag_label"        type="string"  indexed="true"    docValues="true"  multiValued="true" />
<field name="frag_ratio"        type="int"     indexed="true"    docValues="true"  multiValued="true" />
```

```
<field name="frag_node"         type="text"    indexed="false"   docValues="false" multiValued="true" />
<field name="frag_offset"       type="int"     indexed="true"    docValues="true"  multiValued="true" />
<field name="frag_text_id"      type="string"  indexed="true"    docValues="true"  multiValued="false"/>

<!-- searchable fragment fields -->
<field name="frag_text"         type="text"    indexed="true"    stored="false"  multiValued="true" />
<field name="frag_text_shingle" type="shingle" indexed="true"    stored="false"  multiValued="true" />
```

Figure 4.24: Schéma d'indexation des fragments Web

Détection d'événements

Comme cas d'usage pratique du fragment Web, nous souhaitons maintenant ajouter à notre moteur d'exploration un système de détection d'événements dans les archives. Cette proposition a fait l'objet d'une publication démonstration[48] dont l'application est, jusqu'à présent, limitée aux seules archives du site *yabiladi.com*.

48. Lobbé, Q. (2018), *Revealing Historical Events out of Web Archives*, TPDL 2018

La recherche par événements est une alternative aux méthodes d'exploration classiques qui se font principalement par URL (Section 2.2). Des travaux récents tentent également de s'en affranchir en proposant des analyses par catégories (Holzmann et Anand, 2016), par entités nommées (Spaniol et Weikum, 2012) ou basées sur des tendances issues des réseaux sociaux (Risse *et al.*, 2014). Dans l'ensemble, tout indique que, face à des corpus d'archives Web si vastes, une exploration dirigée ou guidée (par exemple sur la base d'événements) peut être bénéfique.

Nous pensons que tout explorateur d'archives (Web ou autre) poursuit à un moment donné la recherche d'événements singuliers qu'il puisse mettre au regard de l'histoire (Chaney *et al.*, 2015). Ainsi, pour J. Baschet (Section 4.1) l'étude historique de lignes processuelles passe par une pensée de l'événement comme « *surgissement* » ou « *rupture* » (Baschet, 2018, p.227).

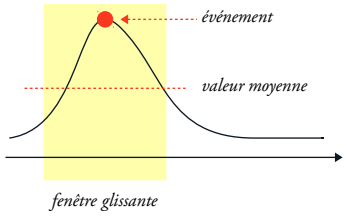


Figure 4.25: Détection d'événements à partir d'un seuil

Au sein d'une distribution temporelle d'éléments donnés, un événement peut être caractérisé de trois manières différentes comme le propose (Viard, 2016, p.106) : par sa détection, par son identification et par son explication.

Un événement **détecté** est une anomalie présente dans une distribution. Il peut s'agir d'un instant singulier ou d'une période entière durant laquelle le niveau de messages postés sur un blog est supérieure à une valeur de référence (une moyenne par exemple). Un événement **identifié** est un événement pour lequel un élément de causalité aura été trouvé dans les données sources. Un pic d'activité dans un forum en ligne est, par exemple, causé par un nombre important de messages postés. Enfin, un événement **expliqué** est un événement pour lequel

une explication aura été trouvée et validée à l'extérieur des données sources. Ici, une expertise et une analyse humaine est nécessaire. Un pic de messages dans un forum peut, ainsi, être expliqué par un contexte social ou politique particulier qui aura fait réagir certains membres de la communauté.

Continuant sur notre logique exploratoire, nous ne voulons pas ici cibler une forme particulière d'événements, nous éviterons donc les méthodes de détection par motif (Chaney *et al.*, 2016) ou classification (Dodds *et al.*, 2011). Nous nous orientons plutôt vers une méthode de détection par **seuil** (Fung *et al.*, 2005) à l'intérieur d'une fenêtre glissante d'une semaine[49]. Nous définissons ici un événement comme une valeur aberrante détectée au sein d'une distribution de fragments Web (Figure 4.25).

Pour ce faire, le contenu textuel de chaque fragment Web indexé est divisé en **bigrammes**[50]. Un bigramme est une séquence de deux mots consécutifs extraite d'un même ensemble textuel (Figure 4.26).

49. Sur *yabiladi.com*, la durée de vie moyenne d'un thread de messages est d'un peu plus d'un jour, le choix de prendre la semaine comme granularité nous a donc semblé judicieux

50. Le champ `frag_text_shingle` de notre schéma d'indexation (Figure 4.24)

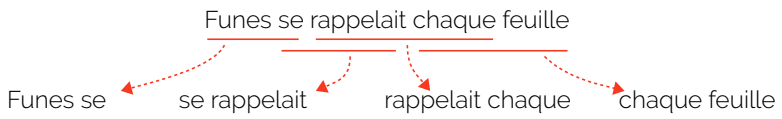


FIGURE 4.26: Division d'une phrase en bigrams

Dans le cadre de cette démonstration, la recherche plein texte porte sur le contenu textuel des bigrammes. Le moteur ne retourne donc plus que les fragments Web dont certains bigrammes auraient correspondu à des mots clés proposés par un chercheur. Dans notre cas, identifier un événement revient, en réalité, à détecter un pic soudain de bigrammes dans le temps. Les bigrammes sont souvent utilisés pour observer des tendances ou des évolutions lexicales au sein de grands corpus de textes. On citera par exemple, le système *ngrams viewer*, conçu par les équipes de Google books, qui permet de suivre l'évolution temporelle de l'utilisation de certains mots dans leur base de données de livres numérisés (Michel *et al.*, 2011).

Pour terminer, nous essayons d'expliquer nos événements en trouvant des corrélations avec certains titres d'articles de news, extraits des archives de la section actualités du site *yabildai.com*. En effet, nous faisant l'hypothèse que les utilisateurs du forum sont susceptibles de réagir, par le biais de messages, à un événement précis de l'actualité (avant ou après que celui-ci ait été rapporté dans la presse).

Nous construisons ainsi, à la volée, un nouvel index d'événements potentiels en utilisant le titre et la date d'édition de ces articles archivés.

Lorsqu'un pic de messages est détecté, une requête secondaire est automatiquement adressée à notre index d'événements possibles. Si l'un des titres correspond également à la recherche du chercheur et que sa date d'édition se trouve à moins d'une semaine de la date du pic détecté, alors nous proposons ce titre comme potentielle explication de cette soudaine recrudescence de messages.

FIGURE 4.27: Ajout de la détection d'événements à notre interface d'exploration

Notre interface de visualisation est modifiée en conséquence et permet maintenant au chercheur de choisir la granularité d'exploration qui l'intéresse le plus, entre la page et le fragment (Figure 4.27). Dans la partie fragment Web, apparait un nouvel histogramme affichant la distribution dans le temps des bigrammes ayant correspondu à la requête du chercheur. C'est sur cet histogramme que sont affichés les événements détectés et leur possible explication. Dans une vidéo de démonstration [51], nous décrivons différents cas d'usages de ce système et, plus généralement, nous y présentons le fonctionnement du moteur d'exploration.

51. Consultable ici : `https://youtu.be/snW4O-usyTM`

Nous donnons ainsi à voir, comme exemple d'événements détectés dans les archives, des pics de discussion autour de diverses actualités liées au roi du Maroc Mohammed VI et des réactions à la destitution de l'ancien dirigeant tunisien Z. Ben-Ali, au début de l'année 2011 (Figure 4.27, tracé vert). Dans le Chapitre 5, nous reviendrons en détail sur la manière avec laquelle le forum de *yabiladi.com* a réagi à ce moment historique particulier qu'a été le Printemps arabe.

***

Malgré notre idée première, l'extraction des fragments Web n'a pas pu se faire sans une forte dose d'heuristiques non génériques. Le Web est ainsi fait qu'il nous oblige, sans cesse, à ré-adapter l'échelle de nos analyses. Passant de vastes traitements automatisés à des moments de pur travail manuel. Au cas par cas. Nos explorations à venir ne contrediront pas cet état de fait : notre méthode d'exploration chemine dans un entre-deux constant entre approche quantitative et validation qualitative. Il nous faut ainsi mettre les mains dans les archives, les ouvrir et y plonger.

C'est, au final, toute l'ambition du fragment Web, tel que nous l'avons introduit dans ce chapitre. Redonner au chercheur les moyens théoriques et techniques d'une plus grande maniabilité des archives. Le terrain théorique ayant été préparé, dans le courant des années 2000, par les travaux pionniers de N. Brügger, il existait un espace analytique à explorer entre l'élément Web et la page Web. Nous définissons ainsi le fragment Web comme un sous-ensemble sémantique et syntaxique d'une page Web.

Avec le fragment Web et la grammaire qu'il introduit, le chercheur quitte les interfaces d'exploration vitrines et s'assoit à la table de montage où il peut découper, déplacer et mettre en relation des éléments épars du Web passé. Nous avons ainsi montré comment, en s'appuyant sur les dates d'édition plutôt que sur les seules dates de téléchargement, les archives en arrivent à basculer d'une temporalité à l'autre. Elles quittent le temps des crawlers pour retrouver le temps du Web tel qu'il a été. Le chercheur accède alors à une mémoire antérieure aux collectes des archivistes, mémoire jusqu'ici retenue derrière le verrou des fichiers sauvegardés. Le fragment témoigne ainsi directement du geste des auteurs, lecteurs et bloggers du Web passé et replace, de fait, l'humain au cœur de l'étude des archives Web.

Intégrés à notre moteur d'exploration, nous nous appuierons sur les fragments Web pour mener, dans le chapitre suivant, deux explorations à travers notre corpus de sites marocains. Nous étudierons l'histoire de collectifs migrants en ligne, depuis longtemps éteints et dont la trace ne subsiste aujourd'hui plus que dans les archives.

Chapitre 5

| Exploration de collectifs migrants éteints

Le processus de cartographie de l'e-Diaspora marocaine s'est étalé sur près de deux ans, de 2008 à 2010. L'archivage de ces sites s'est alors présenté comme une nécessité, leurs traces commençant déjà à disparaitre de la toile. Dix ans plus tard, en 2018, sur les 153 sites observés à l'époque, seuls 53 peuvent encore être qualifiés de vivants, c'est-à-dire : toujours en ligne et ayant été mis à jour au moins une fois au cours de la dernière année passée. Les deux autres tiers sont pour la plupart hors-ligne, tombés à l'abandon ou récupérés par des cybersquatteurs [1].

Le Web est un environnement éphémère (Section 2.2) et les e-Diasporas ne sont, au final, que des images instantanées d'un espace qui, une fois la capture effectuée, n'est déjà n'est plus tout à fait le même. Il a continué à évoluer et s'est transformé.

Parmi les nombreux sites du corpus marocain, les plateformes institutionnelles (Section 1.3) sont celles qui ont le mieux résisté au passage du temps. En effet, 50% d'entre elles sont toujours vivantes et leur ancrage durable en ligne semble résulter d'une stratégie planifiée. La petite dizaine de sites consulaires identifiés en 2008 s'est ainsi muée en un seul et même portail Web *consulat.ma* [2] sous l'impulsion du gouvernement marocain. Puis, derrière elles, viennent les sites des associations et ONG dont 30% sont encore accessibles. Mais c'est le cas de la blogosphère qui, en ce début de chapitre, nous interpelle le plus : sur les 48 blogs cartographiés, seuls 5 sont aujourd'hui vivants (soit à peine 10%).

Au sein de l'e-Diaspora marocaine, la blogopshère avait été caractérisée [3] comme une communauté densément liée entre chacun de ses membres. Soit un espace d'expression et de représentation où circulaient, tout autant, commentaires politiques que témoignages de la vie quotidienne des marocains de l'étranger. Pourquoi ce groupe de blogs s'est-il à ce point réduit, perdant au cours du temps son influence et son dynamisme ? Les blogueurs ont-ils simplement disparu et quitté la toile ? Ou ont-ils migré vers d'autres territoires du Web ? Peut-on remonter le fil des potentielles traces archivées de cette cybermigration

1. Technique de hacking consistant à prendre possession d'un nom de domaine existant et de publier à sa place un contenu autre (e-commerce, publicités, arnaques, pornographie...).

2. Voir `http://www.consulat.ma/fr/aide.cfm`, site lancé en 2011 dans le cadre du Plan Maroc Numeric 2013 (`http://www.egov.ma/sites/default/files/Maroc%20Numeric%202013.pdf`)

3. Voir l'intervention de M. Renault (`https://www.youtube.com/watch?v=1sE5PZVG6iM`)

et en déterminer la destination ?

Au cours de ce chapitre, nous mènerons deux explorations successives, proposant chacune une utilisation différente des fragments Web. La première nous permettra de comprendre la mutation progressive des blogs marocains vers les plateformes de réseaux sociaux à la fin des années 2010. Cette transformation, nous le découvrirons, est autant le fait d'une évolution technologique inhérente au Web, que le résultat de l'influence d'événements socio-politiques extérieurs à la toile. Parmi ces événements, le Printemps arabe et, plus particulièrement, la manifestation marocaine du 20 février 2011 semblent avoir été des moments déterminants de l'histoire de la blogosphère.

En suivant cette nouvelle piste, nous questionnerons la réception de ce même événement par les membres d'un autre collectif en ligne, central dans la vie de l'e-Diaspora marocaine : le forum de *yabiladi.com*. Nous comprendrons la manière avec laquelle les membres de ce site se sont momentanément organisés et retrouvés emportés par la vague révolutionnaire qui déferlait alors sur le Maghreb et le Moyen Orient.

Comme un bilan de ces deux explorations nous introduirons, pour terminer, la notion de **moments pivots du Web** : des instants particuliers de l'histoire de la toile où celle-ci change soudainement de direction par la rencontre d'une avancée technologique et d'un groupe d'utilisateurs capables de s'en saisir [4].

4. Ce chapitre a fait l'objet d'une publication : Q. Lobbé, (2018), *Where the dead blogs are*, ICADL, 2018

Mais, avant de débuter notre analyse, nous souhaitons revenir sur les aspects théoriques de ce que nous nommons **exploration** depuis le début de ce manuscrit. Il s'agit pour nous de présenter comment, ce courant des statistiques qu'est l'analyse exploratoire de données (AED), peut aujourd'hui servir de cadre méthodologique à une recherche portant sur le Web passé.

5.1 À la recherche de l'étonnement

Dans le courant du 19e siècle, la théorie statistique prend un tournant rigoriste majeur en préférant, à la longue tradition exploratoire [5], des analyses et des inférences toutes tournées vers la seule quête de l'optimalité, de la moyenne et de la loi normale (Ladiray, 1997). Les statisticiens tombent alors dans une logique **confirmatoire**, faisant des hypothèses fortes sur la nature même des données étudiées qui, dès lors, ne sont plus sources de calculs statistiques, mais un moyen vers la validation d'un modèle. La donnée est subordonnée au modèle et le réel doit venir confirmer la théorie. La domination des statistiques confirmatoires s'établit ainsi au détriment des statistiques exploratoires, souvent décriées par l'appellation péjorative de statistiques **descriptives**.

5. Les statisticiens ont longtemps été des explorateurs de données, palliant certaines lacunes théoriques, par l'utilisation d'outils de recueil des données simples et visuels, comme le comptage par Quipu chez les civilisations précolombiennes (`https://fr.wikipedia.org/wiki/Quipu`).

L'analyse exploratoire de données

L'**analyse exploratoire de données** (AED) est un courant de la statistique moderne introduit, au cours des années 60 et 70, par le statisticien J.W. Tukey (Tukey, 1977). Voulant réintroduire le réel au cœur des statistiques, Tukey préfère la résolution (même approximative) d'un problème véritable à la recherche d'un indicateur optimal sur une question de peu d'intérêt :

« *The most important maxim for data analysis to heed, and one which many statisticians seem to have shunned, is this : Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise. Data analysis must progress by approximate answers, at best, since its knowledge of what the problem really is will at best be approximate. It would be a mistake not to face up to this fact, for by denying it, we would deny ourselves the use of a great body of approximate knowledge, as well as failing to maintain alertness to the possible importance in each particular instance of particular ways in which our knowledge is incomplete.* » — (Tukey, 1962, p.13-14)

Pour Tukey, les statistiques doivent ainsi retrouver le gout de l'incertitude et traiter à nouveau de la réalité. Ainsi tout modèle ou toute analyse doit, selon lui, partir d'une réflexion empirique, sensible. Les données doivent guider le choix des méthodes utilisées pour les étudier, et non l'inverse. C'est par la connaissance intime des données, que l'explorateur peut faire émerger des pistes de recherche et des hypothèses de travail. Il faut éprouver (qualitativement) les données dans la durée et retrouver une forme de rapport artisanal vis-à-vis d'elles. En cela, l'AED est un processus fondamentalement itératif, basé des boucles d'explorations successives, où l'analyste est amené à conjuguer intuition, capacités visuelles et expérience des données.

Là où les statistiques confirmatoires s'inscrivent dans l'administration de la preuve, l'analyse exploratoire, elle, s'articule autour d'une logique d'observation et de découverte.

Suivre et s'attacher aux indices

L'AED avance par investigation, guidée par la découverte de l'inattendu. L'explorateur multiplie les facets et les vues lui permettant de décrire les données dans leur ensemble et se place ainsi dans une position d'étonnement. Il s'agit de s'attacher autant aux tendances générales qu'aux détails, pour progresser pas à pas, par hypothèses successives en suivant la piste des indices précédemment révélés (Figure 5.1). Mais, dans le l'esprit de Tukey, l'AED ne doit pas remplacer l'analyse confirmatoire, il convient plutôt de conjuguer les deux : les boucles d'exploration (Figure 5.1, a) préfigurent et induisent toute modélisation (Figure 5.1, b).

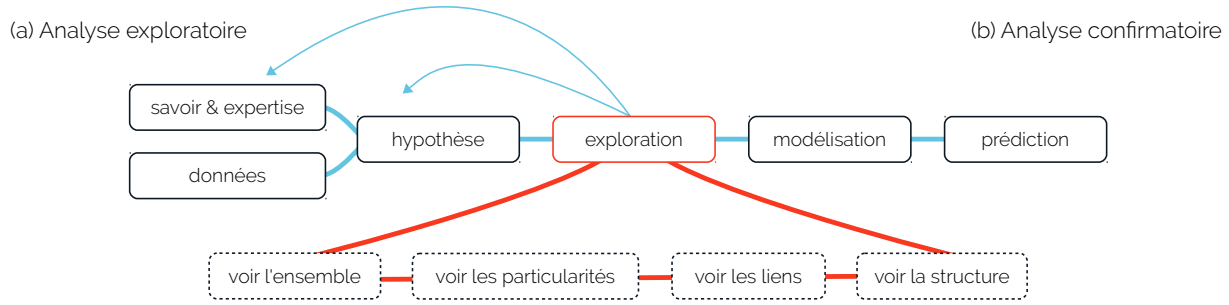


Figure 5.1: L'analyse exploratoire et les diverses phases de l'exploration préfigurent toute analyse confirmatoire

L'AED rencontre rapidement un écho grandissant, porté par le développement des techniques de calculs et de visualisations informatiques. Plusieurs années après, enrichissant sa réflexion par la lecture des travaux de J. Bertin [6] (Bertin *et al.*, 1967), B. Fry propose de définir ce qu'il nomme **conception computationnelle d'information** [7] (Fry, 2004), comme une version moderne de l'AED, explicitement orientée vers la visualisation de données et la programmation informatique. Tout autant processus d'exploration itératif (Figure 5.2) qu'art de la représentation de la donnée, la méthode de Fry accompagne et motive le développement d'outils neufs, tels que l'environnement de programmation Processing [8] ou encore la librairie Javascript D3 créée par M. Bostock [9].

6. Inspirateur de la sémiologie graphique ou science de la représentation graphique des données (`https://visionscarto.net/la-semiologie-graphique-a-50-ans`).

7. *Computational information design* en anglais

8. `https://processing.org/`

9. https ://d3js.org/

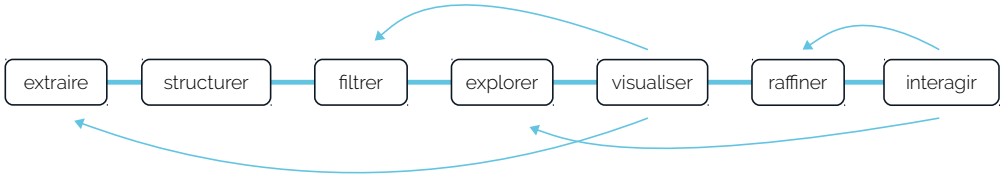


Figure 5.2: Étapes de la conception computationnelle d'information et ensemble des boucles d'analyses possibles selon (Fry, 2004)

Ce mode d'analyse par explorations offre un cadre méthodologique adapté à l'étude que nous souhaitons ici mener sur les archives Web. Concernant le devenir des blogs marocains après 2008, nous ne disposons au début de notre étude que d'un faisceau assez faible d'indices susceptibles d'expliquer la disparition de la blogosphère. De plus, la taille de l'espace d'exploration (48 blogs segmentés en un multitude de fragments Web) et son étendue (4 années d'archives) nous pousse à cheminer par à-coups et à inférer de nos trouvailles de nouvelles hypothèses de travail. En cela, l'AED est souvent comparée (par Tukey lui même [10]) au travail du détective qui tisse la trame de ses recherches

10. « *(EDA) is detective work – numerical detective work – or counting detective work – or graphical detective work... unless exploratory data analysis uncovers indications, usually quantitative ones, there is likely to be nothing for confirmatory data analysis to consider* » — (Tukey, 1977, p.1-3)

à travers un réseau d'indices.

Ce raisonnement par inférence, qui émerge de notre description de l'AED, peut être rapproché d'un autre mode de pensée découvert par C. Ginzburg en 1979 : le **paradigme indiciaire** (Ginzburg, 1980). En effet, l'historien révèle que, dans nombre de disciplines liées aux sciences humaines (psychanalyse, histoire de l'art, médecine...), s'est installé au fil du temps, un paradigme permettant de révéler à partir d'effets ou d'éléments singuliers des événements plus généraux[11]. Le psychanalyste progresse par l'étude des symptômes du patient et l'historien de l'art étudie les traces du geste du maître derrière le pinceau.

11. « *Si la réalité est opaque, des zones privilégiées existent (traces, indices) qui permettent de la déchiffrer.* » — (Ginzburg, 2012, p.290)

Née à la suite des travaux de Ginzburg, la **micro-histoire** s'attache à analyser historiquement des objets de tailles limitées. Les chercheurs cheminent d'indices en indices, à la manière d'un Sherlock Holmes, pour compléter l'étude de trajectoires et de motivations individuelles. La micro-histoire se veut, ainsi, le contre-poids focalisé d'une pratique plus quantitative du travail de l'historien.

Comme pour l'AED face aux statistiques confirmatoires, on reproche à la micro-histoire la non représentativité de ses travaux, trop précis et détachés de l'étude des tendances historiques normales et des grandes structures de masse. Mais, sur ce point, et c'est l'élément qui nous intéresse ici tout particulièrement, selon Ginzburg l'anomalie ou l'objet singulier peut se révéler porteur d'une résonance plus globale :

« *La possibilité de passer du cas isolé à la généralisation part d'une hypothèse qui a gagné en clarté à travers le temps. Aujourd'hui, je proposerai de considérer un individu comme le point d'intersection d'une série d'ensembles différents qui ont chacun une dimension variable. [...] L'historien doit partir de l'hypothèse que chez tout individu quel qu'il soit, coexistent des éléments partagés par un nombre variable (entre plusieurs milliards et zéro) d'individus. L'anomalie sera le résultats des réactions réciproques entre tous ces éléments.* » — (Ginzburg, 2012, p.359-360)

Dans les explorations à venir nous nous attarderons autant sur les tendances générales que sur les détails singuliers, suivant la tradition de l'AED. Mais l'incomplétude des archives Web, leur caractère partiel et détaché (sous certains aspects) de la réalité du Web vivant, nous conduiront à affiner petit à petit notre échelle d'analyse, à nous concentrer sur de plus petits objets dont certains détails nous auront révélé l'importance.

Ce faisant, il peut alors être vertigineux de penser que, partant d'une archive de 70 téra-octets, on en vienne à explorer, *seulement*, le devenir en ligne d'une poignée d'internautes. Mais, en nous inspirant des travaux de Ginzburg, nous chercherons à repositionner ces éléments singuliers par rapport aux grandes vagues historiques dont ils ont été témoins ou vecteurs privilégiés.

5.2 Les traces d'une mutation numérique

Concentrons nous maintenant sur le devenir de la blogosphère marocaine, entre 2008 et 2018. Essayons de comprendre, à travers les archives Web, le ou les processus qui ont amené à sa disparition quasi-complète. Sommes-nous témoins d'une extinction ? Ou d'une migration d'un territoire du Web vers un autre ? Peut-on encore trouver, dans nos collectages, les traces archivées d'une possible mutation des blogs vers les plateformes de réseaux sociaux ?

D'une communauté de blogs...

Entre 2008 et 2010, D. Diminescu et M. Renault cartographient et recensent 47 blogs [12] qu'ils intègrent à l'e-Diaspora Marocaine, alors en construction. Cette blogosphère est l'une des trois communautés [13] majeures du réseau de sites marocains tel que nous l'avons introduit en Section 1.3.

12. Voir la carte `http://www.e-diasporas.fr/wp/moroccan.html`

13. En théorie des graphes, une communauté de nœuds est un groupe fortement lié en son sein et faiblement lié avec le reste du réseau (Wasserman et Faust, 1994; Scott, 2017).

Le terme **blogosphère** est un néologisme forgé, au début des années 2000, pour désigner les communautés de blogs alors florissantes sur la toile. À cette époque, d'anciens concepts, comme logosphère ou graphosphère [14], ne suffisent plus à traduire la nature hypertextuelle des liens établis entre ces sites. La connexion par liens de citation étant l'une des mécaniques de base des blogs et de leurs environnements. En effet, les auteurs en ligne s'agrègent en listes et communautés de blogs amis, qu'ils affichent de manière ostensiblement visibles sur leurs sites, comme marque consciente d'appartenance à un groupe délimité (Keren, 2006). On n'entre pas dans une blogosphère, on s'y intègre. On cherche à la rejoindre en se connectant à ses membres qui, en retour, nous reconnaissent comme l'un des leurs.

14. Communautés humaines basées sur le langage oral d'une part, et l'écrit d'autre part.

Type particulier de site Web, un blog est une plateforme de publication, en ligne, de billets ou d'articles courts. Les blogs sont alimentés en contenu de manière régulière et se rapprochent en cela du journal intime : chaque billet est daté et signé par un auteur (le créateur du blog ou une personne partageant les droits d'édition), l'ensemble des articles se lit ensuite du plus récent au plus ancien. Ces textes (incluant aussi images, liens, vidéos...) peuvent être vus et commentés par d'autres blogueurs ou par de simples visiteurs.

En cela, un blog est, avant tout, un médium destiné à une population d'amateurs de la toile. Sa facilité de création et de maintenance a fortement joué en faveur de la diffusion de ce support d'écriture auprès d'un large public. Il était ainsi possible, en quelques clics, de créer via des plateformes dédiées et souvent gratuites [15] son premier blog, et ce, sans être un professionnel de l'informatique. De fait, les blogs

15. On peut citer Overblog, Blogger, Skyblog...

ont permis une forme de démocratisation de la prise de parole sur le Web, atteignant leur apogée [16] à la fin des années 2000. Les sujets de publication y sont divers et fonction du goût de chacun : on y parle de son quotidien, on écrit des billets d'humeur, voire des commentaires et des analyses politiques. Le tout partagé et diffusé au sein de sa ou de ses communautés d'appartenance.

16. Les statistiques varient d'une source à l'autre, mais on estime généralement qu'entre 150 et 200 millions de blogs avaient été crées en 2011 `https://www.statista.com/statistics/278527/number-of-blogs-worldwide/`.

Sur ce point, de premières études ont rapidement montré le rôle grandissant des blogs quant à la diffusion et à la cristallisation d'opinions politiques sur le Web, souvent en rupture avec les discours tenus par les médias classiques (télévision, radio...). Par exemple, (Adamic et Glance, 2005; Adar *et al.*, 2004) caractérisent les divergences de pensées démocrates et républicaines en ligne, lors de la campagne présidentielle de 2004 aux États-Unis. En France, l'étude du collectif RTGI [17] révèle la forte structuration et l'influence, sur la toile, des partisans du *non* avant la tenue du référendum sur la Constitution Européenne de 2005 (Fouetillou, 2008).

17. *Réseaux, Territoires et Géographie de l'Information*, collectif fondé à l'Université de Technologie de Compiègne (UTC) par F. Ghitalla (`https://web.archive.org/web/20060702021013/http://www.utc.fr:80/rtgi/`).

Passé le tournant des années 2010, les blogs perdent peu à peu de leur influence, (Weltevrede et Helmond, 2012) constatent ainsi, en 2012, la lente disparition de la blogosphère néerlandaise. Cette perte de vitesse se fait au profit des plateformes de réseaux sociaux (Facebook, MySpace, Instagram...) et de micro-blogging (Twitter, Tumblr, Sina Weibo...). Ces dernières reprennent à leur compte, et améliorent, les mécaniques de publication introduites par les blogs, accélérant toujours la diffusion et le partage de contenu sur le Web.

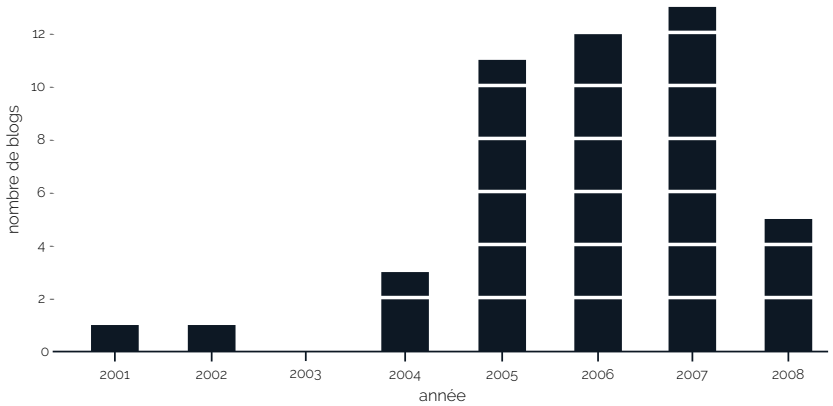


FIGURE 5.3: Répartition des blogs de la blogosphère marocaine par année de création

En 2008, la **blogosphère marocaine** [18] est une communauté fortement connectée de 47 sites Web crées ou maintenus par des citoyens marocains de l'étranger. Ce collectif est relativement jeune, la majorité des blogs ayant moins de trois années d'existante (Figure 5.3) au moment

18. Par abus de langage, nous appelons ici blogosphère marocaine le groupe de blogs de l'e-Diaspora marocaine. Il existe et a existé, bien évidement, d'autres blogosphères au Maroc, ce pays comptant près de 30 000 blogs en 2008 (`https://mg.co.za/article/2008-01-08-moroccos-blogosphere-takes-off`).

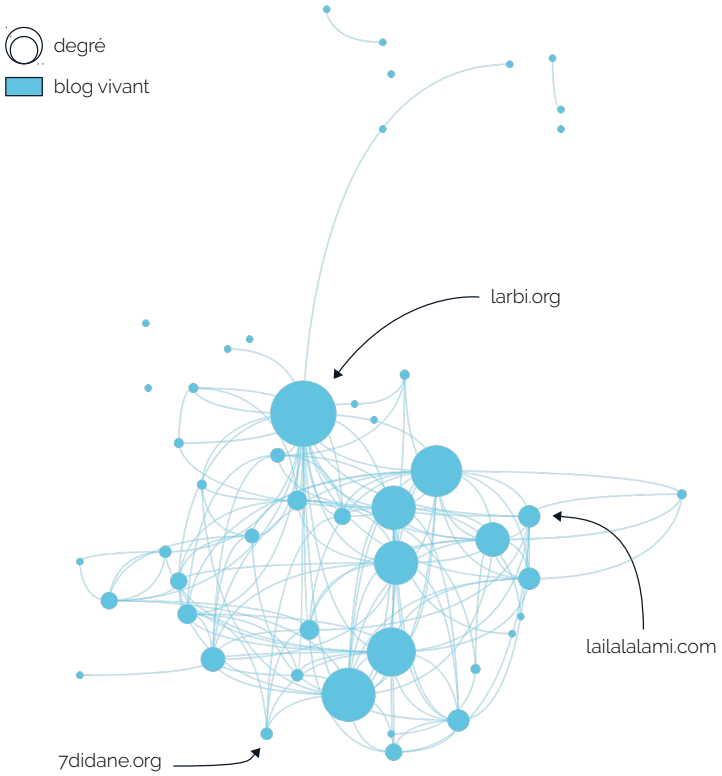
Préfixe d'URL	Catégorie	Langue principale	Pays du blog
7didane.org	intimiste	Français	USA
9afia.blogspot.com	littérature	Français	USA
adilski.blogspot.com	mixe	Anglais	USA
al9adiya7amda.blogspot.com	social – politique	Français	USA
anasalaoui.com	social – politique	Français	France
badrryadi.centerblog.net	littérature	Français	France
blogreda.blogspot.com	mixe	Français	USA
boubouh.over-blog.com	social – politique	Français	France
cabalamuse.wordpress.com	mixe	Anglais	USA
eatbees.com/blog	intimiste	Anglais	USA
emigrant.canalblog.com	intimiste	Français	France
enmarruecos.blogspot.com	intimiste	Espagnol	USA
fatima-salma.over-blog.com	cuisine	Français	France
kalima.hautetfort.com	littérature	Français	France
karlamassini.blogspot.com	littérature	Français	USA
kennza.wordpress.com	mixe	Français	USA
kingstoune.com	mixe	Français	France
klamia.canalblog.com	mixe	Français	France
kugelschreiber.canalblog.com	humeur	Français	France
labelash.blogspot.com	mixe	Anglais	USA
lailalalami.com	littérature	Anglais	USA
lallamenana.free.fr	humeur	Français	France
larbi.org	social – politique	Français	France
lemythe.com	intimiste	Français	Maroc
lesamismarocains.blogspot.com	humeur	Français	USA
louladekhmissbatata.wordpress.com	mixe	Français	USA
magiaenmarruecos.blogspot.com	intimiste	Espagnol	Espagne
makhoudjit.blogspot.com	social – politique	Arabe	USA
marocainsdalgerie.sosblog.com	social – politique	Français	Algérie
marouki.joeuser.com	social – politique	Français	USA
mlouizi.unblog.fr	littérature	Français	France
mohblog.blogspot.com	littérature	Français	USA
monagora.fr	humeur	Français	France
murmures.net	intimiste	Français	USA
myrtus.typepad.com	humeur	Anglais	USA
nowbi.over-blog.com	intimiste	Français	France
oef75.blogspot.com	littérature	Français	USA
poliquonautemarocain.blogspot.com	social – politique	Français	USA
purplemind-candysha.blogspot.com	intimiste	Français	USA
saad.amrani.free.fr/blog	intimiste	Français	France
sahara-libre.blogspot.com	social – politique	Français	USA
sebti.fr	mixe	Français	France
slimane.canalblog.com	humeur	Français	France
sonofwords.blogspot.com	intimiste	Français	USA
soumiaz.blogspot.com	intimiste	Anglais	USA
supertimba.skynetblogs.be	intimiste	Français	Belgique
taha.fr/blog	mixe	Français	France

Table 5.1: Liste des 47 blogs de la blogosphère marocaine tels que référencés dans l'Atlas e-Diasporas

de sa cartographie pour l'Atlas.

La Table 5.1 recense les blogs par préfixe d'URL, type de publication, langue et localisation. Mais situer géographiquement un site Web, quel qu'il soit, est toujours une étape sujette à discussion. Durant la création de l'Atlas, les chercheurs se sont servis, à la fois, d'indications présentes sur les sites eux même, et d'une méthode de localisation par adresse IP [19]. Ainsi, beaucoup de blogs sont hébergés aux USA, du fait principalement des plateformes choisies pour créer ces sites. Mais la combinaison de cette information, associée à la langue principale de chaque site (Table 5.1), nous donne tout de même une idée générale de la répartition géographique du réseau : partagé entre la France et les États Unis.

19. La recherche par IP renvoie le pays d'hébergement du site (là où sont situés les serveurs), pas le pays de résidence de son auteur.



La Figure 5.4 présente la blogosphère sous la forme d'un graphe, telle qu'elle fut cartographiée en 2008. La taille des nœuds y est fonction du degré de chaque blog.

Au cœur de la blogosphère, le blog politique *larbi.org* fait autorité (Figure 5.5) : sa position est centrale et son degré est le plus élevé du réseau. En 2008, il est élu blog marocain et politique de l'année par un jury d'internautes et inaugure, de fait, un cycle de marques de reconnaissance des blogueurs entre eux. Cette forme de mise en avant

FIGURE 5.5: *larbi.org* (rouge) dans l'e-Diaspora marocaine

des siens, pour promouvoir par extension l'ensemble du collectif, est monnaie courante dans la blogosphère [20]. On y cite volontiers les blogs qui nous semblent influents [21], ou ceux que l'on souhaite promouvoir [22].

La présence de certains membres sur le Web est parfois antérieure à la formation de la blogosphère, *7didane.org* par exemple en est déjà à son second site [23]. Et si *larbi.org* se propose, principalement, de commenter la vie politique marocaine à travers son regard de marocain émigré, les billets publiés sur les autres blogs sont divers et variés : allant de recettes de cuisines (*fatima-salma.over-blog.com*), à des récits de vie quotidienne (*murmures.net*), et en passant par des chroniques littéraires (*lailalalami.com*). Ainsi, la blogosphère se vie et se pense comme un environnement vibrant et dynamique où circulent des auteurs, des lecteurs et des visiteurs.

20. https://fr.wikipedia.org/wiki/Maroc_Web_Awards#Maroc_Blog_Awards_2008

21. L'auteur de *manal.over-blog.com* se réfère explicitement à *larbi.org* comme source d'inspiration (http://manal.over-blog.com/article-517669.html).

22. Les lecteurs de *larbi.org* citant *makhoudjit.blogspot.com* parmi d'autres blogs crées par des femmes (https://web.archive.org/web/20130319234735/http://www.larbi.org/post/2006/03/05/183-hommage-aux-femmes).

23. https://web.archive.org/web/20110810230023/http://7didane.blogspot.com/

...à un collectif éteint

En 2015,après une première enquête, (Khouzaimi, 2015) constate la disparition de 47 des 153 sites de l'e-Diaspora marocaine. La blogosphère est alors réduite à 17 membres actifs [24]. Par ailleurs, une série d'entretiens, enrichissant ce travail, nous apprend que les blogueurs toujours *actifs* ne sont pas tous de jeunes étudiants comme le pensait en premier lieu l'auteure. Nombre d'entre eux sont des quarantenaires ou des cinquantenaires qui ont choisi, malgré les évolutions du Web, de maintenir ouverts leurs sites sur la toile.

Mais une frontière entre étudiants de 2015 et anciens blogueurs se dessine : chez les étudiants, l'attachement en ligne à un passé vécu au Maroc se traduit au mieux par l'appartenance à certains groupes communautaires sur Facebook et Whatsapp [25], plus que par la consultation des anciens blogs d'émigrés, dont ils ne connaissent pas toujours l'existence d'ailleurs.

24. Pour comprendre le décompte de J. Khouzaimi, un blog actif est (selon sa méthodologie) un blog toujours en ligne.

25. Tout comme Telegram, l'application de messagerie Whatsapp permet de créer des salons de discussions virtuels à plusieurs, voir https://frama.link/JnjBqw8u

Début 2018, nous choisissons, à notre tour, de revenir en détail sur le cas des blogs de l'e-Diaspora marocaine. Pour ce faire, nous visitons chacune des 47 URL associées et classons les sites selon trois catégories :

1. vivant : le blog est en ligne et a été mis à jour au moins une fois au cours de l'année passée (courant 2017 donc)
2. abandonné : le blog est en ligne mais n'a pas été mis à jour récemment
3. mort : le blog n'est plus en ligne ou a été victime de cybersquatting (le blog n'est plus celui qu'il a été au moment de sa cartographie, il a été volé ou squatté)

La Figure 5.6 présente l'état [26] de la blogosphère, telle que nous la trouvons en 2018. Nous ne recensons plus que 5 blogs vivants (Fi-

26. La position des nœuds est conservée par rapport à la Figure 5.4

gure 5.6, (a)), 23 ont été abandonnés (Figure 5.6, (b)) et les 19 autres sont morts. Parmi les blogs toujours en vie, deux sont, en fait, des transformations de blogs cartographiés en 2008 mais déplacés depuis par leurs auteurs : *louladekhmissbatata.wordpress.com* devient *louladekhmissbatataprise2.com* en 2009 et *cabalamuse.wordpress.com* devient *amoroccandrifter.wordpress.com* en 2012. Le collectif en ligne de 2008 s'est éteint.

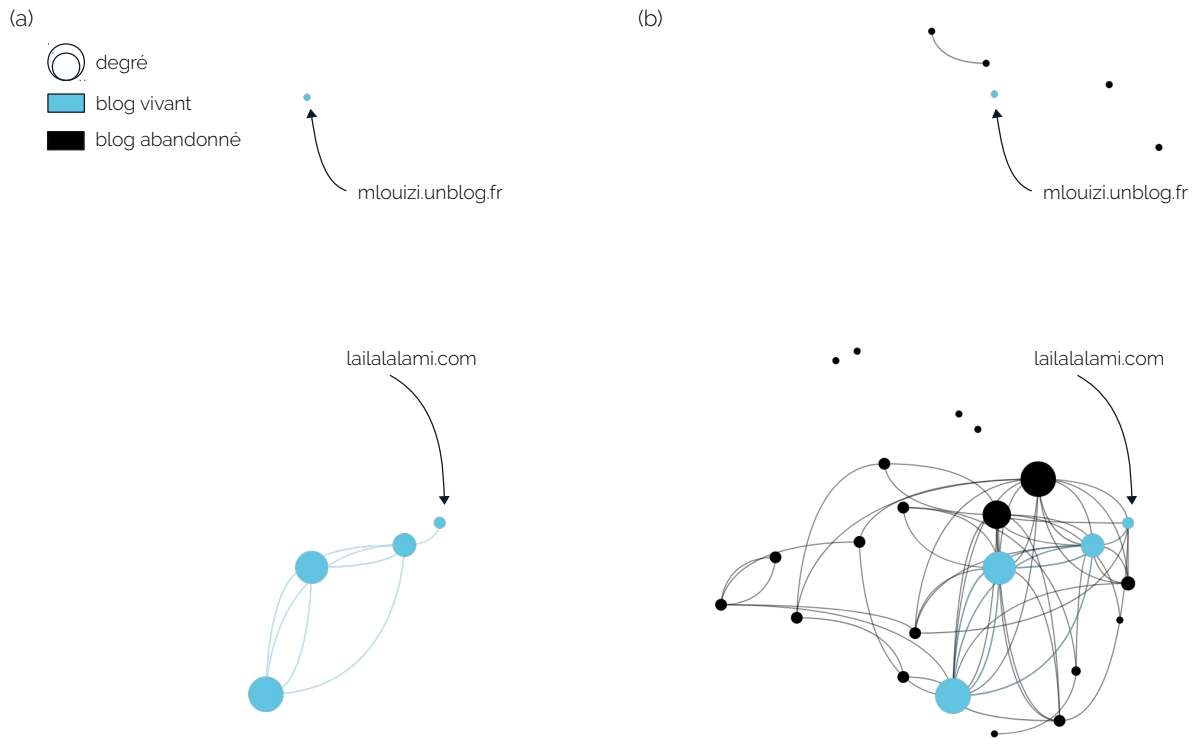


FIGURE 5.6: La blogosphère en 2018, (a) blogs vivants, (b) blogs vivants et abandonnés, la position est conservée

Définir l'espace d'exploration

Partant de cet état de fait, nous pouvons, en rassemblant les informations collectées, commencer à formuler quelques hypothèses. Nous savons qu'en 2015 certains étudiants marocains préféraient les groupes Facebook aux blogs d'émigrés pour maintenir le lien avec leur vie passée au Maroc (Khouzaimi, 2015). Ces groupes Facebook étaient, par ailleurs, déjà contemporains de la blogosphère au moment de sa cartographie. En effet, S. Marchandise a réalisé, dans le cadre de l'Atlas, une e-Diaspora (Marchandise, 2014) entièrement centrée sur la présence[27] des étudiants émigrés marocains au sein de groupes Facebook. De plus, les mécanismes des réseaux sociaux (Facebook, Twitter...) ne sont pas

27. `http://www.e-diasporas.fr/wp/marchandise.html`

si éloignés que ça des fonctionnalités démocratisées par les blogs en leur temps (Kwak *et al.*, 2010) : ils en reprennent certaines fonctionnalités (publication, création de liens, création de communautés...), en améliorent d'autres (vitesse de diffusion, intégration mobile...) et introduisent leurs propres règles (partage, retweet, flux d'informations et timeline...).

Arrivés à ce niveau de notre exploration, nous nommons maintenant **auteur** la personne ou l'ensemble de personnes ayant créé ou maintenu l'un des 47 blogs de la blogosphère marocaine. De même, nous nommons **lecteur** une personne qui visite, réagit ou laisse un commentaire sur l'un de ces sites.

Notre tâche d'exploration a ici pour but de comprendre, autant que faire ce pourra, le devenir d'un **collectif en ligne éteint**, c'est-à-dire une communauté pour laquelle peu ou plus aucune trace ne subsiste encore sur le Web vivant. Deux grandes hypothèses peuvent être formulées : **1)** les blogs ont purement et simplement disparu de la surface du Web, les auteurs qui les ont créés n'ont pas voulu ou pu donner suite à cette présence sur la toile **2)** les blogs se sont mués en une autre entité, migrant d'un territoire du Web (la blogosphère) à un autre (ici, les plateformes de réseaux sociaux). Nous chercherons donc, dans un premier temps, à identifier les indices archivés de cette possible mutation.

L'espace d'exploration est donc le suivant : un nombre connu et limité de sites Web observés pendant 10 ans, de 2008 à 2018. Grâce aux archives de l'atlas e-Diasporas, nous disposons de 4 années de collectes (2010-2014), que nous allons étendre à 6 années grâce aux dates d'éditions des fragments Web (2008-2014), la période 2014-2018 sera, elle, couverte par les collectages d'Internet Archive.

Pour commencer, nous partons de la liste des 47 sites de la blogosphère que nous fragmentons (Section 4.3) et envoyons dans notre moteur d'exploration (Section 2.2). L'idée est ici de réaliser une même requête plein texte sur deux champs différents de nos fragments Web. Cette requête, nous la construisons à partir de mots clés que nous pensons être en rapport avec les réseaux sociaux, tels que :

```
facebook, twitter, instagram, pinterest, youtube, flicker, medium, myspace, google+, googleplus, tumblr,
share, like, retweet, tweet, partager, aimer, social...
```

Ces mots-clés sont, ensuite, testés par rapport au contenu des champs `frag_text` et `frag_label` de chaque fragment. Les requêtes à `frag_text` permettent de détecter d'éventuelles références aux réseaux sociaux dans le corps d'un texte : par exemple, un blogueur parlant de Twitter

dans l'un de ses billets. Les requêtes à `frag_label` sont supposées capturer et retourner les nœuds HTML dont les labels[28] correspondant à l'un des mots-clés : par exemple, un bouton *partager sur facebook* présent sur une page Web. Nous cherchons donc, à la fois, du contenu textuel et des éléments de mise en forme.

28. Pour rappel, le label d'un nœud HTML est la concaténation de son nom, de son identifiant et de sa classe.

```
#### bouton like facebook chez larbi.org

<iframe id="iframe_like" name="fbLikeIFrame_0" class="social-iframe" scrolling="no"
src="http://www.facebook.com/widgets/like.php?
width=300&amp;show_faces=0&amp;layout=button_count&amp;href=http://www.larbi.org/post/2011/02/Morocco-Feb20-
Maroc-20Fev" title="facebook" width="500" height="81" frameborder="0">
</iframe>

#### liste de comptes sociaux chez 7didane.org

<ul>
   <li>
      <span>Sur twitter :</span>
      <span style="font-size: 11px; font-style: italic;"><a href="https://twitter.com/7didane">clique-
ici</a></span><br>
   </li>
   <li>
      <span>Sur Facebook :</span>
      <span style="font-size: 11px; font-style: italic;"><a href="https://www.facebook.com/7didane">clique-
ici</a></span><br>
   </li>
</ul>

#### widget twitter chez anasalaoui.com

<div class="textwidget">
<p style="text-align:center;"><a class="twitter" href="http://twitter.com/AnasAlaoui" title="Suivez-moi sur
Twitter - @AnasAlaoui" target="_blank">

#### badge flickr chez murmures.net

<div id="flickr-badge" class="widget widget_flickrbadge">
   <h2 class="widgettitle">Mes Photos</h2>
   <div class="flickrbadge" align="left">
      <!-- Start of Flickr Badge -->
      <table id="flickr_badge_uber_wrapper" cellspacing="5" cellpadding="0" border="0">
         <tbody>
            <tr>
               <td style="padding:0" class="flickr_badge_image" id="flickr_badge_image1" valign="center"
align="center"><a href="http://www.flickr.com/photos/kenzamurmures/561452263/"><img
src="http://farm2.static.flickr.com/1308/561452263_0b79c5fe3c_s.jpg" alt="A photo on Flickr" title="Bord du
lac" width="75" height="75"></a></td>
            </tr>
         </tbody>
      </table>
   </div>
   <!-- End of Flickr Badge -->
</div>
```

FIGURE 5.7: Exemples de fragments Web associés à divers réseaux sociaux

Migrer d'un territoire du Web à un autre

Trois types de fragments Web répondent à nos requêtes : **1)** des fragments de textes extraits d'articles ou de billets **2)** des boutons et *widgets like*, *share*, *follow me*... près-construits et mis à disposition des blogueurs par les réseaux sociaux **3)** un mélange de nœuds HTML non conventionnels, confectionnés par les auteurs eux même, comme système de redirection vers Twitter ou Facebook.

Ces derniers témoignent d'un age du Web balbutiant (2008-2010),

Figure 5.8: *eatbees.com* (rouge) dans l'e-Diaspora marocaine

où les auteurs de blogs devaient eux même *bricoler*, avec un peu de HTML et de CSS, des connexions vers leurs nouveaux comptes sociaux. Avec le temps, cette pratique s'est faite oublier, supplantée par l'arrivée d'éléments hypertextes standardisés et prêts à l'emploi. La Figure 5.7 donne à voir quelques exemples de ces fragments Web avec, en rouge, les parties ayant correspondu à nos requêtes.

Mais, en y regardant de plus prêt, ces fragments ont d'autres informations à nous offrir : il est possible de trouver, au cœur du HTML (Figure 5.7, bleu), des noms de comptes sociaux liés aux blogs. En effet, lorsque l'on souhaite s'inscrire sur une plateforme de réseaux sociaux, il faut d'abord et avant tout se créer un **avatar** : le *@user* de Twitter ou le *pseudo* de Facebook par exemple. Ces avatars deviennent alors l'incarnation de notre identité sur les réseaux. Ce faisant, nous constituons rapidement une liste d'association entre blogs passés et noms de comptes sociaux. À la fin de cette étape, nous identifions 20 blogs, parmi les 47 de départ, liés à un ou plusieurs réseaux sociaux.

Afin d'élargir nos résultats, nous faisons ici l'hypothèse que les avatars, pseudos et autres noms d'utilisateurs peuvent être réutilisés par un même auteur, d'une plateforme à l'autre, moyennant une légère modification. À titre d'exemple, le blogueur *eatbees.com* (Figure 5.8) se nomme *@eatbees* sur Twitter et *eatbees* sur Medium. Aussi, nous normalisons les 20 avatars déjà extraits avec un petite opération de stemming [29].

29. Les majuscules deviennent minuscules et l'on ne conserve que la racine des pseudonymes (l'avatar moins les deux derniers caractères), voir la Figure 3.10.

L'objectif de cette nouvelle phase est maintenant d'explorer le Web vivant à la recherche de comptes sociaux (Youtube, Flickr, Medium...) liés, de prêt ou de loin, à l'un de ces 20 pseudonymes. Mais, cette tâche ne peut malheureusement pas être automatisée. Si la recherche par fragment Web, dans notre moteur, nous a permis d'identifier rapidement quelques indices, c'est maintenant au chercheur d'entrer en scène. La fouille (via des moteurs de recherche : Google, Bing...) et la vérification des informations tirées du Web vivant se font à la main. Ainsi, nous récoltons un total de 33 comptes sociaux actifs différents, liés par leurs avatars aux 20 blogs ayant correspondu aux requêtes par mots clés initiales.

La Figure 5.9 présente la répartition de ces 33 comptes, entre anciens blogs (gauche) et nouveaux réseaux sociaux (droite). Twitter et Facebook dominent cette liste, ils agrègent à eux seuls 23 comptes. L'usage de Youtube, Pinterest, etc., reste plus confidentiel. Bien sûr, ce classement ne peut se prévaloir d'une quelconque exhaustivité, par construction cela lui est impossible. En effet, si l'un des auteurs des blogs a préféré utiliser un pseudonyme entièrement nouveau, sans en faire mention dans les pages de son ancien site, alors nous ne sommes pas ici en mesure de le retrouver.

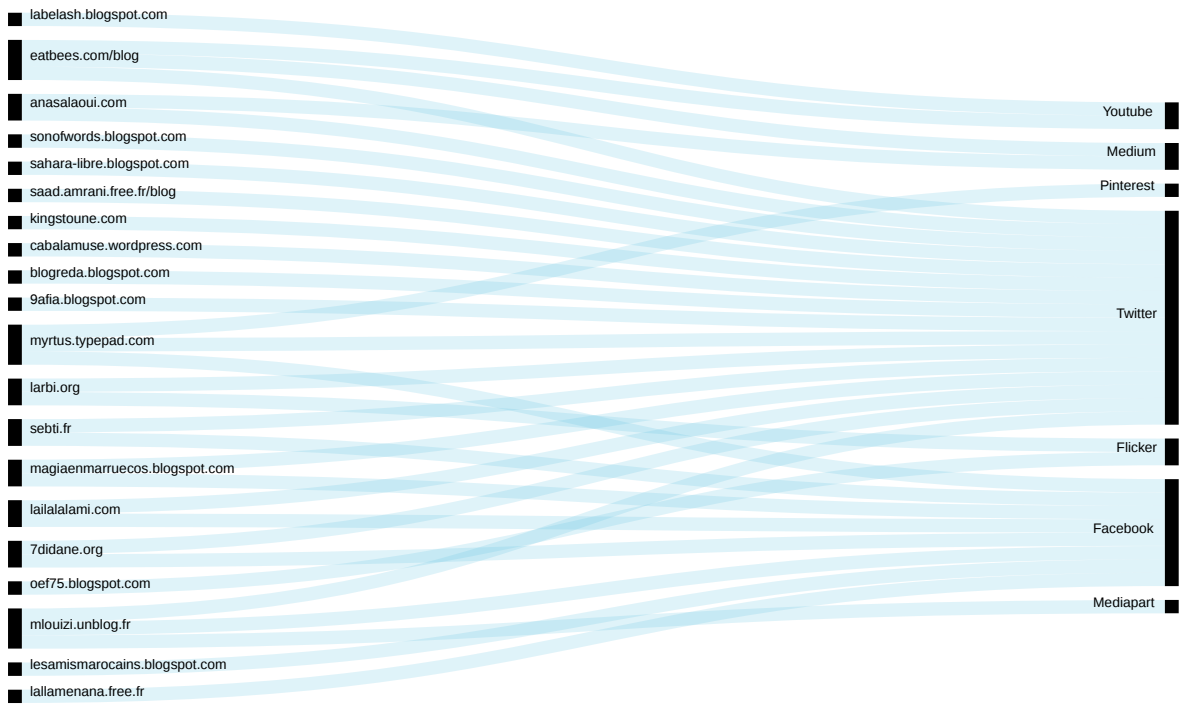


FIGURE 5.9: Liste des blogs de 2008 liés à des réseaux sociaux en 2018

Si l'on s'en tient aux seuls blogs subsistants (Figure 5.6), la blogosphère est bel et bien un collectif éteint. Crées entre 2005 et 2008, la quasi totalité de ces sites ont progressivement été fermés ou laissés à l'abandon par leurs auteurs. La blogosphère en tant qu'espace de communication n'a pas réussi à durer, du fait même de ceux qui la constituaient. Mais en lieu et place d'une disparition, c'est en fait à une transformation que nous avons ici affaire. Les blogs se sont mués en une multitude de comptes sur les réseaux sociaux et au cours de cette migration, les anciens blogueurs ont choisi de conserver leurs identités en ligne. En gardant leurs avatars, ils leur a été plus facile de se retrouver et de se reconnecter après coup.

En effet, en procédant à un rapide scraping des pages Facebook et Twitter, nous pouvons récolter l'ensemble des liens d'amitié et d'intérêt[30] qui sont aujourd'hui tissés entre ces différents comptes. La Figure 5.10 se veut, ainsi, une vue de la blogosphère, après sa mutation et telle que nous la redécouvrons en 2018. Nous plaçons les comptes sociaux là où se situaient leurs anciens blogs respectifs (Figure 5.6) et nous traçons entre eux les nouveaux liens de citations (liens rouges).

30. Lien entre *followers* sur Twitter

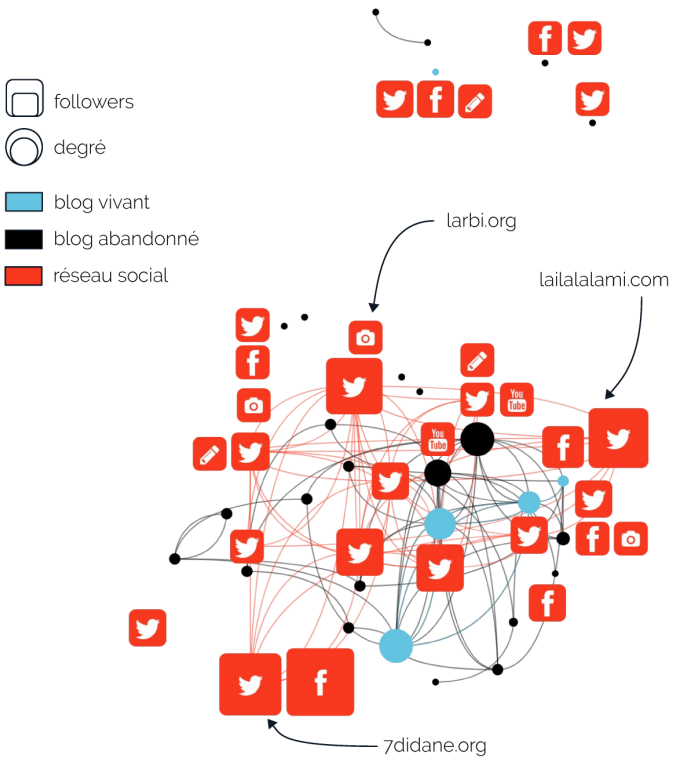


Figure 5.10: La blogosphère en 2018, les emplacements des blogs sont conservés

Sur la nouvelle blogosphère, l'expression en ligne se fragmente et se spécialise par type de réseau et de médium. Certains, comme *7didane.org*, choisissent d'avoir un compte Facebook en parallèle d'un compte Twitter. D'autres utilisent Youtube ou Flickr pour publier des photos ou des vidéos comme *larbi.org*. On peut également observer l'utilisation conjointe de Twitter et de Medium[31] (ou d'un blog Médiapart) sur lequel l'auteur va choisir d'écrire un texte réfléchi et structuré avant de le promouvoir à sa communauté via Twitter, comme *eatbees.com*.

31. Plateforme de partage de billets se voulant longs et argumentés.

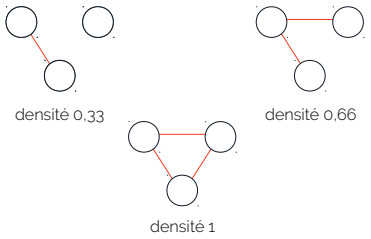


Figure 5.11: Densité d'un graphe

Mais la migration vers les réseaux sociaux n'est pas, pour autant, synonyme d'éclatement ou d'isolation des anciens membres de la blogosphère. Prenons, ainsi, le cas des 16 auteurs ayant désormais un compte Twitter. Il est possible de comparer la densité du sous graphe qu'ils formaient en 2008, avec celle de leur graphe d'intérêt, tel qu'il se trouve sur Twitter en 2018.

En théorie des graphes, un graphe est dit **dense** si le nombre de ses liens est proche du nombre de liens maximum possibles. Dans le cas d'un graphe $G = (V, E)$, avec V représentant l'ensemble de ses nœuds et E l'ensemble de ses liens, sa densité D est définie comme le rapport suivant :

$$D = \frac{2 \times |E|}{|V| \times (|V| - 1)}$$

Si les nœuds ne sont pas du tout connectés entre eux, alors la densité vaudra 0, mais s'ils sont tous liés celle-ci vaudra 1 (Figure 5.11). Entre 2008 et 2018, la densité du sous graphe des utilisateurs de Twitter, anciennement blogueurs, passe ainsi de 0.16 à 0.24. L'aspect *communauté* de l'ancienne blogosphère est conservée par le passage sur Twitter. Notons néanmoins que la différence n'est pas véritablement significative, le graphe Twitter étant plus petit que le graphe des blogs.

FIGURE 5.12: *lailalalami.com* (rouge) dans l'e-Diaspora marocaine

Communautés de lecteurs et identité en ligne

Sur la Figure 5.10, la taille de chaque nœud social est fonction de l'étendue de sa communauté de followers (pour Twitter) ou d'amis (pour Facebook). À titre d'exemple, *7didane.org* est suivi par 43 512 sur Twitter et a 141 947 amis sur Facebook. De son côté, *lailalalalmi.com* (Figure 5.12) dépasse les 35 000 followers sur Twitter.

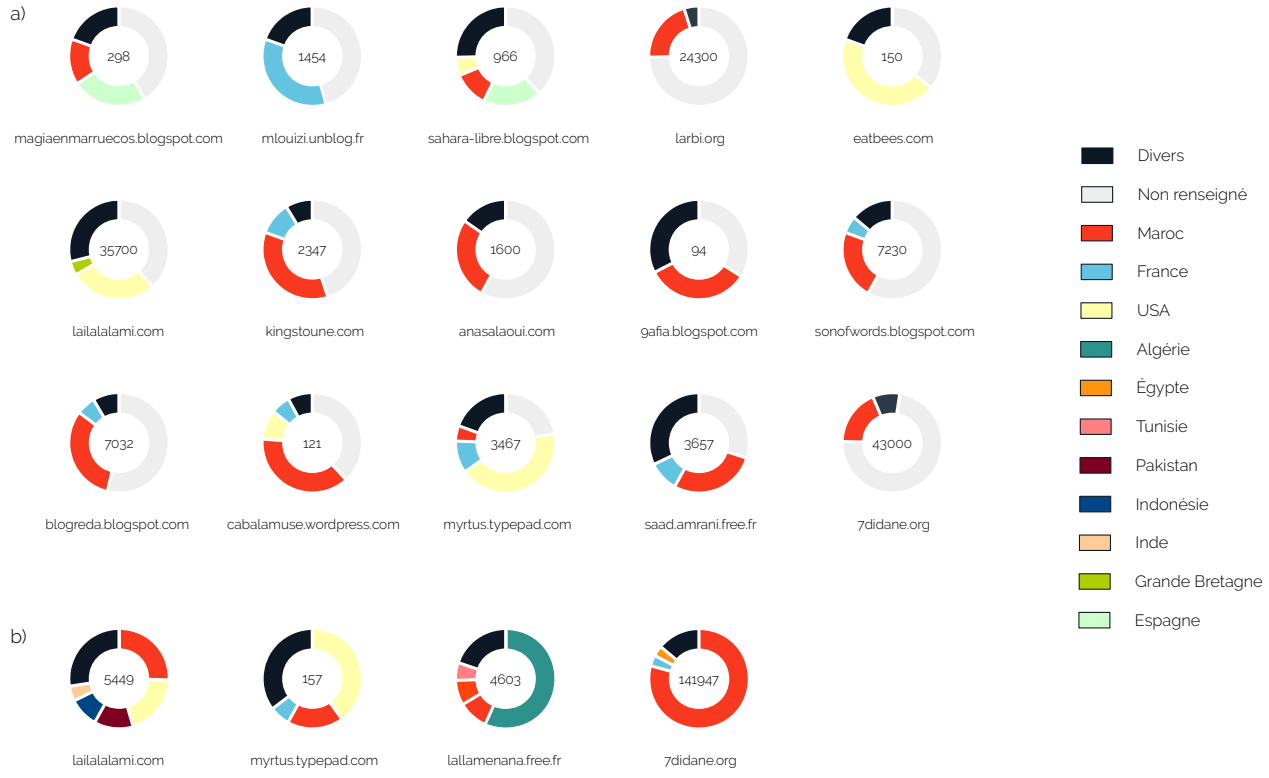


FIGURE 5.13: Répartition géographique des followers Twitter (a) et des amis Facebook (b)

À l'ère des réseaux sociaux, l'influence d'un auteur est souvent liée au nombre de lecteurs ou de visiteurs auxquels il est capable de s'adresser.

La communauté devient une audience. De fait, la dynamique interne à la blogosphère évolue [32] : *larbi.org* n'est plus hégémonique, *7didane.org* et *lailalalalmi.com* le dépassent désormais. Par ailleurs, l'origine des followers Twitter et des amis Facebook peut nous renseigner sur le caractère diasporique de la communauté des lecteurs et des visiteurs de 2018. D'où lisent celles et ceux qui suivent les anciens blogueurs sur les réseaux sociaux ?

32. Basé sur un crawl de Twitter de Janvier 2018

Pour répondre à cette question nous suivons une double approche. La première est adaptée à Facebook, la seconde à Twitter : **1)** nous utilisons l'application Netvizz qui associe à chaque personne ayant aimé une page facebook donnée, une situation géographique [33] **2)** nous développons un programme analysant les pages de l'ensemble des followers des comptes Twitter [34].

33. Cette information est issue des méta-données renseignées lors de l'inscription à Facebook ou de la modification de sa page personnelle (`https://apps.facebook.com/netvizz/`)

34. L'API Twitter est limitée par un système de quota temporel nous obligeant à créer ici notre propre scraper.

En effet, le champ `location` [35] de ces pages peut, s'il est renseigné, renfermer une information géolocalisée. Nous choisissons, par hypothèse, de faire correspondre cette donnée au lieu depuis lequel Twitter est utilisé. Nous croisons, ensuite, la valeur de ce champ avec la base d'entités géographiques Geonames [36] afin de désambigüiser l'information extraite.

35. `https://help.twitter.com/fr/using-twitter/tweet-location`

36. `http://www.geonames.org/`

Ces deux méthodes ne sont pas exhaustives, mais elles nous donnent à voir une coloration et une tendance générale. Nous essayons, néanmoins, de limiter au maximum les faux positifs dans nos résultats. Ainsi, la Figure 5.13 présente pour chaque compte, la répartition géographique (si elle est connue) de ses followers Twitter ou amis Facebook. En pratique, *7didane.org* est principalement suivi sur Facebook par des personnes situées au Maroc (82%) mais aussi par quelques Français (2%) et des Égyptiens (2%). Toujours sur Facebook, *lailalalami.com* continue de s'adresser à des Marocains (24%), à des Américains (15%) et à des Pakistanais (8%). Son compte Twitter semble plutôt attirer un public d'anglophones : des Américains, des Anglais, des Canadiens... Dans l'ensemble, au moins 19 pays [37] sont ainsi représentés, témoignant du caractère géographiquement dispersé de cette communauté de lecteurs.

37. Pour clarifier la Figure 5.13 certains pays faiblement représentés n'apparaissent pas, voici la liste complète : Algérie, Maroc, Égypte, Tunisie, Libye, France, Chine, Espagne, Pays-Bas, USA, Pakistan, Indonésie, Bangladesh, Nigeria, Inde, Belgique, Grande Bretagne, Canada et Corée du Sud.

Comme nous l'expliquions plus tôt, les auteurs ont su conserver leur identité en ligne en réutilisant les mêmes pseudonymes, d'une plateforme à l'autre, et ce depuis le temps de leurs premiers blogs. Nous faisons maintenant l'hypothèse de retrouver le même comportement chez leurs lecteurs. En effet, un blog est un médium tourné vers l'intéressement d'un public donné. Aussi, lorsqu'un lien fort se crée entre un auteur et ses lecteurs, il nous semble raisonnable d'avancer que ces derniers peuvent chercher à conserver à tout prix cette connexion. Nous supposons donc que les lecteurs ont réutilisé, eux aussi, leurs anciens avatars sur Twitter, pour suivre les auteurs qu'ils supportaient au temps des blogs.

Pour tester cette hypothèse, nous nous focalisons sur le site *larbi.org* et sur le lien qu'il entretenait avec les personnes commentant ses articles, lorsqu'il faisait encore autorité (Figure 5.4). Nous définissons alors un modèle de requête capable de nous retourner l'ensemble de ces commentaires archivés. Il s'agit ainsi de cibler des fragments Web ayant la forme suivante (champ `frag_type`) : un nom d'auteur, une date et un contenu textuel publié sous l'article principal d'une page. Ce faisant, nous obtenons 4177 résultats dont nous ne conservons que les noms d'utilisateurs.

Nous croisons cette liste à celle des followers actuels de *larbi.org* sur Twitter [38] via une égalité stricte pour ne pas multiplier les faux positifs. Au final, nous obtenons une borne inférieure de 647 lecteurs ayant, à priori, choisi de suivre *larbi.org* sur Twitter, soit près de 15% de son audience passée. Rapportés à leur nombre de commentaires publiés, ces lecteurs représentent au final 26% des réactions postées sur le site. Les auteurs des blogs ont donc été accompagnés durant leur migration vers les réseaux sociaux. En conservant leur identité en ligne, ils ont emporté, à leur suite, la part la plus fidèle de leurs lecteurs et de leurs communautés.

38. `https://twitter.com/larbi_org/followers`

Le Printemps arabe comme un moment clé ?

Avant de clôturer leurs blogs, seuls 6 auteurs ont déposé sur leur site un message d'adieu, ou d'explication [39]. Passé un certain temps, les blogs ne sont plus mis à jour et finissent alors par disparaitre sans crier gare. Le blogueur *7didane.org* est de ceux qui ont souhaité indiquer à leur communauté qu'il partait. En avril 2012 la page d'accueil de son blog vire au noir (Figure 5.14, a), une adresse mail est laissée pour ceux qui souhaiteraient garder le contact. Si *7didane.org* n'a pas donné plus d'explication, *larbi.org*, au contraire, argumente sa volonté de clôturer son blog par un dernier article (Figure 5.14, b).

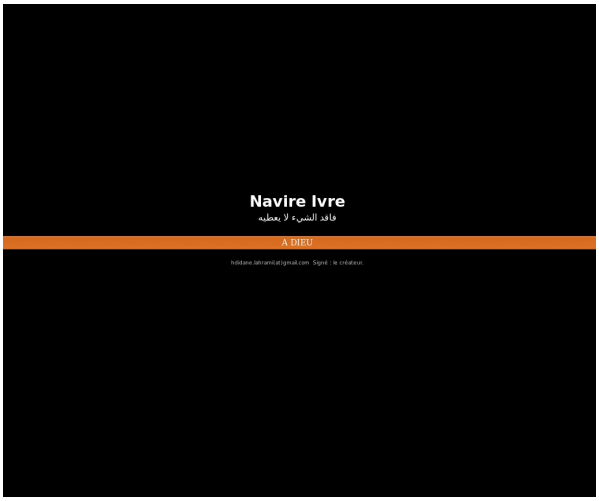
39. Pour *7didane.org* `https://web.archive.org/web/20120415100250/http://www.7didane.org:80/`, pour *larbi.org* `https://web.archive.org/web/20140119233126/http://www.larbi.org/post/2013/12/A-nos-apprentis-dictateurs-redacteurs-du-Code-du-Numerique`

Fin 2013, il y met en garde ses lecteurs contre l'application à venir, au Maroc, du *projet de loi formant Code numérique*. Ce texte, selon lui, porte une atteinte grave au droit de citation hypertexte et, de manière plus générale, vise à limiter la liberté d'expression sur le Web [40]. Quelques semaines après, sans rien ajouter de plus, l'auteur décide de fermer son blog. Il ne s'exprime aujourd'hui plus que par son compte Twitter [41]. Ce dernier billet témoigne d'un fort attachement aux libertés fondamentales et d'une crainte de voir le Web, comme espace d'expression, être de plus en plus censuré. Mais le contenu de ce texte peut être désormais mis au regard d'un autre article.

40. Voir `https://www.esi.ac.ma/Dossiers/20131217031227.pdf`

41. `https://twitter.com/Larbi_org`

(a)



(b)

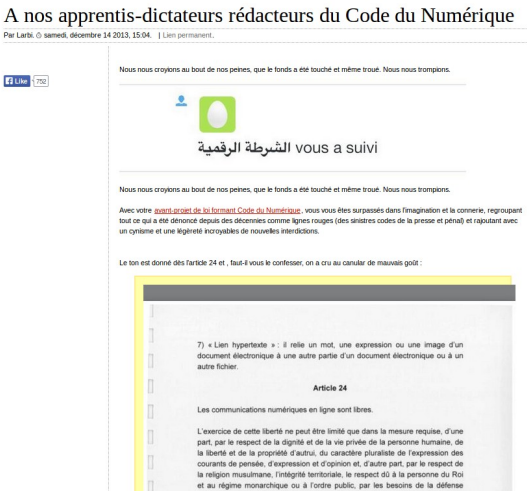


FIGURE 5.14: Annonce de la fermeture de *7didane.org* (a) et article précédant la clôture de *larbi.org* (b)

En effet, nous n'avons pas encore parlé des fragments de billets ayant, plus tôt, correspondu à notre première liste de mots-clés de réseaux sociaux. Parmi ces résultats, les premiers fragments Web archivés faisant mention de Twitter chez *larbi.org* et *7didane.org* attirent particulièrement notre attention, la Figure 5.15 transcrit ces morceaux de textes.

```
#### chez larbi.org, 14 février 2011

« Non, le printemps arabe ne peut pas s'arrêter aux frontières du maroc - Revivez la journée de manifestations du 20 Février »
Morocco #Feb20 Maroc #20Fev lundi, février 14 2011
Morocco #Feb20 Maroc #20Fev
En direct de Tunisie 14/01/2011
En direct de Tunisie La révolution en Marche
En direct de Twitter Join the conversation

#### chez 7didane.org, 16 juin 2009

Jun 16
L'épiphénomène Twitter (droit de réponse)
Exactement une semaine après mon billet sur Twitter (L'épiphénomène Twitter), et suite à une intervention de la maison blanche, le site me
répond personnellement sur leur blog. Voici le texte :

    Down Time Rescheduled (cliquez-ici)
    A critical network upgrade must be performed to ensure continued operation of Twitter. In coordination with Twitter, our network host
had planned this upgrade for tonight. However, our network partners at NTT America recognize the role Twitter is currently playing as an
important communication tool in Iran. Tonight's planned maintenance has been rescheduled to tomorrow between 2-3p PST (1:30a in Iran).
    Our partners are taking a huge risk not just for Twitter but also the other services they support worldwide—we commend them for being
flexible in what is essentially an inflexible situation. We chose NTT America Enterprise Hosting Services early last year specifically
because of their impeccable history of reliability and global perspective. Today's decision and actions continue to prove why NTT America is
such a powerful partner for Twitter.

Hat off to the youth of iran !
```

FIGURE 5.15: Premiers fragments Web archivés faisant mention de Twitter chez *larbi.org* *7didane.org*

En 2009, le blogueur *7didane.org* dit suivre [42], via Twitter, le déroulé des manifestations [43] qui secouent alors l'Iran. Cette année là, la contestation des élections du mois de Juin provoque de nombreux mouvements de protestations à Téhéran et dans le reste du pays. Mais surtout, les manifestations trouvent une résonance inédite sur le Web. L'information étant contrôlée par les médias gouvernementaux, les citoyens s'emparent de Twitter pour diffuser leurs témoignages, leurs photos et leurs vidéos. Cette vague de manifestations porte encore aujourd'hui le nom de *Révolution Twitter*.

42. https://web.archive.org/web/20090627012354/http://www.7didane.org:80/2009/06/16/1453/

43. https://fr.wikipedia.org/wiki/Soulèvement_postélectoral_de_2009_en_Iran

Chez *larbi.org*, Twitter fait pour la première fois son apparition dans un billet [44], au moment où le Printemps arabe [45] commence à se propager au Maroc, début 2011. Dans cet article *larbi.org* propose de suivre puis de revivre, via un hashtag dédié (Figure 5.15, bleu), la manifestation marocaine du 20 Février 2011.

44. https://web.archive.org/web/20110319191709/http://www.larbi.org:80/post/2011/02/Morocco-Feb20-Maroc-20Fev

45. https://fr.wikipedia.org/wiki/Printemps_arabe

Pour ces deux auteurs, Twitter n'a pas véritablement été découvert dans le contexte très précis de ces deux moments révolutionnaires, *larbi.org* et *7didane.org* (Figure 5.16) ayant tous deux ouverts leurs comptes Twitter en 2008. Mais, les manifestations Iraniennes et le Printemps Arabe ont pu jouer un rôle de révélateur, quant à un nouvel usage possible de Twitter, chez des personnes sensibles aux questions de libertés et de démocraties. À leurs yeux, Twitter est peut être, à ce moment là, devenu un nouvel espace d'expression libre sur la toile.

Figure 5.16: *7didane.org* (rouge) dans l'e-Diaspora marocaine

Nous pouvons ici émettre l'hypothèse que le Printemps Arabe, dans le cas particulier du Maroc, s'est inscrit dans un faisceau complexe d'événements qui ont, à un moment donné, influencé le devenir de la blogosphère, voire du reste de l'e-Diaspora marocaine. Bien sur, à l'échelle de nos archives, ces deux textes sont des détails statistiques, mais ils sont suffisamment intéressants pour nous mettre maintenant sur la piste de la manifestation du 20 février 2011 et de ce hashtag `#20Fev`. Cet événement a-il-pu avoir un impact visible sur d'autres espaces de notre corpus ?

5.3 Un soulèvement en ligne éphémère

En Section 4.5, partant d'une démonstration technique, nous avions commencé à entrevoir le contenu des archives du forum *yabiladi.com*. Notre système de détection d'événements nous révélait un pic de conversations portant sur la démission de l'ancien dirigeant tunisien Z. Ben Ali, comme un trait d'union entre un fait du Web et un bouleversement du réel. En effet, cette démission marque l'un des moments forts de la vague de contestations populaires qui a déferlé, fin 2010, sur l'ensemble du monde Arabe.

Dans cette thèse, nous tenons à insister sur la nécessité d'identifier des points d'entrée et de trouver des indices pour orienter nos explorations des archives Web. Le corpus marocain à lui seul est tellement vaste qu'il convient selon nous de s'attacher d'abord à l'étude de moments singuliers, bien définis et cadrés, sur lesquels nous sommes capables, à notre échelle et avec nos outils, de porter un regard autant quantitatif que qualitatif.

En cela, la manifestation du 20 février 2011 révélée dans la section précédente nous semble être un bon fil conducteur, pour tenter de comprendre la manière dont les utilisateurs de *yabiladi.com* ont pu réagir au Printemps arabe. Peut-on remonter les traces archivées d'un potentiel soulèvement du forum en amont ou en aval de cette manifestation ? Ses membres se sont-ils organisés d'une quelconque manière ? Ou ont-ils regardé passer les protestations sans y prendre part ?

Figure 5.17: *yabiladi.com* (rouge) dans l'e-Diaspora marocaine

Yabiladi.com : une porte d'entrée sur la diaspora

Le site Web *yabiladi.com* se présente donc comme le candidat idéal pour cette nouvelle exploration de notre corpus d'archives. Son rôle est central dans l'organisation de l'e-Diaspora marocaine (Section 1.3), puisqu'il fait le lien entre les trois grands clusters du réseau. C'est un *hub*, un **pont** entre des constellations de sites qui ne se parlent pas (Figure 5.17).

Site hybride, mélangeant forum de discussion en ligne, actualités, petites annonces et rencontres amoureuses, *yabiladi.com* fait figure de dinosaure à l'échelle du Web. Créé à la toute fin de l'année 2001 par un jeune ingénieur marocain travaillant en France [46], le site s'inscrit dans la longue tradition des portails diasporiques en ligne. Dans son analyse de l'e-Diaspora Indienne [47], E. Leclerc décrit avec précision le rôle de ces vastes sites protéiformes. La génèse de *yabiladi.com* est similaire à celle de ces sites indiens apparus au début des années 2000 sous l'impulsion d'une diaspora influencée par le succès de certains de ses membres, nouveaux entrepreneurs de l'informatique faisant carrière aux USA (Leclerc, 2012).

Un portail Web peut ainsi se voir comme une **porte d'entrée** ou un relais focalisé sur une thématique, une zone géographique ou régionale, une ethnie, une caste, un culte... et incarnant un double rôle : celui de l'intermédiaire entre une masse d'internautes et des grappes de sites inféodés à son autorité et celui du focalisateur d'une ou plusieurs communautés dont il se considère comme le représentant.

46. voir `http://lavieeco.com/news/portraits/mohamed-ezzouak-le-mre-qui-a-lance-yabiladi-com-6596.html`

47. `http://e-diasporas.fr/wp/leclerc.html`

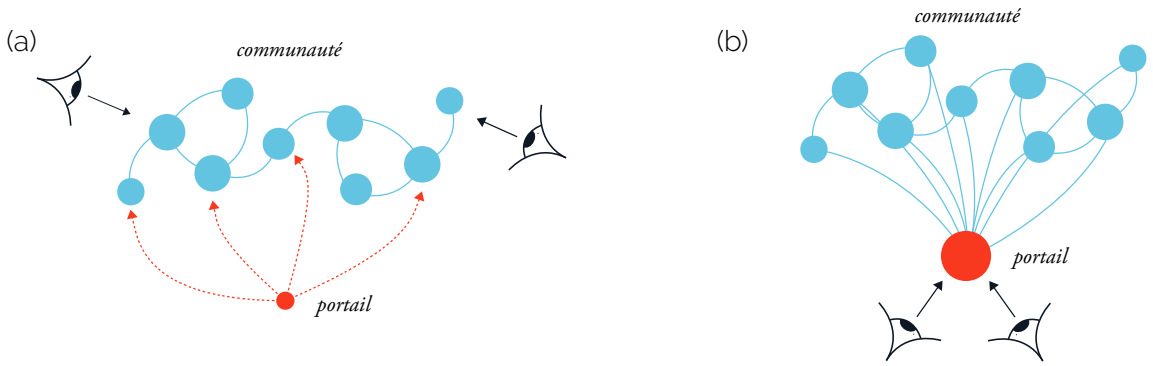


FIGURE 5.18: Dynamique de création d'un portail Web autour d'une communauté de sites

Un portail Web est donc dans la recherche constante d'une position d'autorité. Il commence par se mettre au service de sites existants en leur servant d'annuaire, en les pointant (Figure 5.18, a). Puis, une fois que les internautes l'ont identifié comme porte d'entrée unifiée, le portail peut prospérer sur le Web, fort du monopole acquis et bascule alors dans la seconde phase de son existence : la conservation de sa position et de son audience (Figure 5.18, b). Tout l'enjeu est alors de ne plus être seulement un site par lequel on passe, mais un espace où l'on reste. Pour ce faire, les portails s'adaptent aux exigences des internautes et proposent de nouveaux services (forum, petites annonces, radio, tv...), voire deviennent éditeur ou hébergeur de contenu. Un portail diasporique peut ainsi chercher à créer, en son sein, un véritable environnement dédié aux rencontres et aux discussions entre émigrés, un lieu de partage et d'entre-aide.

Mais le lien qui nait entre l'internaute émigré et son portail Web devenu espace, ne suffit pas à comprendre la longévité de ces sites. Eux qui, contrairement aux blogs, ont su résister et s'adapter à l'arrivée des réseaux sociaux. Pour continuer à être la vitrine culturelle de sa communauté, le portail doit y insuffler du dynamisme en testant toutes sortes de nouvelles technologie créatrices de lien. De fait, un portail Web, peut être vu comme un parfait témoin de l'évolution du Web et de ses outils. En cela, la genèse de *yabiladi.com* est tout à fait exemplaire. En interrogeant nos archives et en les associant aux collectages d'Internet Archive, nous créons la timeline de la Figure 5.19 qui présente les transformations successives de *yabiladi.com* au cours de son histoire.

Yabiladi signifie littéralement *Oh mon pays* (Ya biladi). Mis en ligne fin 2001, le portail inaugure, en 2002, une toute nouvelle section, un forum [48] contenant quatre espaces de discussions *Général*, *Femmes marocaines*, *Investir au Maroc* et *La corruption*. Il se présente alors comme le « *Le premier portail destiné aux Marocains Résidant à l'Etranger (RME) !* » mettant en avant un annuaire complet et éditorialisé (classement par

48. `https://web.archive.org/web/20020615201721/http://www.yabiladi.com:80/forum/list.php?f=4`

catégories) d'adresses Web marocaines [49], ainsi que quelque rubriques mélangeant informations et divertissements.

49. https://web.archive.org/web/20020531015840/http://www.yabiladi.com:80/

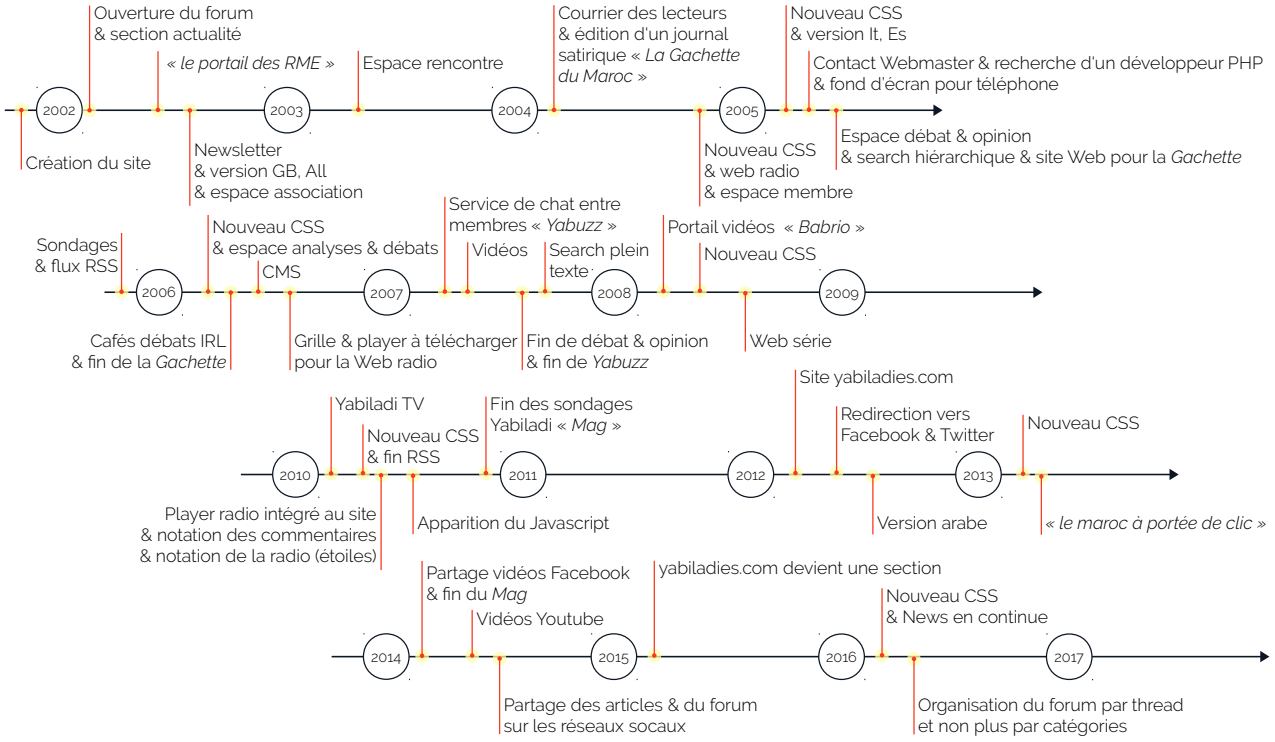


Figure 5.19: Timeline des évolutions successives de *yabiladi.com*

Suivre le fil des évolutions de *yabiladi.com*, c'est remonter l'histoire du Web dans son intégralité. Avant 2005, le site est maintenu par sa propre équipe de développeurs PHP, puis dès 2006 un premier CMS est utilisé pour moderniser le forum [50] et des éléments dynamiques de Javascript sont progressivement intégrés à partir de 2010. Les flux RSS [51] sont ajoutés dès 2005 mais disparaissent 5 années plus tard ; le search hiérarchique (à la Yahoo) est lui remplacé en 2007 par un search plein texte (à la Google). Et il faut attendre 2012 pour voir une redirection se faire vers les comptes Twitter et Facebook de *yabiladi.com*, puis patienter jusqu'en 2014 pour pouvoir partager du contenu sur les réseaux sociaux. Des technologies maisons sont testées puis abandonnées au profit de solutions extérieurs : un service de chat *Yabuzz*, un player radio intégré [52] et une plateforme de partage vidéo *Babrio* [53].

À plusieurs reprises, le portail tente de proposer du contenu éditorialisé en complément de ses seuls articles Web. Une Web radio, une Web télé et une Web série [54] sont lancées. Deux magazines sont publiés : le premier, un journal satirique et politique *La Gachette du Maroc* [55]

50. https://web.archive.org/web/20070406192841/http://www.phorum.org:80

51. https://web.archive.org/web/20051217073725/http://www.yabiladi.com:80/rubrik/rss.php

52. https://web.archive.org/web/20060717190039/http://www.yabiladi.com:80/radio/download.php

53. http://babrio.com/

54. https://web.archive.org/web/20081220014828/http://www.yabiladi.com:80/webserie/, https://web.archive.org/web/20101202113409/http://yabiladi.com/tv/, https://web.archive.org/web/20050411011853/http://www.yabiladi.com:80/radio/

55. https://web.archive.org/web/20050831073003/http://www.yabiladi.com:80/magazine/

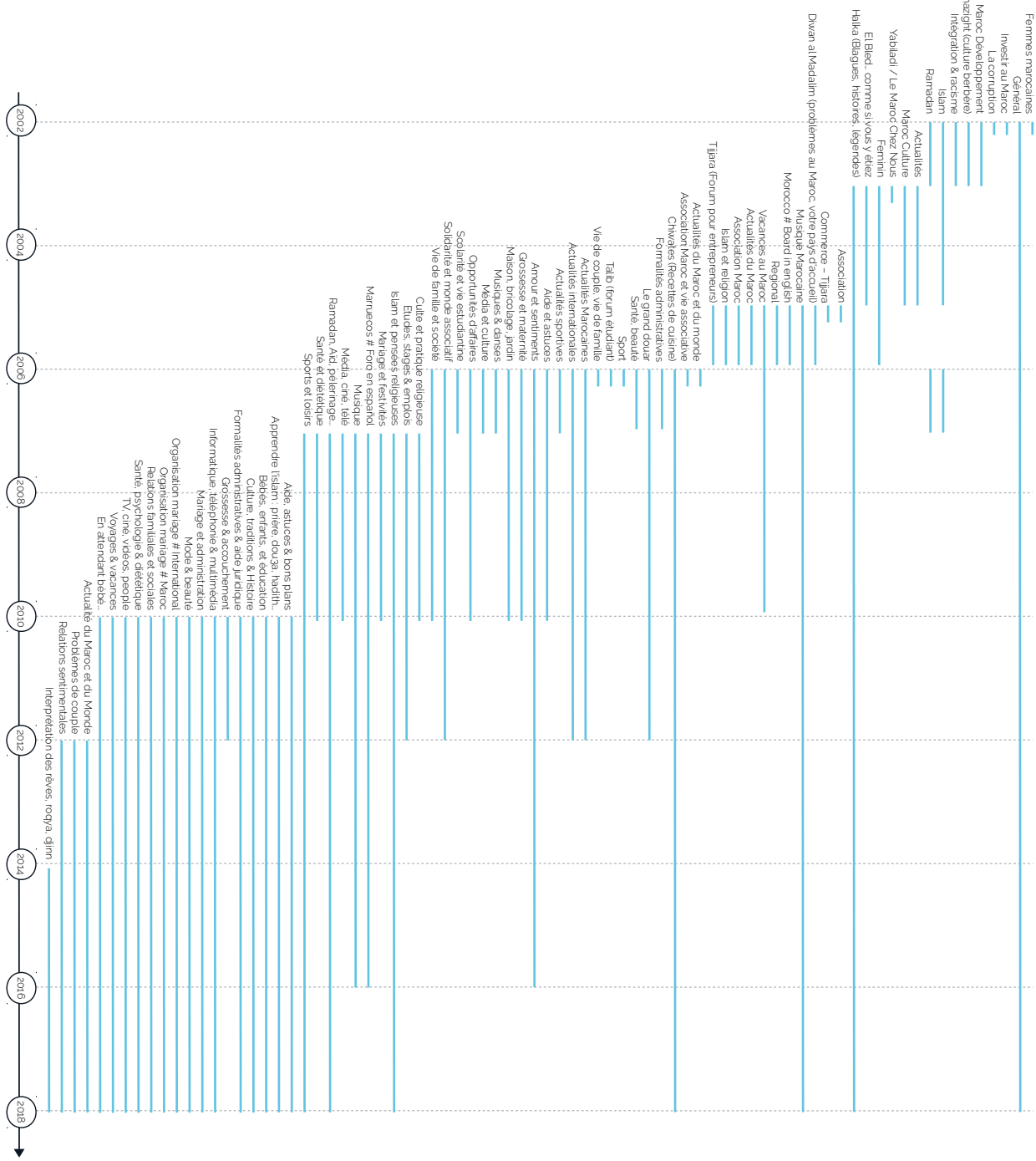


FIGURE 5.20: Évolution des catégories du forum de *yabiladi.com*, par date d'apparition

ne durera que 2 ans ; le second *Yabiladi Mag* plus orienté business et tourisme [56] sera édité pendant 4 années.

Espace de rencontre en ligne [57], *yabliadi.com* propose à ses membres de se croiser en dehors du site, via des cafés-débats [58]. Mais ces moments d'échanges ne durent pas et les espaces d'analyses et d'opinions [59], présents depuis les débuts du portail, sont peu à peu fermés, remplacés par des sections plus généralistes : *Divertissement*, *Vie pratique*, *Famille*...

Le site évolue et se transforme. Six changements visuels et structurels majeurs sont opérés successivement (Figure 5.19, *nouveau CSS*), impactant notamment le forum. Si l'organisation locale de cette section reste la même (les messages postés y sont triés par date de publication et associés à un sujet unique, formant ce que l'on appelle des *threads* [60]), sa structure générale, quant à elle, est régulièrement bouleversée. Les catégories de discussions qui prédéfinissent l'ordre et l'agencement des threads changent, disparaissent ou sont réintégrées au grès des nombreuses mutations du site.

En 16 années d'existence, 81 catégories successives viennent, une à une, donner une nouvelle orientation au forum (Figure 5.20). Des thématiques entières apparaissent (*Grossesse, Maternité, Vie de Famille*...) ou perdent en influence (*Opportunité au Maroc, Forum étudiant, Solidarité & associations*...). Seules les catégories généralistes et légères : *Général*, *Halka* (c.-à-d., blagues) et *Vacances* restent inchangées, alors que d'autres sont en perpétuelles mutations : *Islam* devient *Islam et religions*, puis *Islam et pensées religieuses* et enfin *Apprendre l'Islam*. L'organisation du forum se stabilise en 2010 et compte aujourd'hui 25 catégories différentes. Chaque jour, ce sont ainsi plusieurs dizaines de sujets de discussions et plusieurs centaines de messages qui sont ouverts et échangés entre les membres du forum.

56. https://web.archive.org/web/20101118010916/http://www.yabiladi.com:80/mag

57. https://web.archive.org/web/20030626013657/http://www.yabiladi.com:80/rencontres/

58. https://web.archive.org/web/20060209041225/http://www.yabiladi.com:80/cafe-debat.php

59. https://web.archive.org/web/20050711003647/http://www.yabiladi.com:80/sommaire-debat.html

60. Suite de messages inféodés à un même sujet de discussion.

La manifestation du 20 février 2011

Le **Printemps arabe** est une suite de protestations et de soulèvements des populations du monde arabe (Maghreb, Proche & Moyen-Orient), débutée en décembre 2010. L'ampleur et la finalité des contestations varie d'un pays à l'autre, mais la contiguïté de leurs déclenchements contribue à véhiculer l'image d'un embrasement spontané de la région, traversée dans son intégralité par une vague révolutionnaire (Table 5.2).

Le printemps arabe débute le 17 décembre 2010, en Tunisie, dans la ville de Sidi Bouzid. Suite à la confiscation répétée de son matériel par l'administration locale, le vendeur ambulant M. Bouazizi s'immole devant le siège du gouvernorat. Il décèdera deux semaines plus tard des suites de ses blessures. La portée de son geste trouve immédiatement un écho parmi la population et devient le déclencheur d'une vague

de protestations inédite dans tout le pays. Le peuple tunisien descend dans la rue, scandant son mécontentement d'une situation économique devenue invivable. Au ras-le-bol généralisé, contre le chômage et l'austérité, s'ajoute la soif de renouvellement démocratique d'une jeunesse tunisienne qui cristallise sa colère contre la personne de Z. Ben Ali, dirigeant en place depuis 1987 (Salmon, 2016). L'usage déterminant d'Internet, des réseaux sociaux et des téléphones portables, comme moyens directs d'organisation collective, permet au mouvement de gagner rapidement en ampleur (Lotan *et al.*, 2011; Khondker, 2011).

La fuite de Ben Ali, un mois plus tard le 14 janvier 2011, rapportée sur le Web et amplifiée par des canaux de diffusion comme Al Jazeera, entraine la contagion des révoltes aux pays voisins marquant, de fait, le premier temps fort du Printemps arabe. Et si la démission de H. Moubarak, le 11 Février suivant, après plusieurs semaines d'occupation populaire de la place Tahrir au Caire, est vécue comme une nouvelle victoire démocratique, les assauts contre-révolutionnaires à venir donneront un coup d'arrêt aux protestations. Courant 2011 et 2012, les soulèvements sont réprimés (Égypte, Yémen) et, pour certains, tournent à la guerre civile (Libye, Syrie).

Pays	Déclenchement des contestations	Finalités
Tunisie	17 décembre 2010	fuite du président et changement de régime
Algérie	28 décembre 2010	promesses de réformes
Yémen	29 décembre 2010	départ du président
Jordanie	14 janvier 2011	démission du gouvernement
Mauritanie	17 janvier 2011	manifestations
Oman	17 janvier 2011	réforme constitutionnelle
Arabie saoudite	21 janvier 2011	réformes sociales
Liban	24 janvier 2011	manifestations
Égypte	25 janvier 2011	départ du président et changement de régime
Syrie	26 janvier 2011	guerre civile
Palestine	28 janvier 2011	manifestations
Maroc	30 janvier 2011	réforme constitutionnelle
Soudan	30 janvier 2011	manifestations
Djibouti	1er février 2011	manifestations
Bahreïn	14 février 2011	manifestations
Irak	10 février 2011	manifestations
Libye	13 février 2011	mort du dirigeant et guerre civile
Somalie	13 février 2011	manifestations
Koweït	18 février 2011	manifestations

TABLE 5.2: Liste des pays touchés par le Printemps arabe (source : `https://fr.wikipedia.org/wiki/Printemps_arabe`)

Tout au long de cette période, le Maroc cultive une forme d'*exception marocaine* vis-à-vis du Printemps arabe. Suite à une première manifestation en demi-teinte (le 30 Janvier 2011), de jeunes militants marocains des droits de l'homme voient, dans les mobilisations tunisiennes et égyptiennes, l'occasion de redonner de l'élan à d'anciennes revendications démocratiques. L'analyse stratégique de ces événements, les amène à déduire que les victoires de leurs voisins sont, d'abord, le fait de la mise en avant d'une jeunesse dépolitisée, évoluant sans leader clairement identifié et agissant sur la base d'actions spontanées et innovantes (sit-in, réseaux sociaux, comités locaux...) (Bennani-Chraïbi et Jeghllaly, 2012). Partant de ce constat, un appel est alors lancé pour un grand rassemblement national le 20 février 2011, baptisé *journée de la dignité*. Dans la foulée, le *mouvement du 20 février* (M20F) voit le jour pour coordonner et incarner les différents aspects de la contestation à venir. Contrairement à ses voisins, la mobilisation marocaine n'est donc pas née d'un embrasement soudain, mais d'une construction méthodique, orchestrée par des acteurs hétérogènes.

En effet, plusieurs associations et organisations rejoignent rapidement le mouvement[61] qui déclenche déjà l'enthousiasme de militants historiques de la cause démocratique, comme le blogueur *larbi.org*[62]. Un préavis unitaire est déposé le 17 février, agrégeant derrière le slogan « *Liberté, Dignité, Équité* » une multitude de revendications nationales ou régionales : pour une justice indépendante, pour la reconnaissance de l'amazigh[63] comme langue officielle et pour la dissolution du gouvernement, en passant par la lutte contre la corruption, contre la mauvaise gestion administrative et communale et contre le chômage (Bennafla et Seniguer, 2011). Bien que l'objectif affiché ne soit pas le renversement de la monarchie, les signataires n'en réclament pas moins de grands changements démocratiques et une modification en profondeur de la constitution marocaine, pour « *un roi qui règne mais ne gouverne pas* »[64].

Relayée localement par les comités du M20F, la mobilisation se prépare aussi, et avant tout, sur la toile : un site Web est mis en ligne et associé à des comptes Twitter et de groupes Facebook[65] qui tiennent les internautes informés grâce aux hashtags `#Feb20` et `#20Fev`. La diaspora s'empare elle aussi du sujet via les blogs et les canaux d'informations du M20F. Des manifestations de soutien sont organisées, notamment à Paris[66] et à Bruxelles.

Mais une campagne de contre-communication est lancée en parallèle par le gouvernement. Les membres du M20F sont ciblés personnellement sur les réseaux, dénigrés et intimidés. Anti- et pro- 20 février s'affrontent par vidéos interposées sur Youtube[67] et Dailymotion. À deux jours de l'échéance, le M20F publie une dernière vidéo, où de jeunes militants appellent, face caméra, à descendre dans la rue, incarnant à visage découvert l'élan d'une jeunesse marocaine qu'ils souhaitent voir

61. https://www.yabiladi.com/articles/details/4596/marche-fevrier-maroc-associations-defense.html

62. https://frama.link/cTMRPS5a

63. Langue berbère

64. https://frama.link/V7R8y1SH

65. https://web.archive.org/web/20110220014401/http://mamfakinch.posterous.com/, https://twitter.com/mamfakinch, https://www.facebook.com/Movement20

66. https://youtu.be/g7muGDOGoto

67. https://www.youtube.com/user/mouvement20fevrier/feed et https://www.youtube.com/user/MoroccanYouth/videos

défiler massivement[68].

Le 20 février 2011, entre 37 000 (selon le ministère marocain de l'intérieur) et 100 000 personnes défilent dans 53 localités différentes (Figure 5.21). Si le succès de la journée est mitigé d'un point de vue quantitatif, le M20F marque symboliquement les esprits et veut pousser plus loin la mobilisation, prévoyant de redescendre dans la rue les 20 de chaque mois à venir. Mais le gouvernement marocain et le roi Mohammed VI prennent, à titre préventif, une série de mesures visant à apaiser les mécontentements (Desrues, 2012). Le 9 Mars, Mohammed VI prononce un discours à la télévision et annonce une modification de la constitution, dont les détails[69] sont donnés le 17 Juin : le premier ministre sera dorénavant désigné par des élections, mais le roi gardera le pouvoir de dissoudre le parlement qui, de son côté, gagne en indépendance. Par ailleurs, la langue berbère devient, comme l'arabe, langue officielle d'état. Ce faisant, le régime marocain arrive à gérer la menace contestataire et étouffe une possible amplification du mouvement en reprenant à son compte l'agenda politique des événements.

La nouvelle constitution est approuvée par référendum le 1er juillet 2011, malgré l'appel au boycott[70] lancé par le M20F qui n'y voit que des changements de façade. Quelques manifestations émaillent le reste de l'année, mais le M20F peine à mobiliser et à retrouver sa légitimité. Lors des élections législatives de novembre 2011, les conservateurs du Parti de la justice et du dévéloppement (PJD) arrivent en tête des suffrages, mettant ainsi un terme au volet marocain du Printemps arabe.

68. https://youtu.be/6Y_J-2S_1m8

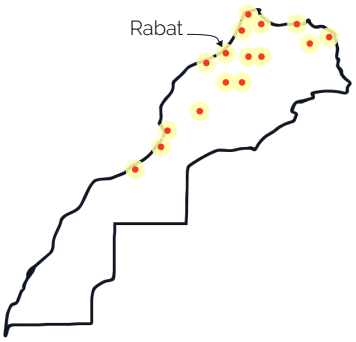


FIGURE 5.21: Principaux foyers de contestation, lors des manifestations du 20 février 2011 au Maroc (source : https://globalvoices.org/2011/02/20/morocco-across-the-nation-demonstration/)

69. https://frama.link/oB6K7-hJ

70. https://youtu.be/irbHL8Io--Q

Topologie d'un site Web

Comment évolue un site ? Comment s'étend-il sur le Web ? Comment naissent et disparaissent certaines de ses sections ? Peut-on voir un site croitre, à mesure que de nouvelles pages s'ajoutent à son arborescence ?

Ces questions sortent du cadre strict de notre étude sur le forum de *yabildai.com*, mais nous souhaitons tout de même nous y attarder ici, afin d'ajouter une nouvelle brique à nos outils d'exploration. En effet, le Web est tout autant un flux continu d'informations qu'un espace en pleine expansion, dont nous voulons maintenant retracer la genèse (à l'échelle d'un site tout du moins). Aussi, plusieurs pistes de réflexion s'offrent à nous : Qu'est ce qu'un site Web, d'un point de vue topographique ? Comment le représenter ? Comment traduire ses dynamiques internes ? Comment rendre compte de l'évolution de ses diverses ramifications depuis les archives Web ? Notre corpus suffit il, à lui seul, à couvrir toute l'étendue d'un site ?

Sur ce point, M. Toyoda et M. Kitsuregawa proposent de se baser sur les URL d'un site pour en reconstituer le graphe (Toyoda et Kitsuregawa, 2003, 2005). Chaque élément d'une URL (compris entre deux /) forme

un nœud particulier du graphe, dont l'évolution est retranscrite par une série de captures successives. On appelle cette méthode une **série temporelle de graphes**. C'est l'approche que nous avions utilisé pour notre prototype *en oursin* (Figure 3.13), présenté en Section 2.2. Mais, avec un peu de recul, il nous semble difficile de comparer visuellement de telles séries, surtout lorsque les sites (comme *yabiladi.com*) se révèlent être très verbeux en URL à afficher. De plus, la période de capture et de représentation reste sujette à discussion et peut varier sensiblement d'un site à l'autre.

À l'opposé, la méthode employée par B. Fry, dans son application Anemone[71] (Fry, 2000, p.76-82), permet de saisir de manière dynamique l'évolution d'un site, vu ici comme un être vivant en pleine croissance. Le système, conçu par Fry, part de la première page archivée d'un site Web donné, à laquelle il ajoute toutes les n secondes une nouvelle ramification, collant ainsi à l'arborescence du site telle qu'elle a été collectée (Figure 5.22) :

71. `http://benfry.com/anemone/`

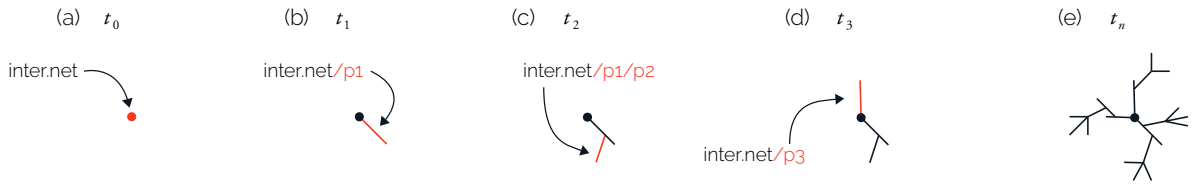


Figure 5.22: Fonctionnement du système de visualisation Anemone, pour un site fictif *inter.net*

Mais si l'idée de croissance est bien restituée par Anemone, la spatialisation (en haut, en bas, à gauche...), de chaque branche est, quant à elle, codée pour suivre un comportement aléatoire : la structure du site reste la même, mais l'agencement local des pages varie à chaque relance de l'application. Ce procédé écarte malheureusement toute possibilité d'étude répétée et systématique. De plus, le temps est ici pensé comme une mécanique de visualisation (le temps sert à mettre à jour l'Anemone) et non comme une véritable dimension d'analyse.

Sur ce point, en matière de visualisation des dimensions spatio-temporelles d'un processus historique, la célèbre *Carte figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813* de C. J. Minard reste la meilleure source d'inspiration. Elle permet de suivre le trajet (tracé beige) de l'armée napoléonienne et ses pertes humaines (épaisseur du tracé), à travers la Russie (indications géographique), au cours de l'hiver 1812-1813 (temps en abscisse). Deux dimensions (temps et espace) servent ici de cadre à l'étude de l'évolution d'une troisième (le nombre d'hommes).

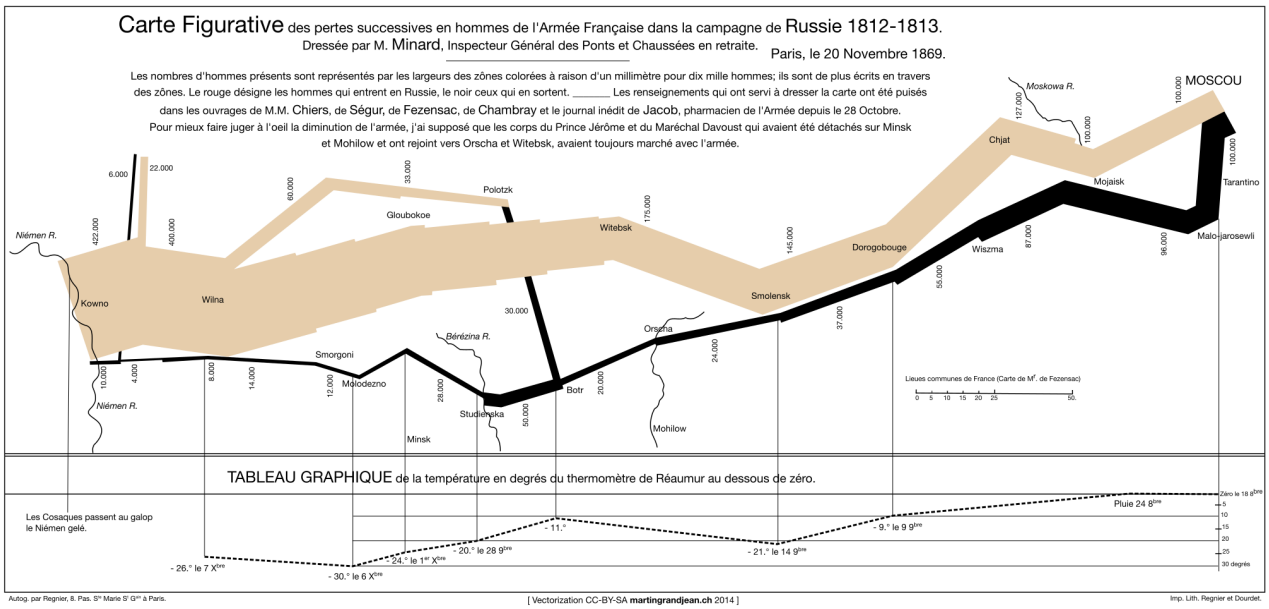


FIGURE 5.23: C.J. Minard, (1869), Carte figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813

Au regard de ces exemples, il nous semble judicieux de choisir de représenter un site comme un arbre d'URL, se déployant depuis sa racine (la première page mise en ligne) et grandissant à mesure que de nouvelles pages y sont publiées. En effet, un arbre croit suivant une direction donnée qui peut, dans notre cas, être associée à un axe temporel, placé en abscisse à la manière d'une timeline (Figure 5.24) :

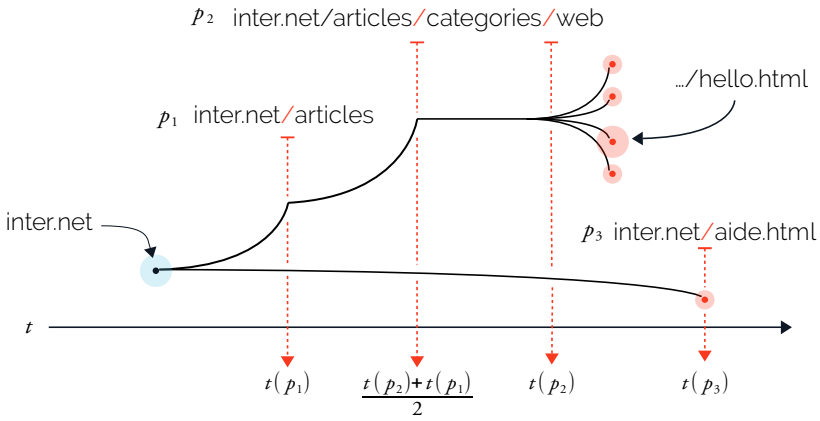


FIGURE 5.24: Proposition de visualisation de l'évolution de la structure d'un site fictif *inter.net* dans le temps

Chaque URL du site représente ici une branche de l'arbre et les pages Web font office de feuilles (Figure 5.24, points rouges). L'abscisse d'une URL est déterminée suivant la date de création (dans le meilleur des cas) ou de téléchargement (dans le pire des cas) de la page qu'elle adresse. Pour positionner les embranchements mécaniques[72] de notre arbre,

72. les portions d'URL entre deux / ne correspondant à aucune page réelle.

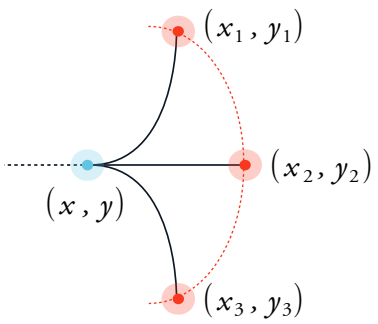


Figure 5.25: Positionnement de 3 pages (rouge) autour d'un embranchement parent (bleu)

nous leur attribuons arbitrairement une date médiane (Figure 5.24, voir p_2). L'ordonnée d'une URL est, quant à elle, définie de manière arbitraire (l'axe vertical n'a donc pas de signification particulière). Les URL sont placées, du haut vers le bas, de la plus ancienne à la plus récente, suivant une demi-ellipse centrée sur l'embranchement parent (Figure 5.25). L'épaisseur d'une branche de l'arbre d'URL est déterminée par son nombre d'enfants. La taille des marqueurs représentant les pages peut être ramenée à une mesure donnée, comme par exemple, un nombre de messages postés.

Ceci étant défini, nous développons alors un système[73], permettant de récupérer, depuis notre moteur d'exploration et depuis le corpus d'Internet Archive, l'ensemble des URL archivées d'un site donné. En effet, afin de garantir une plus grande couverture de l'étendue d'un site nous choisissons de coupler plusieurs sources de données. Nous implémentons cette méthode et construisons une interface de visualisation Web. L'utilisateur peut y zoomer à volonté et sélectionner, s'il le souhaite, le détail d'une branche ou d'une feuille (Figure 5.26).

73. Open-source et téléchargeable ici `https://github.com/lobbeque/archive-viz/tree/master/stayingAlive`

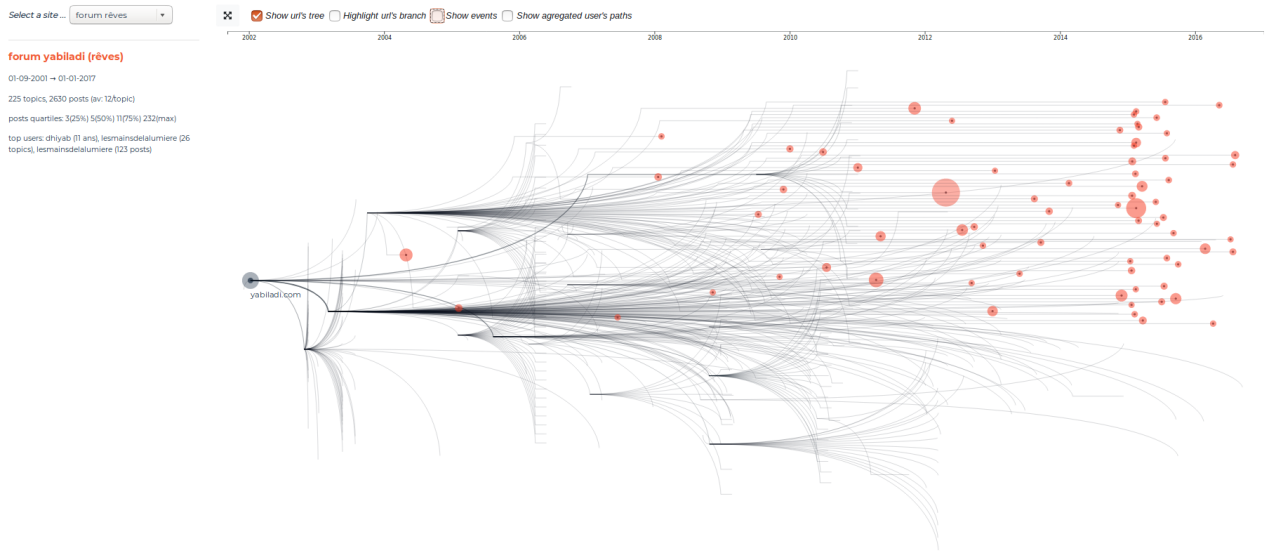


Figure 5.26: Capture d'écran de l'évolution topologique de la section *Interprétation des rêves, roqya, djinn* du forum de *yabiladi.com*

Revenir au 20 février

La mobilisation de la jeunesse marocaine sur les réseaux sociaux a servi d'amorce aux manifestations du 20 février 2011. Le temps d'une journée, les plateformes Twitter et Facebook ont prolongé les rues de Casablanca, de Rabat et de Marrakech via les hashtags `#Feb20` et

`#20Fev` (Section 5.2). Porté par une communication inédite, l'élan de ce mouvement a atteint la communauté des marocains de l'étranger, qui s'est organisée pour apporter son soutien (Desrues, 2012). Parallèlement à ces événements, le portail en ligne *yabiladi.com* s'est imposé comme un point de rassemblement et de discussion privilégié pour la diaspora marocaine sur le Web.

Partant de là, notre exploration se tourne maintenant vers la masse des archives de *yabiladi.com*. Dans quelle mesure la mobilisation du 20 février 2011 a-t-elle pu impacter le quotidien de ce site et de ses utilisateurs ? Y a-t-il eu une forme d'organisation collective, en amont de la manifestation ? Une communauté, même éphémère, de soutiens ou de détracteurs a-t-elle pu y voir le jour ?

Aussi, commençons par définir notre espace d'exploration. Nous partons d'un site Web unique, *yabiladi.com*, dont nous possédons des collectages hebdomadaires, allant de mars 2010 à septembre 2014. Les corpus d'Internet Archive peuvent, ici, servir de données d'appoint ou complémentaires. *Yabiladi.com* est un site vaste, possédant plusieurs sections, dont un forum divisé lui même en une dizaine ou une vingtaine de catégories (Figure 5.20). Chaque catégorie englobe plusieurs threads de messages, hébergés sur des pages Web dédiées. Si, d'un point de vue temporel, notre cible est connue, il s'agit de l'avant et de l'après 20 février 2011, le nombre de messages ou de threads s'y référant reste pour le moment inconnu. L'idée est donc : **1)** d'identifier les messages parlant de la mobilisation, **2)** de remonter aux utilisateurs qui les ont publiés, **3)** de voir si ces membre forment une communauté.

Tout d'abord, nous construisons un arbre d'URL qui servira de support à nos visualisations. Nous partons des 2 683 928 pages archivées du corpus e-Diasporas et des 887 981 pages collectées par Internet Archive, pour proposer une représentation, la plus complète possible, de la structure du site et de son évolution. Les données venant de nos archives sont associées, si possible, à une date de création ; les pages fournies par Internet Archive sont, elles, datées par rapport à leur collecte. En cas de doublon, nous privilégions la datation la plus précise (Table 4.1).

Les archives DAFF de *yabiladi.com* sont ensuite envoyées dans notre moteur d'exploration qui les fragmente et les indexe. Là, un filtre est appliqué, pour ne conserver que les pages appartenant à la seule section forum du site : celles contenant la mention `/forum/` dans leurs URL. Passée cette étape, nous retenons un groupe de 109 534 pages Web, associées à une date de création unique et segmentées en 422 906 fragments Web.

Mais il ne nous semble pas nécessaire d'analyser toutes les catégories du forum. Les sections *Général* et *Actualités* sont plus susceptibles de renfermer des threads et des messages associés au 20 février. Or comme

nous savons que la section *Actualités* a subi de nombreux changements au cours du temps (Figure 5.20)[74], nous choisissons de les agréger en une seule et même section renommée *Actualités*.

74. *Actualités*, *Actualités du Maroc*, *Actualités Marocaines*, *Actualités internationales* et *Actualité du Maroc et du Monde*

Par ailleurs, en se basant sur la structure des URL du forum, il est possible d'en filtrer les pages par catégorie. En effet, depuis la création du forum en 2002, les URL suivent toutes le même schéma :

```
.../forum/thread_title-category_id-thread_id.html
```

Par exemple, la `category_id` de la section *Général* est 1, information qui peut facilement être extraite. Nous appliquons ainsi ce filtre et obtenons en sortie de notre moteur les résultats de la Table 5.3.

Table 5.3: Statistique des archives des sections *Général* et *Actualités* de *yabiladi.com*

	Général	Actualités
Nombre de threads	17 025	20 553
Nombre de messages	352 231	328 965
Nombre de membres	19 745	10 819
Premier message	janv. 2002	janv. 2002
Max. messages par membre	5019	6041
Max. threads par membre	1316	2057
Messages moyens par thread	21	16

Pour construire cette table, nous demandons à notre moteur de nous retourner les fragments Web ayant la forme d'un message générique, publié sur le forum, soit : un nom d'utilisateur, une date d'édition et un ensemble d'éléments textuels postés au fil d'un thread. Dans la suite de cette section, et par abus de langage, fragments Web et messages publiés seront confondus et utilisés de manière équivalente. Aussi, pour identifier les messages[75] se référant aux événements du 20 février 2011, nous utilisons la recherche plein texte de notre moteur et partons de l'ensemble de mots clés suivants :

75. Le forum est écrit en majorité en français

```
février, 20 février, manifestation, manifestation février,
M20F, #Feb20, #20Fev
```

Nous agrégeons, ensuite, par thread les fragments retournés. Après une validation manuelle, nous n'en conservons que 12, selon qu'ils traitent directement du 20 février ou de ses conséquences[76]. Nous présentons, ci-dessous, la liste des différents titres (ou sujets) de ces 12 threads :

76. Réception par la diaspora, suites à donner au mouvement, discréditation d'un des organisateurs...

```
Qui assistera le 20 février au Maroc....???
Marocains résident à l'étranger que pensez-vous de ce soulèvement...
2 millions de manifestants a la place jame3 lfna
Photos Des Manifestations
```

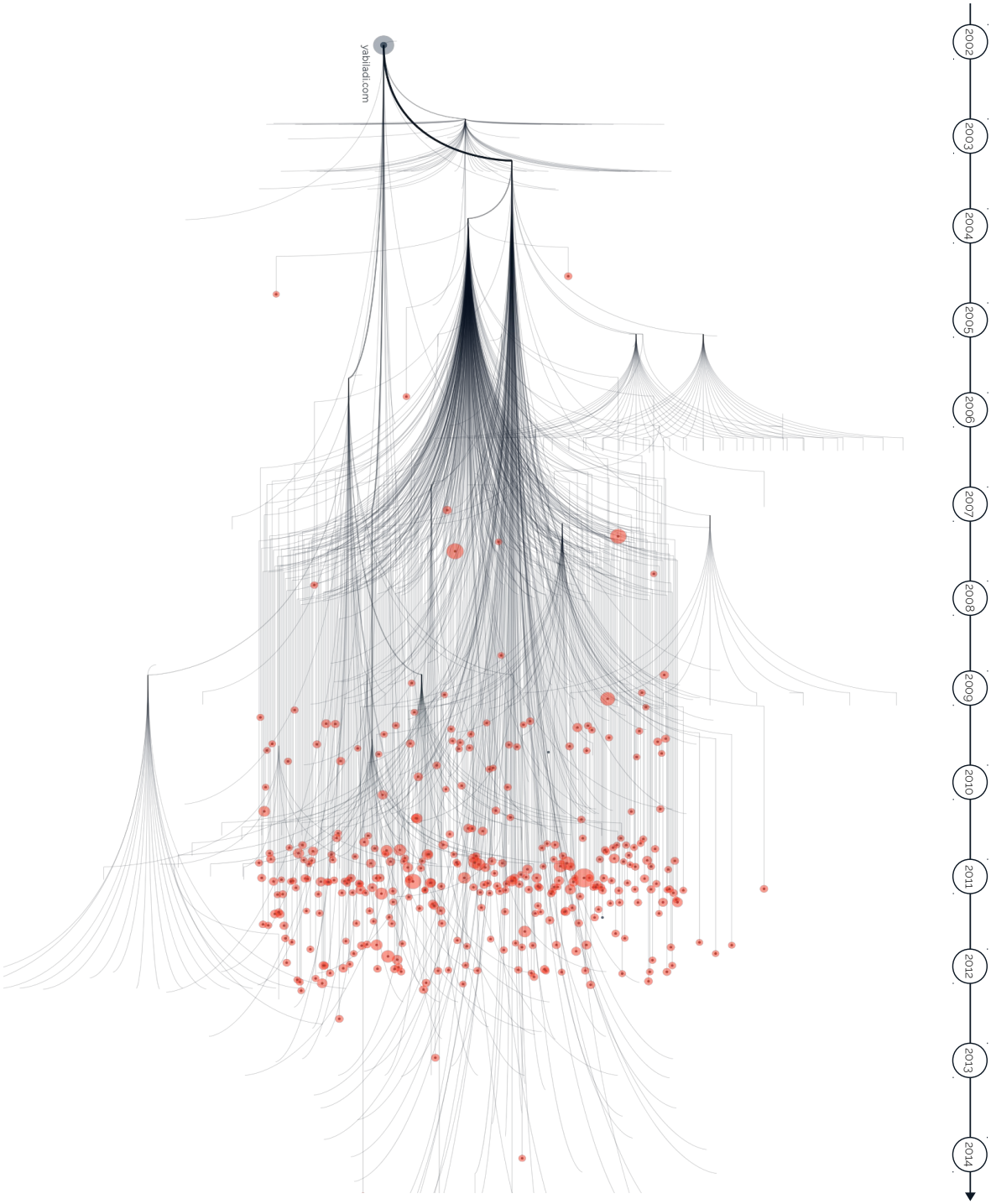


FIGURE 5.27: Évolution de la structure de *yabiladi.com* et ensemble de threads V_1

```
Anecdotes du 20 Février
Arrestations, violences, et un mort suite au 20 février...
Demain Manifestation à 15 h, PLACE DE LA REPUBLIQUE PARIS
Communiqué du comité marocain de suivi du 20 février à Paris
Le mouvement du 20 février et le 20 mars
Le "20 Fevrier" est il democratique
L'un des initiateurs du 20 fév est un arnaqueur
Mouvement du 20 février France
```

Nous appelons V_0 ce groupe initial de 12 threads, consistant en un ensemble de 196 messages, écrits par un panel de 94 auteurs uniques nommé E_0. L'idée est alors d'élargir ce premier groupe aux messages écrits en amont et en aval de la manifestation. Pour ce faire, nous utilisons E_0 comme nouveau point de départ et filtre de sélection. Aussi, nous demandons à notre moteur de nous retourner tous les threads du forum dans lesquels au moins deux auteurs de E_0 ont publié un message. En effet, c'est un collectif d'utilisateurs que nous cherchons ici à étudier. Il nous faut donc découvrir s'ils se connaissaient avant la manifestation et s'ils ont continué à collaborer par la suite, notamment en alimentant les même threads à plusieurs mains. Au final, nous obtenons un nouveau groupe de 343 threads appelé V_1. La Figure 5.27 donne à voir la répartition de ces threads (points rouges) dans le temps, parmi les pages archivées du forum de *yabiladi.com*.

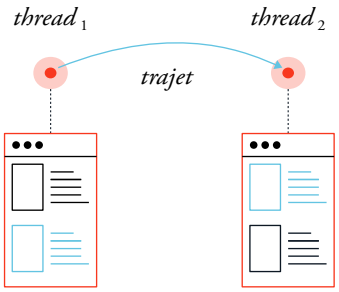


Figure 5.28: Trajet d'un utilisateur ayant posté des messages (bleu) sur deux threads (rouge)

Des trajets de contributeurs

Il faut ajouter à la Figure 5.27 une nouvelle information : la trace du trajet d'un utilisateur, naviguant d'un thread à l'autre du forum. En effet, en utilisant le fragment Web (Section 3.1) comme unité principale d'exploration, nous souhaitons réintégrer l'humain au cœur des archives Web, revenir aux gestes de celles et ceux qui alimentent quotidiennement les sites collectés. Aussi, en nous servant des noms d'auteurs associés aux messages que nous manipulons comme moyen d'identification, nous pouvons afficher de manière dynamique le trajet d'un membre donné (Figure 5.28). La Figure 5.30 présente ainsi l'un de ces trajets individuels, agrégé ici au niveau des threads.

Mais bien plus qu'un individu isolé, c'est une communauté que nous cherchons d'abord à caractériser. Or, le fragment Web nous permet justement de conjuguer le devenir historique de multiples lignes processuelles (Section 3.1), afin de raisonner autour de moments singuliers, à l'image du 20 février 2011. Nous définissons ainsi l'ensemble E_1 comme l'agrégation des trajets empruntés par le groupe d'utilisateurs E_0 et le graphe $G = (V_1, E_1)$ comme un réseau de threads liés entre eux par les trajets de contributeurs communs (Figure 5.29).

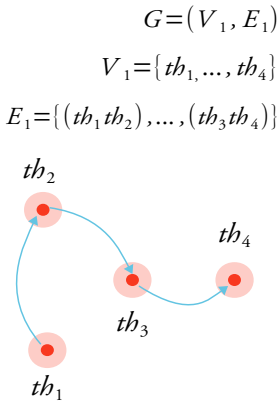


Figure 5.29: Le graphe $G = (V_1, E_1)$ pour V_1 contenant quatre threads de messages

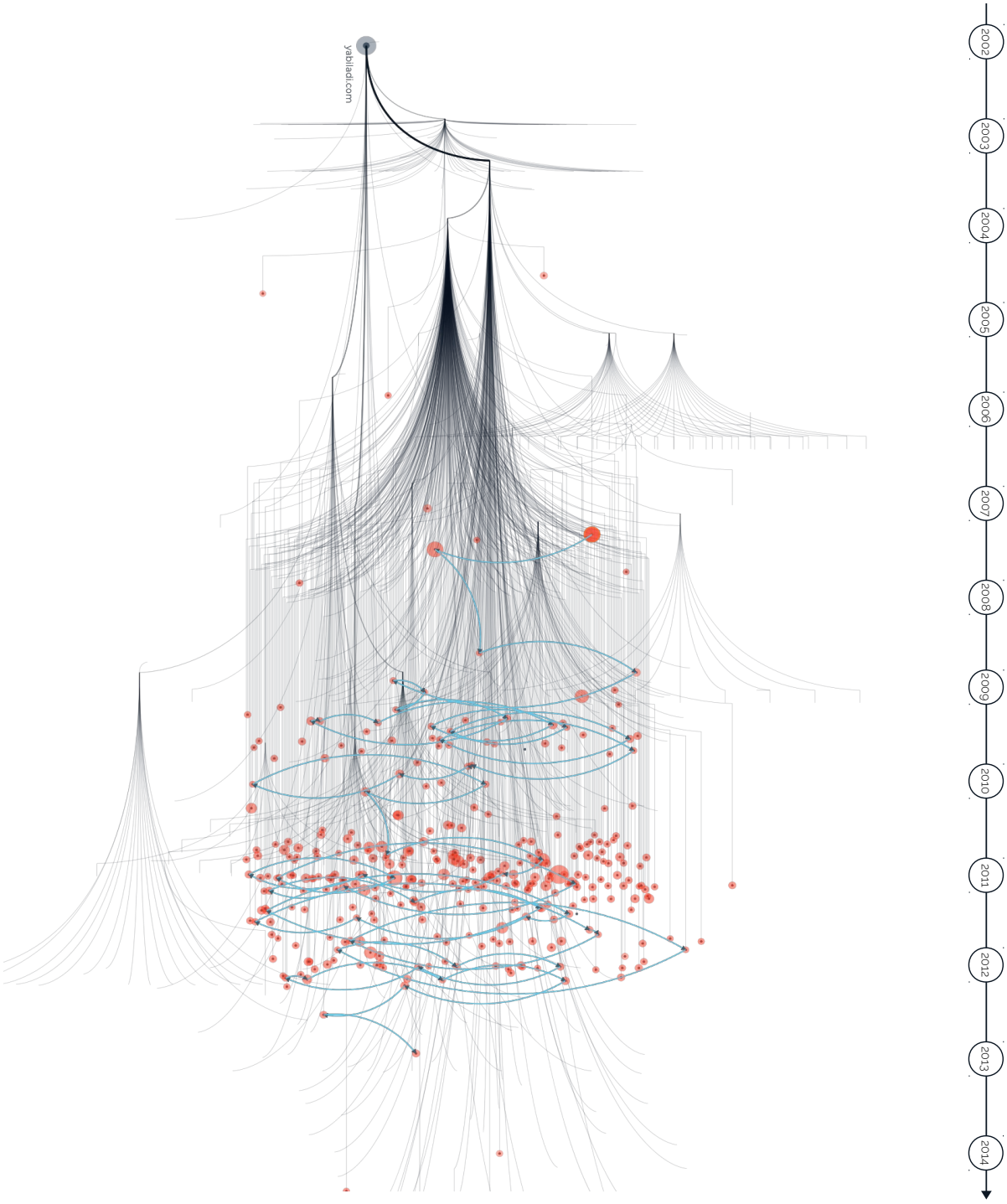


FIGURE 5.30: Exemple de trajet individuel (tracé bleu clair) parmi 66 threads du forum *yabiladi.com*

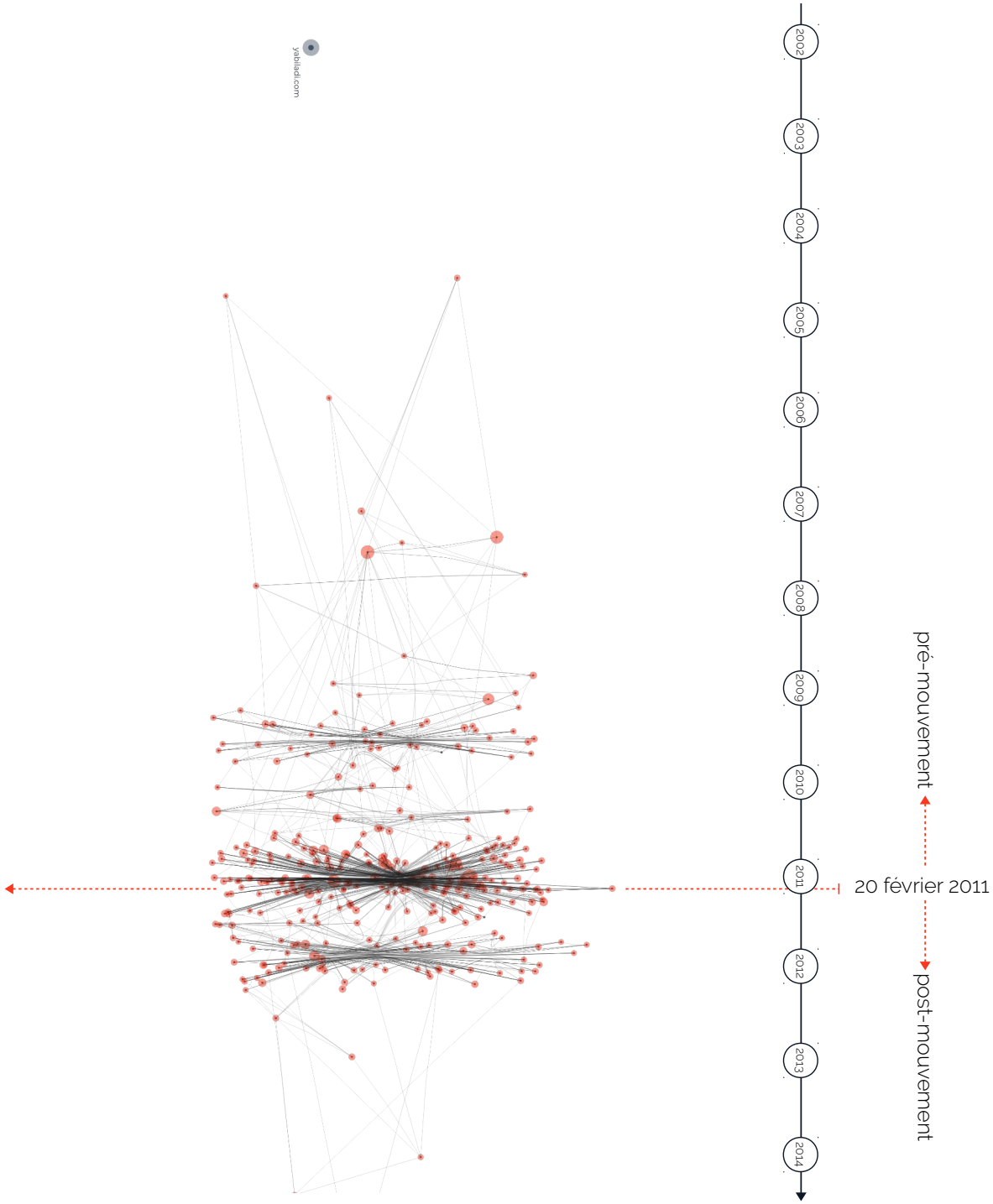


FIGURE 5.31: Trajets agrégés E_1 de l'ensemble des contributeurs des threads V_1 du forum *yabiladi.com*

Nous affichons donc l'ensemble des trajets d'utilisateurs, extraits de nos archives et associés au 20 février. Mais afin de faciliter l'analyse de ces chemins, nous les représentons en utilisant la méthode *edge bundling* (Holten, 2006), qui permet de modifier localement la courbure d'un lien dans un graphe, pour le rapprocher visuellement de ses voisins.

La Figure 5.31 donne à voir ce graphe G et la manière dont il se construit au cours du temps. Pour clarifier, une dernière fois, la lecture de cette figure, nous dissimulons l'ossature de l'arbre d'URL de *yabiladi.com* et n'affichons que les trajets E_1 (tracés noirs) parmi les threads V_1 (points rouges).

Une première analyse de la Figure 5.31, révèle l'apparition de plusieurs **points de fixation**. Ce sont des points précis du graphe G où arcs et nœuds semblent converger tous au même endroit et même moment. Le plus visible de ces points, là où le nombre de threads est le plus important et où les trajets sont les plus densément concentrés, correspond aux voisinages directs du 20 février 2011 (Figure 5.31, flèche rouge). Plus précisément, ce sont 25% des 343 threads de V_1 qui sont crées entre janvier et février 2011.

Cette première observation fait sens au regard notamment de la construction de notre groupe de départ V_0, centré sur la mobilisation marocaine. Mais, nous voyons également que les manifestations du 20 février ont agrégé, autour d'elles, une communauté de membres du forum qui communiquaient déjà en entre eux, en amont des événements (un autre point de fixation est visible en 2009 par exemple, Figure 5.31).

Aussi, si nous divisons maintenant la Figure 5.31 en deux parties, séparées par le 20 février et appelées respectivement parties **pré-mouvement** et **post-mouvement**, nous voyons que la partie pré-mouvement prend la forme d'un sous-graphe de G, clairsemé et étalé sur une période relativement importante, allant de début 2004 à février 2011.

Sur l'ensemble du groupe initial d'utilisateur E_0, 62% écrivent leur premier message sur le forum durant la période pré-mouvement (Figure 5.32). En particulier, 20% d'entre eux sont apparus sur *yabiladi.com* en 2007-2008, suivant une vague de nouveaux arrivants[77]. Ces premiers membres se sont alors réunis et retrouvés pour débattre ensemble dans des threads dédiés au 20 février et ont, semble-il, continué à communiquer entre eux après ces événements. En effet, les 38% restants, du groupe d'utilisateurs E_0 ont contribué d'abord et en premier aux threads directement liés au 20 février. Avant cette date, ils n'apparaissent pas dans les archives de *yabiladi.com*. Nous les retrouvons, néanmoins, dans la partie post-mouvement de notre graphe.

77. Ces auteurs ont notamment contribué au thread « *Accueil des nouveaux yabis ! Marhaba !* ».

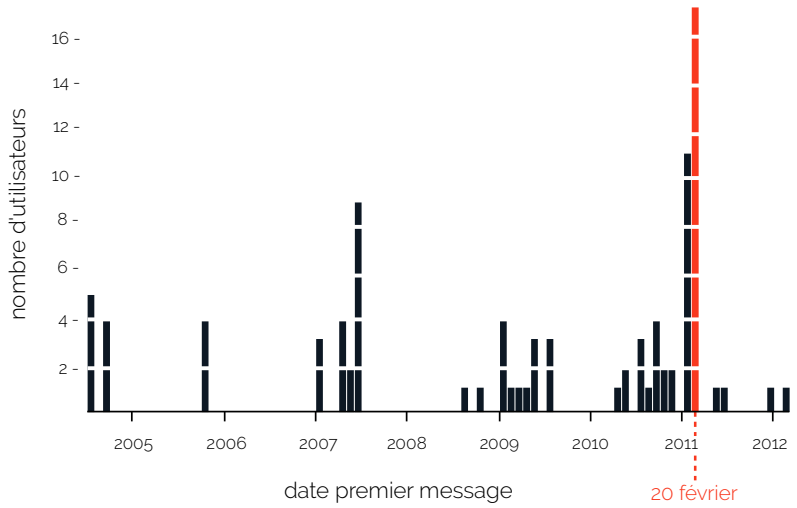


Figure 5.32: Répartition des utilisateurs du groupe E_0 par date de premier message

Ces observations sont ici purement topographiques et basées sur des corrélations temporelles, nous ne rentrons pas encore dans le détail des conversations. Mais, à ce niveau, nous pouvons déjà affirmer être face à un double comportement : **1)** d'anciens utilisateurs du forum convergent collectivement vers les threads mentionnant les manifestations **2)** de nouveaux membres débarquent directement sur *yabiladi.com* pour contribuer aux débats post-mouvement. En revanche, tous disparaissent et s'évanouissent soudainement de la surface du site au début de l'année 2012.

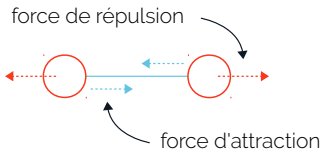


Figure 5.33: Principe de base de Force Atlas, les noeuds d'un graphe sont chargés d'une force de répulsion et les arcs d'une force d'attraction

Vers un embrasement

Afin de mieux saisir la dynamique de convergence des anciens et des nouveaux utilisateurs du forum autour des événements du 20 février 2011, nous souhaitons maintenant affiner notre analyse du graphe G. Avec la Figure 5.31, la spatialisation du graphe, c'est-à-dire la manière dont nous le représentons, est uniquement fonction de la date à la laquelle chaque thread de V_1 a été crée. Mais d'autres types de spatialisation existent, d'autres manières de reprocher ou d'éloigner visuellement les nœuds d'un graphe peuvent être employées.

Nous reprenons donc notre graphe et l'envoyons dans le logiciel de visualisation Gephi [78] où nous choisissons de le spatialiser en utilisant, cette fois ci, la méthode *force atlas* (Jacomy *et al.*, 2014). Appartenant à la famille des algorithmes dits **force-directed** [79], force atlas suit une approche purement topographique, où les nœuds d'un graphe sont supposés être chargés d'une force de répulsion et où les arrêtes sont elles chargées de les retenir (Figure 5.33). L'algorithme itère sur l'en-

78. `https://gephi.org/`

79. `https://en.wikipedia.org/wiki/Force-directed_graph_drawing`

semble du graphe, déplaçant les nœuds et les rapports de force jusqu'à trouver un point d'équilibre global.

La Figure 5.34 présente le graphe G nouvellement spatialisé. De la même manière qu'avec la Figure 5.31, nous voyons que G suit à nouveau une distribution longitudinale, la graphe est étiré confirmant, de fait, son aspect temporel.

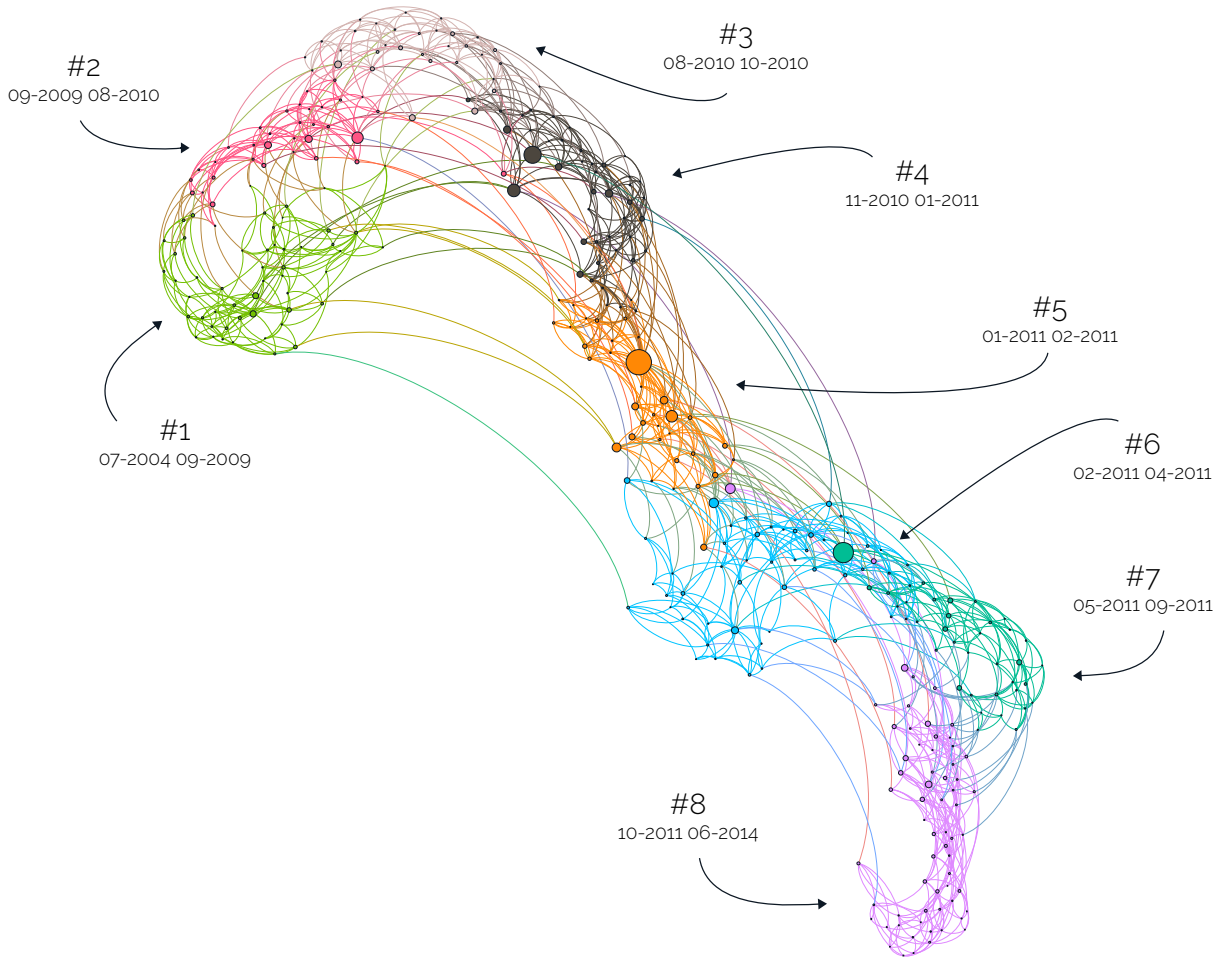


FIGURE 5.34: Spatialisation du graphe G par *force atlas* et clustering des threads par classes de modularité

Les membres du forum contribuent d'abord et avant tout aux threads nouvellement crées. Ils ne reviennent que dans de très infimes cas participer à des discussions passées, construisant leur trajet d'auteur d'un instant t_i vers un instant t_{i+1}, sans retour arrière. Le forum est ainsi vécu comme un flux continu de nouveaux threads, demandant des utilisateurs qu'ils entretiennent une présence régulière sur le site. Le temps moyen espaçant deux threads auxquels participe un auteur est ainsi de 53 jours, la valeur médiane étant de 8 jours. Ils sont, en cela, peu nombreux à marquer de longues pauses entre deux interventions,

seuls 20% des trajets durent plus d'un mois, l'intervalle maximum étant de 6 ans et 2 mois.

Dans un second temps, nous appliquons à notre graphe un algorithme de classification par modularité [80] (Blondel *et al.*, 2008). Nous souhaitons grouper entre eux les threads du forum et ainsi analyser, sous une nouvelle perspective, les points de fixation précédemment révélés (Figure 5.31).

80. `https://fr.wikipedia.org/wiki/Modularité_(réseaux)`

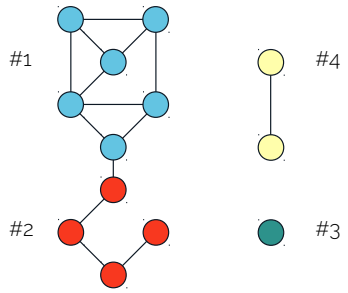


Figure 5.35: Exemple de clustering d'un graphe en 4 classes de modularité (couleurs)

La **modularité** d'un graphe est un score $\in [-1, 1]$ qui cherche à traduire le rapport entre le nombre d'arêtes réel et le nombre d'arêtes espéré (placées de manière aléatoire) d'un groupe de nœuds. Plus ce score est élevé, plus le groupe tend à se rapprocher topographiquement de la structure d'une communauté (Figure 5.35).

Ceci étant fait, le graphe G se voit divisé en 8 clusters distincts (Figure 5.34, $\{\#1, ..., \#8\}$), témoignant chacun d'un moment singulier de l'histoire de cette communauté. À nouveau, rappelons que cette segmentation ne prend pas en compte le contenu des threads, mais uniquement la structure du réseau. Le cluster #*1* couvre ainsi les premières années de G (de juillet 2004 à septembre 2009) et le cluster #*5*, quant à lui, s'étend sur la période la plus courte, de janvier à février 2011.

Afin de donner du sens à la lecture de ces clusters, nous procédons à une analyse manuelle du contenu de chacun des threads les composant. Avertissons simplement qu'il n'est pas aisé de résumer la diversité de ces conversations en une ou deux thématiques, cela n'a peut être même pas beaucoup de sens, aussi ce qui suit doit être pris avec précaution, de futures analyses viendront compléter cette première observation. Ce faisant, la Table 5.4 présente pour chaque cluster, les sujets des 5 threads ayant agrégé le plus d'auteurs (en se basant sur le degré).

Le cluster #*1*, le plus long, est constitué de conversations sur les règles de fonctionnement internes au forum. Les groupes #*2* et #*3* agrègent des propos que l'on peut, faute de mieux, qualifier de quotidien pour ce forum. Le cluster #*4* marque le tournant des années 2010 à 2011, on note néanmoins une focalisation sur l'identité marocaine et des comparaisons entre le Maroc et ses voisins maghrébins. Mais soudainement, le cluster #*5* fait émerger une majorité de threads (19 sur 35) liés au Printemps Arabe et à la manifestation du 20 février. Puis, le cluster #*6* regroupe certains débats post-mouvement et diverses prises de positions politiques. Le cluster #*7* traite de l'héritage politique du 20 février et du changement de constitution à venir. Enfin, avec le cluster #*8*, le forum retombe dans ses conversations quotidiennes.

Cluster	Thread	Degré	Date de création
1	des espions a yabiladi ?	8	25-05-2009
1	Un nouveau forum ? Suggestions ?	9	04-05-2009
1	adopter un chat	10	14-07-2009
1	Assumez-vous vos écrits depuis votre inscription ?	10	26-07-2009
1	Accueil des nouveaux yabis ! Marhaba !	16	25-06-2007
2	Les arabes et les apparences....	9	13-07-2010
2	Franciser son prénom	10	20-06-2010
2	correction orthographique...	12	24-08-2010
2	Double triple pseudo etc...	14	08-02-2010
2	Quels sont les 3 mots que vous associez à la France ?	15	13-08-2010
3	les homo sont musulmans comme les autres	10	15-10-2010
3	Mais qu'est ce qui retiens certains en Occident ? ? ? ?	10	26-09-2010
3	Fautes d'orthographe : Catastrophique	10	19-10-2010
3	porter une barbe au maroc...terrorisme ?	12	15-09-2010
3	Plus beau pays du monde ? ?	12	03-09-2010
4	Bonne Année 2011 Meilleurs Voeux	12	28-11-2010
4	Elections miss and mister yabi 2011.	14	04-11-2010
4	Oh mon Dieu ! Qu'il est beau mon pays	15	16-11-2010
4	L'accueil des MRE au Maroc	15	21-11-2010
4	marocains et algériens	15	29-10-2010
5	Arabe et Tamazight langues officielles !	12	28-01-2011
5	Photos Des Manifestations	13	21-01-2011
5	2 millions de manifestants a la place jame3 lfna	14	20-01-2011
5	Le Couscous est il un plat raffiné ou populaire	14	16-01-2011
5	Qui assistera le 20 février au Maroc.... ? ? ?	30	15-01-2011
6	Kadafi est juif	11	28-02-2011
6	Images d'Algérie	12	04-03-2011
6	Manifestation le 20 Mars au Maroc	12	15-02-2011
6	Le "20 Fevrier" est il democratique	14	17-02-2011
6	Si le Front national passe en 2012 ? ? ?	16	08-02-2011
7	Oui ou Non pour la nouvelle constitution	10	20-05-2011
7	J'aime notre Roi	10	29-06-2011
7	Le roi du Maroc dans la boue !	10	24-05-2011
7	Pourquoi un garçon a droit a plus d'héritages qu'une fille	11	27-08-2011
7	Racisme biologique	12	07-04-2011
8	Mariage à durée..déterminée !	10	07-09-2011
8	Brandissons nos couleurs Rif...	10	19-11-2011
8	J'ai peur de la présence musulmane au Québec	10	21-11-2011
8	Bonne année 2961 et meilleurs voeux à tous !	10	12-12-2011
8	La Grande Classe Marocaine....	11	22-09-2011

TABLE 5.4: Liste des 5 premiers threads (par degré) de chacun des 8 clusters du graphe G

Arrivée à cette étape, notre exploration des archives de *yabiladi.com* nous suggère donc que le mouvement du 20 février 2011 n'a pas réellement été préparé de longue date sur le Web (ou tout du moins sur ce forum). Une étincelle a, soudainement et sans crier gare, embrasé une partie très minoritaire de *yabiladi.com* : 94 utilisateurs actifs sur un total de 30 564. Cette vague a agrégé d'anciens membres établis du forum et de nouveaux venus en brisant leurs habitudes et la routine de leurs conversations quotidiennes. Mais, la mobilisation n'a pas duré dans le temps, ni donné lieu à une autre forme de protestation en ligne ou a une autre incarnation. La réforme de la constitution, menée par le roi du Maroc au printemps 2011, est venue clore ce moment révolutionnaire. Cette communauté d'utilisateurs est, en cela, une construction éphémère et aujourd'hui éteinte.

En effet, la plupart de ses membres ont disparu du forum, au début de l'année 2012, quelques mois seulement après la première apparition de Twitter [81] sur la page principale de *yabiladi.com* (Figure 5.19). Mais, sur les 94 auteurs de l'ensemble E_0, nous pouvons avancer qu'au moins 26 d'entre eux ont créé et utilisent aujourd'hui un compte Twitter avec le même avatar [82].

81. `https://web.archive.org/web/20120627021244/www.yabiladi.com`

82. Comptage et validation manuelle réalisés en avril 2018.

5.4 Mener des recherches historiques depuis les archives Web

Au regard des derniers éléments tirés de nos explorations, peut-être est-il maintenant temps d'interroger le bien fondé d'études basées sur les archives Web comme matière historique. Qu'il s'agisse des auteurs de la blogosphère marocaine ou des membres de *yabiladi.com*, tous ont, peu de temps après le début des opérations de collecte de l'e-Diaspora, choisi de quitter ces territoires du Web pour migrer vers les réseaux sociaux. En explorant d'abord et avant tout ces corpus d'archives Web, ne sommes nous pas passés à côté d'un des aspects majeurs de notre problématique ?

Face aux limites de l'archivage du Web

La mise en place de notre moteur d'exploration (Section 3.3) et la définition du fragment Web comme nouvelle unité d'analyse des archives Web (Section 4.2) ont été guidées par l'idée selon laquelle un site Web devait faire l'objet de recherches historiques dédiées (Brügger et Schroeder, 2017).

Ce faisant, nous montrons qu'en étudiant les archives Web sur une échelle de temps large (Section 5.2) ou à travers un vaste espace d'explo-

ration (Section 5.3), il est possible de comprendre et d'analyser historiquement le devenir de collectifs en ligne aujourd'hui disparus. Nous avons trouvé les indices d'une mutation des auteurs et des lecteurs du Web (blogueurs, commentateurs, membres de forum...) depuis les anciens blogs et les portails en ligne vers les nouvelles plateformes de réseaux sociaux (Facebook, Twitter...).

Mais, partant de nos seuls corpus archivés, nous ne pouvons mettre en lumière l'intégralité du processus de mutation. Seules quelques traces éparses de cette transition (Section 5.2), voire l'absence soudaine de ses dernières (Section 5.3), restent aujourd'hui décelables. À ce niveau, force est de constater que nous touchons ici l'une des limites des archives Web. Il nous faut, dès lors, considérer l'idée que nos corpus peuvent être intrinsèquement incomplets et partiaux.

Cette incomplétude va même bien au delà des cécités et autres artéfacts de crawl hérités de la structure même des archives Web (Section 3.4). Créés et pensés principalement au début des années 2000 (Masanès, 2006b), les systèmes de préservation de la toile ont suivi avec succès les multiples évolutions du Web en tant que médium (Cho et Garcia-Molina, 1999; Oita et Senellart, 2010; Pop *et al.*, 2010), mais pourtant, ils échouent encore à traduire le Web comme un écosystème (Brügger, 2009).

Le Web vivant est un flot continu d'informations autant qu'un espace en perpétuelle expansion. C'est un environnement de créations et de mutations où une multitude d'acteurs peuvent interagir organiquement. Les archives Web, elles, sont des collections figées de captures discrètes, stockées les unes à l'écart des autres. L'INA archive des données Twitter, la BNF collecte des blogs et Internet Archive préserve les premiers pas du Web mais sans jamais chercher à constituer des ponts entre chacun de ces corpus. Les archives des réseaux sociaux et les archives des sites Web classiques sont traitées comme deux entités séparées et irréconciliables. Quand bien même elles ont, toutes deux, été arrachées au même écosystème.

Alors que nous cherchions à remonter les traces archivées des possibles conséquences du Printemps arabe, les acteurs du Web (auteurs, lecteurs, visiteurs...) s'étaient, depuis longtemps déjà, détournés des blogs et des forums (Khondker, 2011; Lotan *et al.*, 2011). La problématique des collectifs en ligne éteints est, ainsi, moins une question de disparition, qu'une question de transition. Malheureusement, nos archives Web ne témoignent que du premier saut de cette transition.

Une micro-histoire

Dans son *Manifeste pour une pensée numérique de l'écrit* (2016), F. Bon questionne les mutations récentes de l'écriture sur support numérique, au regard, notamment, des moments charnières qui ont jalonné l'histoire de l'écriture dite classique, celle du livre papier, des rouleaux et des tablettes. Ces « *moments de transition où deux systèmes techniques se superposent* » et où le « *saut mental* » induit est « *mis à nu* »[83] sont des axes de recherche privilégiés pour resituer des textes et des œuvres numériques dans l'histoire même de leur genèse.

83. Manifeste vidéo `https://youtu.be/8kFntt1l09o`, 5mn 30s

Aussi, l'histoire de l'écriture a-elle été ponctuée de moments-clés, de moments de transition où tout l'écosystème de l'écrit (les modalités techniques, les supports techniques, les auteurs, les œuvres, les moyens de diffusion...) pivote sur lui même et s'engage sur une nouvelle voie. L'historien L. Febvre note ainsi (Febvre et Martin, 2013), parmi d'autres moments : le passage de l'oral à l'écrit, l'apparition des méta-données sur les rouleaux de parchemins, l'invention de la recopie mécanique et de l'imprimerie...

Le Web et l'Internet en général possèdent-ils, eux aussi, de tels moments pivots ? Sur ce point, nous pouvons avancer que le Web s'inscrit déjà dans sa propre micro-histoire. À l'heure où ces lignes sont écrites, le Web a moins de 30 années d'existence (Section 1.1) et son enfance tout autant que son adolescence ne nous ont jamais semblé aussi éloignées. La difficulté à étudier aujourd'hui les premières heures du Web n'est plus à démontrer (Gebeil, 2016b) et, malgré les mesures de préservation mises en place, beaucoup de ressources originales (logiciels, matériels, documentations...) sont désormais perdues (Helmond, 2017). Le simple fait de vouloir remettre sur pied un site Web passé, dans les conditions matérielles de sa première mise en ligne, s'apparente de nos jours à un jeu de piste insoluble ou a un exercice d'archéologie expérimentale[84] (Driscoll et Paloque-Berges, 2017).

84. `https://fr.wikipedia.org/wiki/Archéologie_expérimentale`

L'épopée de l'Arpanet (Russell et Schafer, 2014) et des réseaux antérieurs ou parallèles (Schafer, 2015) à Internet font aujourd'hui figure de pré-histoire du Web. Dans leur enquête sur les pères fondateurs d'Internet, K. Hafner et M. Lyon mettent ainsi en lumière le devenir singulier de la liste de discussion *MsgGroup* dans les années 80 :

« *The dialogue itself in the MsgGroup had always been more important than the results. Creating the mechanisms of e-mail mattered, of course, but the MsgGroup also created something else entirely – a community of equal, many of whom had never met each other yet who carried on as if they had know each other all their lives. It was the first place they had found something they'd been looking for since the Arpanet came into existence. The MsgGroup was perhaps the first virtual community.* » — (Hafner et Lyon, 1998, p.218)

En outrepassant l'objet initial de cette liste[85], pour en faire un espace vibrant de discussions et d'échanges, les membres du MsgGroup ont préfiguré les forums, les communautés de blogs et les réseaux sociaux à venir (Stevenson, 2016; Paloque-Bergès, 2018).

L'histoire du Web est ainsi jalonnée de ces moments où tout bascule, de ces moments pivots : comme la mise en place des systèmes xDSL et ADSL[86] qui participent de la démocratisation du Web à la fin des années 90, ou l'émergence des systèmes de partages et d'échanges de *consumer gifts* (Giesler, 2006) qui tracent des ponts entre économie du Web et du réel au début des années 2000, en passant par le développement des smartphones et du Web mobile au tournant des années 2010...

85. Le MsgGroup est une liste de diffusion créée en 1975, à laquelle souscrivaient les spécialistes de l'e-mailing émergeant, dans le but de définir des bonnes pratiques, des normes, des standards...

86. `https://fr.wikipedia.org/wiki/ADSL`

Les moments pivot du Web

Si F. Bon ou J. Baschet utilisent tous les deux la notion de **moment**, plutôt qu'*instant* ou *point*, pour parler de moments de transitions entre systèmes ou de moments de rencontre de lignes processuelles, c'est que le moment est « *une forme diversifiée du temps vécu* ». C'est une manière singulière de s'inscrire dans le temps et caractérisée par une « *densité spécifique* » (Baschet, 2018, p.186).

Chez ces chercheurs, la notion de moment est issue de la conception chinoise du temps. Contrairement à la pensée occidentale, le temps est ici envisagé comme une suite de moments (*shí*) successifs, s'inscrivant dans la durée (*jia*), à la fois distincts et liés entre eux (Chen, 2011) et non comme un continuum abstrait de points élémentaires (Augustin, 421, livre XI, chap. XV, 20). Dans la pensée chinoise, le temps est un agencement de moments et de durées. Le moment s'éprouve ainsi qualitativement, ce n'est pas une valeur abstraite (un instant, une date, une heure...) mais le marqueur d'une temporalité concrète dont il faut faire l'expérience en la vivant (comme une saison ou une époque). L'intensité d'un moment (douce, tendue, brutale, ou diluée) varie tout autant que sa temporalité (brève, durable, étalée ou soudaine), permettant ainsi de différencier deux moments l'un de l'autre.

Mais pour nous, parler de moment s'inscrit d'abord dans la continuité de notre discussion (Section 3.2) sur les diverses temporalités du Web et de ses archives. Nous avancions alors, qu'il pouvait être possible de conjuguer ensemble plusieurs fragments Web tout en suivant leurs lignes processuelles, et ce, afin de saisir les cristallisations du Web autour de moments singuliers.

Ce faisant, une histoire du Web qui s'attarde sur ces moments est, d'abord et avant tout, une histoire qui s'attache à comprendre qualitativement des processus historiques d'étendue et de durée variables. Sur ce point, les Sections 5.2 et 5.3 témoignent, à travers le prisme particu-

lier des archives de l'e-Diaspora marocaine, du moment spécifique où certains acteurs de la toile ont choisi de quitter le territoire des blogs et des forums pour rejoindre le Web des plateformes de réseaux sociaux. Ils cherchaient, par là même, à satisfaire leurs désirs et leurs besoins grandissant d'expression, d'organisation, de connexion et de partage de l'information en ligne. Nos archives Web (par leur contenu et par leur forme) sont ainsi chargées des derniers soubresauts d'un Web mourant (Stevenson, 2018) qui, lors de sa collecte, ne reflétait déjà plus la réalité de son temps (Helmond, 2015).

De fait, nous appelons **moment pivot du Web**, une époque de transition entre deux systèmes, un moment du Web où de nouveaux usages émergent et s'écartent des pratiques alors en vigueur sur la toile. C'est un moment d'instabilité où deux manières de vivre et d'expérimenter le Web se chevauchent. En cela, un moment pivot du Web nait de la rencontre de trois facteurs, soit : **(1)** la rencontre à un moment donné entre **(2)** une accélération technique et **(3)** des utilisateurs capables de s'en saisir et s'en emparer. La convergence de ces trois facteurs fait alors bifurquer le Web dans une direction inédite.

***

Ce chapitre nous aura permis de mettre en pratique et d'expérimenter une forme désagrégée d'exploration des archives Web, se basant, en premier lieu, sur le fragment Web comme unité d'analyse. La souplesse de cette méthode se conjugue assez naturellement avec l'Analyse Exploratoire de Données, courant singulier des statistiques, dont nous avons présenté les principales caractéristiques (Section 5.1).

En partant d'une question simple, sur le devenir d'une blogosphère de sites migrants aujourd'hui éteints, nous avons cheminé à travers les archives Web, suivant une logique inductive, privilégiant l'observation et la recherche de l'étonnement (Section 5.2). De fait, il nous a été possible de suivre et de comprendre la manière avec laquelle ces blogs et les auteurs, qui les administraient, ont petit à petit migré vers d'autres territoires du Web : vers les réseaux sociaux (Twitter, Facebook...). Dans ce nouvel espace d'expression, ces acteurs du Web se sont retrouvés pour reformer, puis renforcer leur communauté.

Suite à la découverte, dans les archives des blogs, de deux hashtags liés à la manifestation marocaine du 20 février 2011, nous nous sommes demandés si le portail en ligne *yabiladi.com* avait pu être impacté par ce mouvement de contestation (Section 5.3). Site central de l'e-Diaspora marocaine, *yabiladi.com* possède une section forum, fréquentée depuis toujours par des internautes marocains de l'étranger, susceptibles d'avoir été emportés par la vague révolutionnaire du Printemps arabe. Dans les faits, une communauté éphémère d'utilisateurs du forum s'est construite en réponse à ces événements, agrégeant d'an-

ciens et de nouveaux membres du portail. De nombreux messages ont été publiés par ce collectif, mais sans pour autant réussir à s'inscrire dans la durée. Ses membres disparaissent rapidement des pages du forum, alors que le mouvement et les manifestations s'essoufflent, étouffés par le pouvoir marocain.

Comme un bilan de ces deux explorations, nous voulons d'abord mettre en avant le rapport vertigineux entre l'infinité de points de départ possibles pour nos analyses et la dimension restreinte voire réduite des entités sur lesquelles nous conduisons, finalement, nos recherches. D'une part, nous partons d'un ensemble de 47 blogs archivés pendant 4 années et terminons par questionner les messages d'adieu et les hashtags laissés par deux blogueurs uniquement. D'autre part, nous nous positionnons face à la collection des pages archivées de *yabiladi.com*, qui se comptent en millions d'enregistrements, pour aboutir, en fin de compte, à l'étude d'un collectif de 94 membres.

Pour nous, ce rapport est le résultat naturel de notre chaine d'analyse socio-technique, alliant approches qualitatives et quantitatives, déjà à l'œuvre lors de la création de l'Atlas e-Diasporas. C'est par l'articulation constante de ces diverses stratégies que nous sommes en mesure d'identifier, parmi nos vastes corpus archivés, des cas singuliers qui n'en demeurent pas moins aux croisement de multiples processus historiques, sociaux et techniques : le Printemps arabe, les formes d'organisation collective en ligne, le passage d'un âge du Web à un autre, etc.

De fait, ces deux explorations découlent d'une suite non linéaire d'opérations de natures (automatiques ou manuelles) et d'échelles (larges ou réduites) variables, comme le résume la Figure 5.36. Aussi, nos corpus d'archives Web n'étant pas, en eux même, exhaustifs, il a été nécessaire de les enrichir de collectages réalisés par Internet Archive et de données extraites du Web vivant. Au cours de ces boucles d'exploration successives, le contenu et la structure des archives nous ont mené à conjuguer ensemble des techniques se référant autant à la science des réseaux, qu'à la recherche d'information, en passant par l'analyse sociologique d'une communauté et la définition d'un contexte historique.

Mais, si l'ensemble s'apparente effectivement à une chaîne de traitements ad-hoc, cela ne nous empêche pas de concevoir chacune de ces briques, avec pour finalité, de les rendre adaptables à une grande variété de cas. Il en va ainsi des fragments Web, qui sont ici utilisés dans deux explorations distinctes, mais également de notre système de visualisation de la topologie d'un site Web (Figure 5.24). Cet outil de visualisation peut, partant d'un corpus d'archives Web, rendre possible l'étude des grands moments de l'histoire d'un site préservé, comme

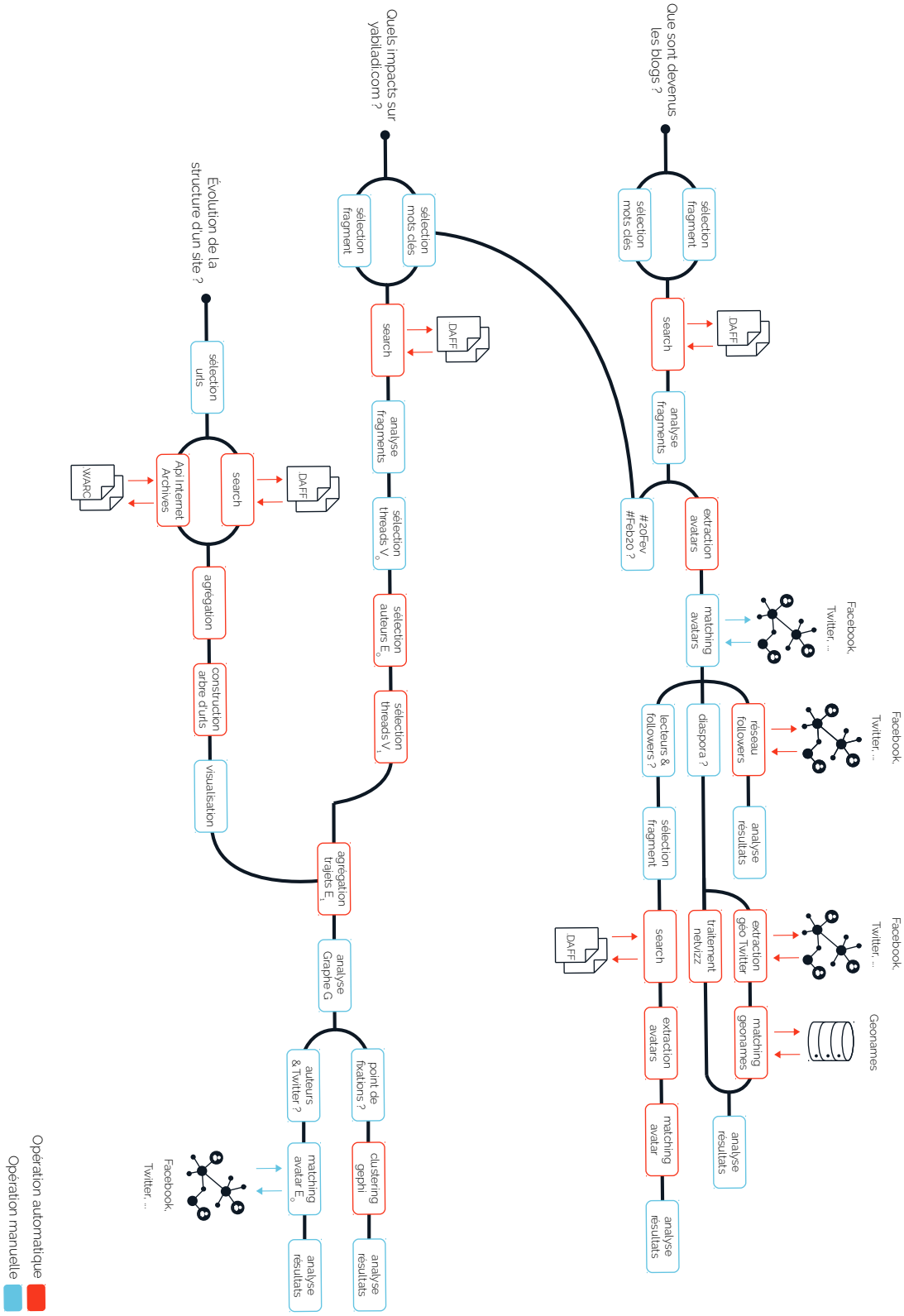


FIGURE 5.36: Résumé des diverses opérations mises en place lors des explorations 5.2 et 5.3, avec distinction entre approches automatiques et manuelles

l'illustre, par exemple, la Figure 5.37 pour les sites *7didane.org* (haut), *atmf.org* (bas, et Figure 5.38).

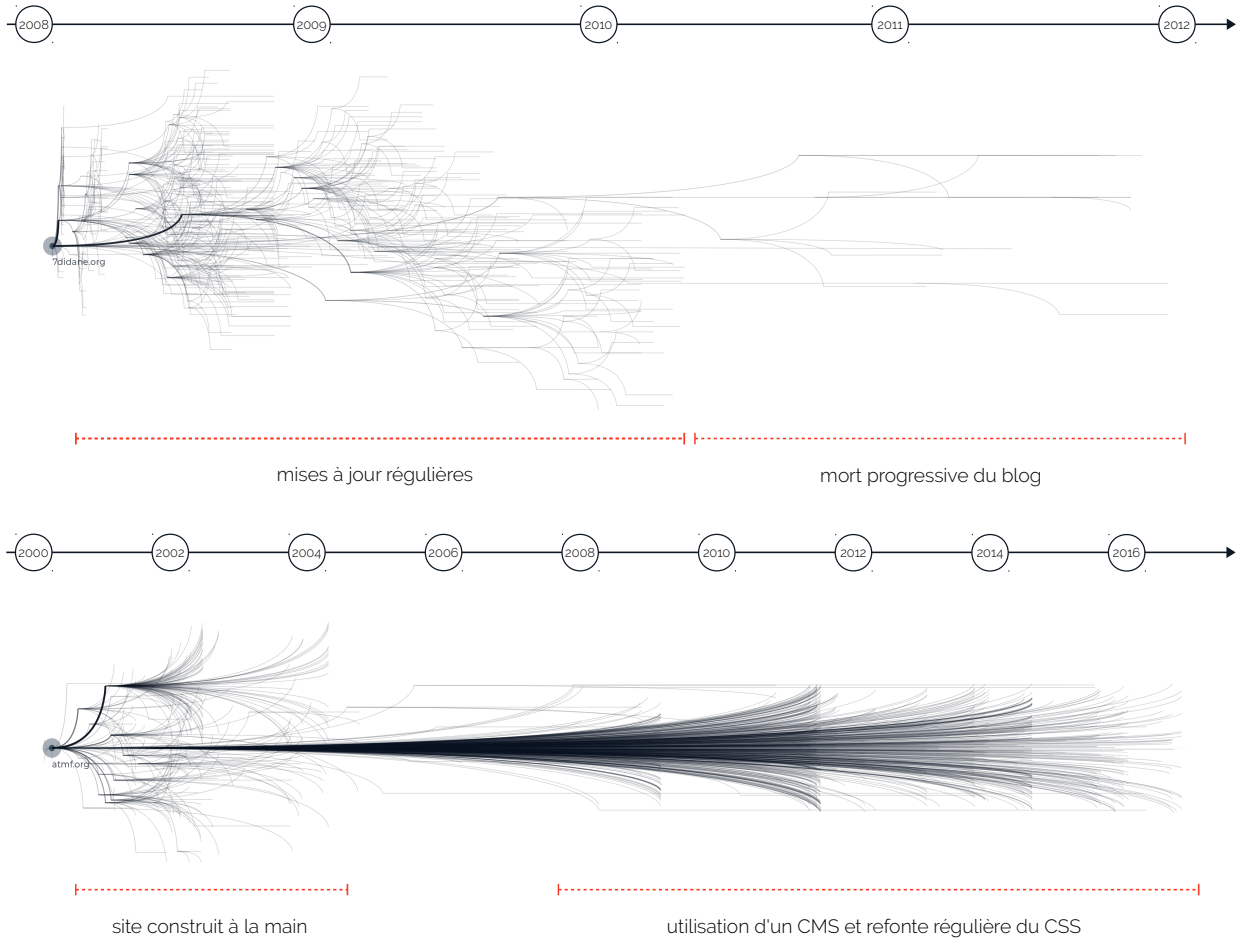


FIGURE 5.37: vue de l'évolution de la structure des sites *7didane.org* (haut), *atmf.org* (bas)

Enfin, ce chapitre aura été l'occasion d'apporter une contribution plus générale à l'étude du Web en tant qu'objet de recherches historiques. Nous proposons, ainsi, de construire une histoire du Web autour de l'analyse de certains de ses moments pivots. L'histoire de ces périodes de transition entre différents systèmes et manières de penser le Web, où les gestes des différents acteurs de la toile étaient chargés d'une intensité particulière. Intensité qu'il est aujourd'hui possible d'identifier à travers l'exploration de nos corpus d'archives Web. Mais, tout au long de ce chapitre, nos collectages n'ont pu témoigner que de la première étape du moment qui nous intéressait, à savoir la migration des blogs vers les réseaux sociaux.

Malheureusement, les archives Web ne seront ni jamais complètes, ni jamais impartiales et ni jamais parfaitement fidèles au Web vivant.

FIGURE 5.38: *atmf.org* (rouge) dans l'e-Diaspora marocaine

Aussi, dans la dernière partie de ce manuscrit, nous nous intéressons à d'autres formes d'archives Web, différentes de celles recueillis à l'INA ou chez Internet Archive et capturant de nouvelles traces numériques des migrations.

Chapitre 6

Au Delà des Archives Web

Le Chapitre 2 nous avait permis d'aborder les différentes techniques de collecte et de stockage des archives Web, chaque stratégie de préservation étant liée à une manière particulière de concevoir les archives, autant sur la forme (fichiers, systèmes d'exploration...) que sur le fond (nature de la ressource Web préservée).

Les réflexions des Chapitres 3, 4 et 5 ont émané de l'analyse d'un corpus d'archives Web déjà existant : les collectages de l'Atlas e-Diasporas, appartenant à la famille de l'**archivage côté client**, celle des crawlers et des scrapers (Section 2.2). Mais bien d'autres formes d'archives existent pour préserver les traces numériques des migrations.

Au cours de ce chapitre, nous présenterons deux autres corpus sur lesquels nous avons pu travailler dans le cadre élargi de notre participation au Diasporas Lab [1] : **1)** les archives des logs de navigation de l'Internet Libre de la Bibliothèque Publique d'Information du Centre Pompidou (BPI), **2)** les archives du formulaire des familles d'accueil potentielles du programme d'hébergement de personnes exilées Comme à la Maison (CALM). Le premier peut être rattaché aux stratégies dites d'**archivage des transactions**, centrées sur les interactions entre un site et un ou plusieurs utilisateurs. Il s'agira pour nous de tenter de comprendre les particularités et les caractéristiques du Web de la BPI, destiné, d'abord et avant tout, à un public précaire de personnes seule, isolée, souvent migrantes. Le second appartient à la famille des méthodes dites d'**archivage côté serveur**, cherchant à préserver les ressources du Web profond. L'idée, pour nous, sera de partir de l'étude d'une base de données (un formulaire de renseignement et ses évolutions) afin de saisir, à travers elle, la dynamique de l'accueil des personnes exilées chez des particuliers à la fin de l'été 2015. Nous chercherons, en particulier, à caractériser les familles accueillantes et leurs motivations, ainsi que l'environnement associatif et technique chargé de faciliter les mises en relation.

Ces deux sujets seront traités de manière plus synthétique que l'analyse des archives e-Diasporas, telle que nous l'avons abordé précédemment. Ce sont des travaux toujours en cours [2], faisant appel à un grand

1. Programme de recherche visant à enrichir les connaissances dans le domaine des migrations et mobilisant des méthodologies de valorisation spécifiques à l'exploration numérique, dirigé par D. Diminescu et prenant la continuité du programme TIC Migration, `http://www.diasporaslab.fr/`.

2. Ayant fait l'objet de communications préliminaires ou pour lesquelles des articles sont en cours de rédaction

nombre d'acteurs (institutions, chercheurs, étudiants...) et contraints par une inertie plus importante que celle des travaux principaux de cette thèse. Mais, nous tenons néanmoins à en faire ici la présentation, car nous pensons qu'ils peuvent enrichir le regard que nous portons sur les archives Web et sur les migrations en ouvrant de nouvelles pistes de réflexions.

6.1 Les archives de navigation Web de la BPI

Implantée à l'intérieur du Centre Pompidou, la Bibliothèque Publique d'Information (BPI) est la principale bibliothèque publique parisienne, accueillant plusieurs milliers de visiteurs uniques par jour [3]. Inaugurée le 31 janvier 1977 en plein cœur de Paris, la BPI offre à ses visiteurs la possibilité de faire des recherches documentaires sur des collections papier et audio-visuelles, réparties parmi les 3 étages qui la compose. La bibliothèque est aussi un lieu de travail et d'occupation quotidienne, où se croisent étudiants, travailleurs, indépendants, personnes en situations précaires, migrants, sans domiciles fixes, etc [4].

3. `https://fr.wikipedia.org/wiki/Bibliothèque_publique_d'information`

4. Cette section a fait l'objet d'une communication écrite : M. Amar, A. Camus, C. Evans & Q. Lobbé, (2016), *Études et recherches récentes sur les usages numériques à la Bpi*, Culture et Recherche n°134, 2016

Les postes Internet Libre

La BPI s'est toujours voulue pionnière de l'accès publique à l'information. S'il est maintenant possible de venir y connecter son ordinateur en Wifi, la BPI met en place dès juin 1995 (très peu de temps après la création du Web) des postes informatiques d'accès au World Wide Web (Amar et Béguet, 2008). Ce service, aujourd'hui connu sous le nom d'**Internet Libre** [5], met à disposition des visiteurs de la bibliothèque entre 200 et 250 postes [6] ouverts sur le Web. L'inscription est anonyme (sans demande de pièce d'identité, ni formulaire à remplir) et donne accès à une session limitée à 40 minutes.

5. `http://www.bpi.fr/informations-pratiques/les-services/internet-et-le-wifi`

6. L'effectif des postes est régulièrement réajusté par les services de la BPI.

FIGURE 6.1: Ticket de réservation d'une session de 40 mn sur l'un des postes Internet Libre de la BPI

Une fois avoir fait la queue, un ticket nous est remis pour nous orienter vers la machine qui nous est attribuée (Figure 6.1). Là, il suffit de se connecter avec le code indiqué pour ouvrir notre session. Un compte à rebours nous indique en permanence le temps qui nous est encore alloué. Néanmoins, cette limitation à 40 minutes est progressivement abandonnée au cours du printemps 2017. Les postes sont aujourd'hui en accès libre et sans limite de temps, une personne peut donc passer toute la journée sur une machine, si elle le souhaite.

Les postes sont répartis sur l'ensemble des trois étages de la bibliothèque, parmi les rayonnages des collections, à l'entrée ou le long des couloirs (Figure 6.2). Mais la consultations du Web à la BPI diffère d'une consultation *à la maison* en cela qu'elle est publique, les écrans étant visibles aux yeux de tous les visiteurs. C'est pour cette raison que

la connection Internet de la BPI est partiellement restreinte, les sites pornographiques, de pari en ligne, etc., sont ainsi filtrés.

Bien plus qu'une bibliothèque, la BPI est un environnement culturel ouvert à tous et devient, de fait, une maille importante du tissu social parisien[7]. Fréquentée par un public en partie constitué de personnes en situation précaire (seules, fragiles, sans domicile fixe, migrants...), bien souvent des hommes, les enquêtes de terrain dressent un portrait singulier du lien qu'entretient la BPI avec ses visiteurs (Paugam et Giorgetti, 2013).

7. Cette observation est commune à plusieurs bibliothèques et centres culturels parisiens comme à la Gaîté Lyrique ou à la Halle Pajol.

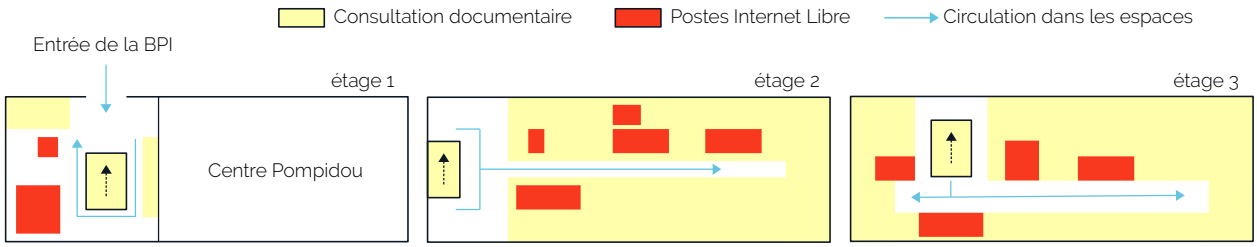


FIGURE 6.2: Implantation des postes Internet Libre (rouge) sur les trois étages de la BPI

Pour beaucoup, la bibliothèque est un refuge où l'on peut s'assoir, passer la journée et y voir du monde. Certaines zones sont dédiées à l'accueil et à l'accompagnement de ces personnes, comme l'espace auto-formation ou les cours de langues et d'initiation à l'informatique. Mais la plupart du temps, ce sont les précaires eux même qui, par leur présence, redéfinissent le fonctionnement de la bibliothèque. Ce phénomène est notamment visible sur les postes Internet Libre. Malgré la présence d'un accès aux ressources électroniques (journaux en lignes, base de données...) depuis ces machines, les observations font principalement remonter un usage *loisir* et *passe-temps* du Web, centré sur la consultation de vidéos en ligne, de sites de réseaux sociaux ou de résultats sportifs[8]. Néanmoins, il reste courant de voir des personnes mener, depuis ces postes, de véritables recherches approfondies ou encore se lancer dans des démarches administratives complexes (CAF, pôle emploi...) et qui pourraient, par exemple, faire l'objet d'un accompagnement dédié.

8. Voir l'intervention : *Les publics, in situ et en ligne*, C. Evans et M. Amar `https://www.dailymotion.com/video/x5mw8i7`

Ainsi, le Web de la BPI peut, aujourd'hui, être vu comme un **Web de première nécessité** (Pasquier, 2018). Tout l'enjeu de cette étude est donc de chercher à caractériser cet usage particulier de la toile à la bibliothèque. Le tout, partant d'une source de données archivée et centrée sur la relation site/usager : les logs de navigation Web.

Les logs de navigation

Dans le contexte sécuritaire de l'après 11 septembre, nombre de gouvernements se dotent d'un arsenal législatif renforcé en matière de captation, de conservation et de fouille de données électroniques personnelles.

En France, la loi du 15 novembre 2001 *relative à la sécurité quotidienne* [9] introduit le principe de rétention des données à des fins notamment d'enquêtes judiciaires et codifié dans le Code des postes et des communications électroniques (CPCE). L'article L.34-1 assimile, de fait, une bibliothèque comme la BPI à un opérateur de communications électroniques en cela qu'elle a pour objet d'« *offrir un accès à des services de communication au public en ligne [...] par l'intermédiaire d'un accès au réseau, y compris à titre gratuit* » [10].

Le CPCE prévoit, tout particulièrement, une obligation de conservation des données dites « *relatives au trafic* ». Il s'agit d'un ensemble de données exclusivement techniques, mais contenant néanmoins des « *informations permettant d'identifier [un] utilisateur* » [11] telles qu'une adresse IP. La durée de rétention de ces données est fixée à une année glissante. Le stockage est lui effectué par et chez l'opérateur de communication qui doit être en mesure de fournir les données à la justice. Aussi, la BPI garde-t-elle en permanence une année de logs de navigation sur ses serveurs, auxquels nous avons eu accès dans le cadre d'un projet de recherche commun avec C. Evans, M. Amar et A. Vigué-Camus du service Études et Recherches de la BPI.

9. https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000222052

10. https://www.legifrance.gouv.fr/affichCodeArticle.do?cidTexte=LEGITEXT000006070987&idArticle=LEGIARTI000006465770

11. Voir le décret d'application du 24 mars 2006 que la CNIL a, par ailleurs, qualifié d'imprécis https://frama.link/jcb9WapK

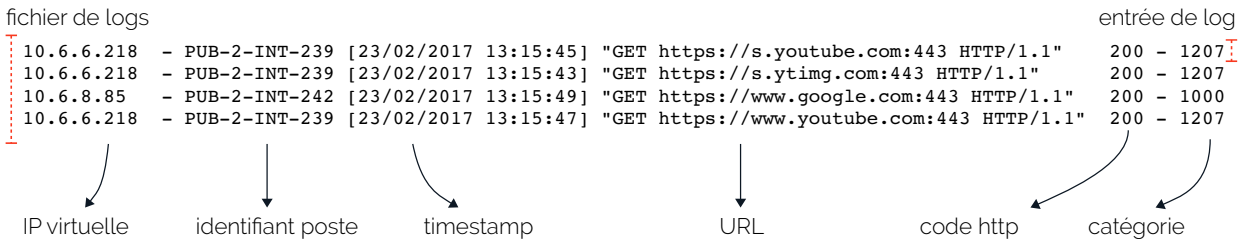


FIGURE 6.3: Exemple de logs de navigation générés par le proxy Web de la BPI

Un **log de navigation** est un enregistrement texte, généré par un proxy Web [12], dès qu'une page Web est chargée sur l'un des postes Internet Libre de la BPI. Or toute page Web est composée d'un ensemble de ressources associées, chacune, à une URL donnée (comme les publicités, les images, les vidéos...). Aussi, il est rare qu'une page s'affiche immédiatement, il lui faut souvent une poignée de secondes pour se charger dans son intégralité, élément par élément. Ainsi, chaque fois qu'un utilisateur visite une page Web à la BPI, cela se traduit, en base de données, par la création d'un ou plusieurs logs (Figure 6.4). Plus une

12. Logiciel intermédiaire entre le réseau de la BPI et le Web, utilisé notamment pour filtrer certains sites.

page Web aura de ressources hétérogènes à charger plus elle produira de logs.

Un log est une capture horodatée d'une action donnée (Figure 6.3). À la BPI, les logs de navigation font correspondre dates de visite et URL affichées. Dans le détail, nous pouvons y trouver les informations suivantes : une adresse IP virtuelle associée à chaque session de navigation, un identifiant physique de poste, l'horodatage du chargement de l'URL, l'URL visitée, son code de retour HTTP et, enfin, une catégorie. Les IP virtuelles ne sont pas uniques. Elle sont issues d'un lot d'adresses disponibles et redistribuées à chaque ouverture d'une session de navigation. En revanche, à un identifiant physique de poste correspond une et une seule machine. Cet identifiant permet de situer le poste Internet dans les locaux de la bibliothèque. Par exemple, l'identifiant `PUB-2-MUS-239` signifie : poste `publique`, étage `2`, espace `musique`, numéro `239`.

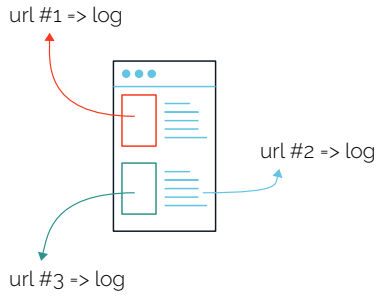


Figure 6.4: Une page Web peut générer plusieurs logs de navigation

Le proxy Web utilisé par la BPI est développé par la compagnie privée Olfeo. Il a pour particularité de fournir une catégorisation à la volée de chaque URL qu'il voit passer. Cette catégorisation se fait sur la base de masques de correspondances de noms de domaines, mis à jour et maintenus par Olfeo [13]. De fait, nous n'avons pas la main sur ces catégories, mais elles peuvent néanmoins servir à enrichir nos analyses ou à servir de filtres. Par exemple, le site *youtube.com* est catégorisé sous l'identifiant 1207 correspondant à l'appellation *Sites de Partage de Vidéo*.

13. Voir le détail des catégories d'URL ici `https://frama.link/ffuy3r1h`

Chaque log est donc une ligne de texte et plusieurs entrées de logs, mis les uns à la suite des autres, forment un fichier de logs. Ces fichiers sont générés quotidiennement, compressés, puis stockés sur les serveurs de la BPI. Pour résumer, les logs de navigations de la BPI archivent la relation entre un utilisateur donné et la page Web (adressée par une URL) qu'il visite.

Il existe plusieurs type de logs de navigation, plus ou moins détaillés, plus ou moins proches des actions de l'utilisateur ou du site Web. En effet, il est possible de générer des logs spécifiquement centrés sur le geste des internautes, souvent limité à un site donné, en collectant chaque clic ou chaque mouvement de la souris. Cette forme enrichie est, par exemple, utilisée pour analyser les usages du site Web *gallica.bnf.fr* par (Beaudouin *et al.*, 2016) qui déterminent plusieurs profils type d'utilisateurs *gallicanautes*. Mais, si les logs de la BPI ne bénéficient pas d'un tel niveau de finesse, ils nous offrent néanmoins un champ plus large de pistes de recherche. Ces logs sont balancés et équilibrés par leur simplicité, dans le rapport site/usager aucune des deux dimensions n'est favorisée. Ainsi, pour caractériser l'usage singulier du Web de la BPI, nous pouvons choisir de placer nos analyses, soit du côté des utilisateurs (profilage temporel, géographique...) soit du côté des sites Web (études topologiques, sémantiques...), soit en associant les deux.

Construire une chaine de traitements

Notre première tâche consiste à penser et mettre en place une chaîne d'extraction, d'analyse et de visualisation des logs de navigation [14]. En collaboration avec le service informatique de la bibliothèque, nous copions quotidiennement les logs de la veille vers l'espace de stockage sécurisé TeraLab [15], où un script automatique se charge d'ordonner les fichiers, de les décompresser, voire d'effacer les données antérieures à une année glissante.

14. Travail réalisé en collaboration avec C. Berrachdi, étudiante à Télécom ParisTech

15. `https://www.teralab-datascience.fr/fr/`

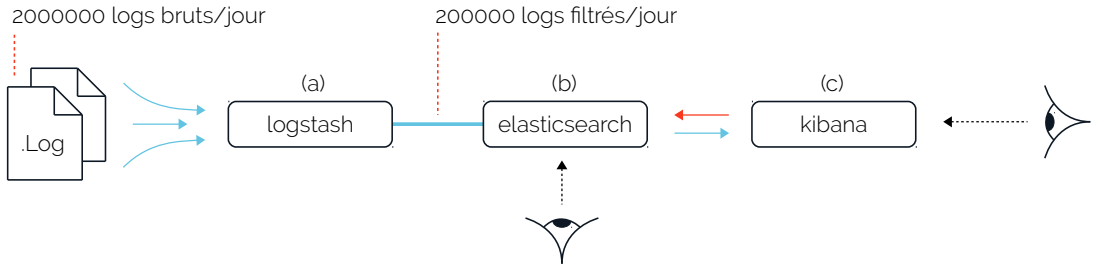


Figure 6.5: Chaine de traitement des logs de navigation de la BPI

Une fois copiés, les logs sont ensuite envoyés dans notre chaine de traitement, constituée de 3 briques logicielles distinctes [16] : **1)** le logiciel logstash se charge d'extraire les logs de leur fichier d'origine, de les nettoyer et de les filtrer **2)** le moteur de recherche elasticsearch indexe et rend interrogeable les logs **3)** l'interface Web kibana permet de réaliser des visualisations à la volée de nos données de navigation (Figure 6.5). Ici nous ne re-développons rien, nous agençons et configurons des solutions déjà existantes et performantes.

16. Logiciels open-sources, voir `https://www.elastic.co/fr/products/logstash`, `https://www.elastic.co/fr/products/logstash`, `https://www.elastic.co/fr/products/elasticsearch`, `https://www.elastic.co/fr/products/kibana`

Ouverte 6 jours sur 7 (de 12h00 à 22h00), la BPI génère entre 2 et 3 millions de logs de navigation par jours. Beaucoup de ces logs ne sont en réalité que du bruit. Or, nous souhaitons le plus possible nous rapprocher d'un ratio 1/1, soit un log généré par page visitée. Pour ce faire, nous filtrons d'abord les logs associés à des URL publicitaires en réutilisant à notre compte les listes des addblockeurs [17]. Certaines catégories du proxy Olfeo peuvent également servir de filtre. Enfin, pour gérer les multiples URL de chargement, nous choisissons d'agréger ensemble, les logs générés par un même nom de domaine, et ce, dans un intervalle inférieur à 4 secondes [18]. Ce faisant, en sortie du logiciel logstash (Figure 6.5, a), nous supprimons 9 logs sur 10 et envoyons au moteur de recherche une moyenne de 200 000 logs/jour. Il nous faut environ 3 heures pour traiter une journée de logs.

17. Par exemple `https://filterlists.com/`

18. D'une étude à l'autre, le temps moyen de chargement d'une page Web est compris entre 3 et 6 secondes.

Les données peuvent être interrogées via l'interface de visualisation kibana (Figure 6.5, c) pour faire, par exemple, du suivi quotidien (top 10 des sites les plus visités, occupation des postes par étage...), ou

encore, procéder à des analyses comparatives [19] (Amar, 2018). Mais, dans la suite de cette étude, nous choisirons d'explorer les données de manière plus large et plus souple, en travaillant directement depuis le moteur de recherche elasticsearch (Figure 6.5, b).

19. En cela, notre chaine de traitement vient remplacer un système existant, créé en 2004 par M. Renault, aujourd'hui obsolète.

D'un point de vue méthodologique, nous cherchons à ne pas couper le lien entre l'image du Web de la BPI, tel que nous le voyons à travers les logs, et la réalité du terrain. Au contraire, nous dressons des ponts entre le réel et le Web, afin de pouvoir questionner à tout moment telle ou telle représentation, tel ou tel résultat. Aussi nous mettons régulièrement en place, ce que nous appelons des **sessions de navigation** documentées et horodatées. Ces sessions ont deux finalités : **1)** vérifier la correspondance entre le ressenti d'une navigation Web derrière un écran et l'écart possible avec les logs archivés **2)** s'immerger parmi les utilisateurs de l'Internet Libre, se mettre dans une position d'observateur, pour déterminer de futures pistes de recherche.

Des profils de navigation Web

Lorsqu'il s'agit de caractériser des utilisateurs partant de traces de navigation, les recherches s'orientent rapidement vers l'étude de profils et de parcours communs (Feng *et al.*, 2006). Quels sont les grands motifs de navigation dominants ? Y-a-il des effets de périodicité temporelle (à la semaine, en fin de journée...) ou de territorialité (par étage ou par effet de voisinage...) ? Cette piste est d'autant plus privilégiée qu'elle pourrait permettre de vérifier par les données, les observations du terrain, à savoir un Web de la BPI avant tout tourné vers le loisir et le passe-temps.

L'idée est ici de ramener les logs à la notion de trajet individuel d'une page à une autre et de voir, par des méthodes de classification, s'il n'existe pas des trajectoires communes à tous les utilisateurs de la BPI [20]. Comme nous sommes dans une phase de découverte des données, notre première stratégie est de nous tourner vers des techniques non supervisées, sans apprentissage. La méthode dite **k-means** [21] permet par exemple de déterminer le centre d'une communauté d'individus donnés représentés par des vecteurs.

20. Travail réalisé en collaboration avec M. Pierru, étudiante à Télécom ParisTech.

21. K-moyennes en français `https://fr.wikipedia.org/wiki/K-moyennes`

Pour initialiser l'algorithme k-means, il faut lui indiquer le nombre de centres recherchés, nombre qui peut être déterminé en utilisant la méthode dite du coude [22]. Ces centres sont positionnés aléatoirement dans l'espace vectoriel (Figure 6.6, b), puis l'algorithme procède par itération. La distance entre les centres et les individus est calculé pour permettre une première assignation (Figure 6.6, c). Les centres se déplacent au milieu de ces communautés temporaires et la distance est à nouveau calculée. L'algorithme chemine ainsi jusqu'à que les centres

22. `https://en.wikipedia.org/wiki/Elbow_method_(clustering)`

ne se déplacent plus (Figure 6.6, c), déterminant ainsi les communautés et leurs individus moyens (ie : centroïdes).

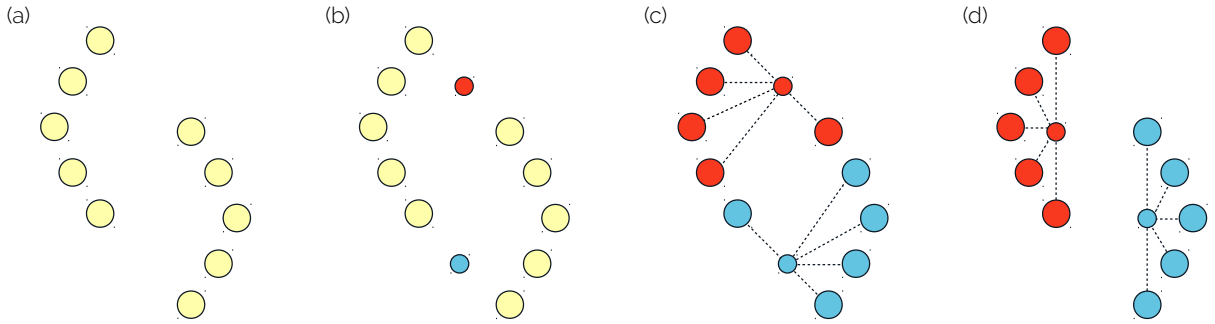


Figure 6.6: Exemple de fonctionnement de l'algorithme k-means avec 2 centres (bleu et rouge)

Avec nos logs, nous commençons par créer un identifiant de session unique pour regrouper les pages visitées par un même individu : soit la concaténation de l'IP virtuelle, de l'identifiant de poste et de la date. Chaque session est ensuite représentée par un vecteur ayant autant de dimensions que de catégories Olfeo existantes. Les parcours sont déterminés par la catégorisation des URL. Nous choisissons, dans un premier temps, de ne pas prendre en compte la dimension temporelle à l'intérieur des sessions, c'est-à-dire le fait qu'une page est visitée après ou avant une autre. Il existe néanmoins des exemples de clustering de sessions de navigation par k-means temporel (Stevanovic *et al.*, 2011). Aussi, pour chaque log de chaque session, nous incrémentons la coordonnée de la catégorie correspondante dans le vecteur. Nous testons cette méthode sur une semaine de données, du 20 au 26 mars 2017, semaine à priori neutre et sans événement particulier.

Le nombre de centres varie d'un jour à l'autre, entre 2 et 4 centroïdes. La Figure 6.7 présente les résultats sous la forme d'une matrice : les 11 catégories d'ULRs les plus représentées en ordonnée, les jours en abscisse. Pour chaque journée, nous ne conservons que les deux classes les plus importantes (les colonnes) et pour lesquelles nous indiquons la proportion de logs correspondants.

Deux grands profils se dégagent, à savoir un profil fait exclusivement de consultations de sites de partage de vidéos (Youtube, Dailymotion...) et un profil fait de consultations vidéos partagées en proportion avec les sites de réseaux sociaux (Facebook, Twitter...). L'observation du terrain est donc confirmée par les données. Notons le cas particulier du mardi, où la bibliothèque est fermée et où les seules consultations Web sont celles des gardiens. Notons aussi, la sur-représentation des sites de traduction le lundi 20 mars 2017, qui après fouille dans les

données, s'avère être liée à des recherche russophones. Là encore, cette information confirme l'observation dans la bibliothèque d'une grande communauté d'usagers d'Europe de l'est et russophiles. La catégorie *services aux entreprises* s'avère, après recherches, être un ensemble d'URL et de sites de types *analytics*, traceurs ou cookies. Cela ne nous renseigne pas beaucoup sur le profil des usagers, mais plutôt sur la nature du Web en général, qui tend à générer pléthore de statistiques et de sondes sur le dos des internautes.

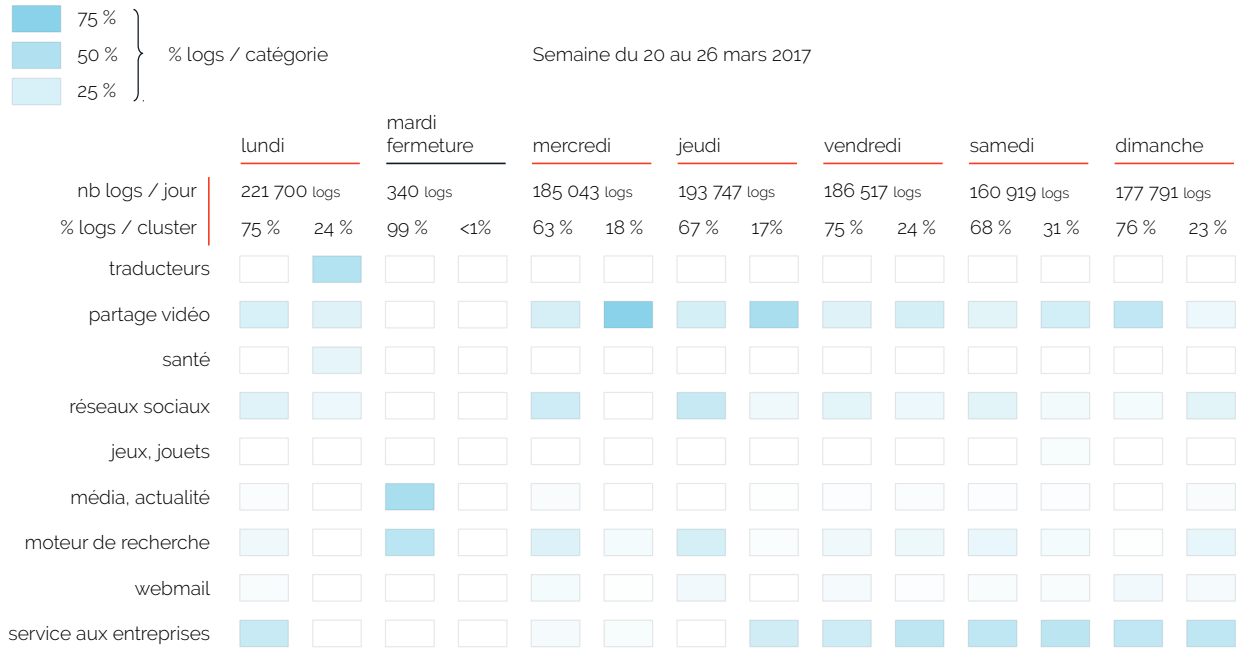


FIGURE 6.7: Classification des sessions de navigation Web par la méthode des k-means pour la semaine du 20 au 26 mars 2017

En mars 2017, les sessions de navigation étaient encore limitées à 40 minutes. Aussi, nous relançons une seconde fois notre algorithme en choisissant maintenant une semaine où l'Internet Libre est en accès entièrement ouvert : du 22 au 28 mai 2017. Mais les résultats sont sensiblement les mêmes, à ceci près qu'ils sont moins précis puisqu'avec la fin des sessions, un visiteur n'est plus obligé de se déconnecter en quittant un poste informatique. À une session peut donc correspondre plusieurs personnes. Notre identifiant de session pourrait être ici amélioré en ajoutant la détection d'un temps de non-activité entre deux usagers. Nous testons également des différenciations par moments de la semaine, de la journée ou par étage, mais là encore les deux mêmes profils ressortent à nouveau.

Cela peut s'expliquer par la simplicité de la méthode employée, une stratégie plus fine qui incorporerait le temps dans ses calculs, comme l'utilisation des chaînes de Markov [23] (Cadez *et al.*, 2003), pourrait nous

23. `https://fr.wikipedia.org/wiki/Chaîne_de_Markov`

donner des résultats plus variés. Mais là encore, ce qu'il nous serait donné de voir, ne pourrait être que la norme et la moyenne, renvoyant à notre discution sur l'AED en Section 5.1. Or l'étude de la longue traine et des profils minoritaires pourrait s'avérer plus intéressante. En effet, il y a un biais majeur dans cette analyse, à savoir que les sites Web ne sont pas *verbeux* de la même manière. Des sites emplis de publicités (même si nous en supprimons beaucoup), de vidéos ou de traceurs vont mécaniquement générer plus d'URL, et donc plus de logs, qu'un site modéré comme le portail d'une administration ou des sites gouvernementaux. Aussi, si un profil semble faible ou écrasé par les mastodontes Youtube et Facebook, cela vient très certainement de notre méthode de comptage et de la nature de ces archives.

Des dimensions particulières

Pour caractériser le Web de la BPI, il peut être intéressant de regarder du côté du contenu des pages visitées, dont nous connaissons les URL. En effet, l'étude directe des usagers précaires de la bibliothèque, via leurs profils de navigation, ne semble nous mener que vers la confirmation de grandes observations. Or, de manière indirecte, ce qui est lu ou ce qui est vu sur les postes Internet Libre, sont autant d'indices à extraire des pages Web pour dresser une image globale de l'environnement complexe qu'est le Web de la BPI.

Sur ce point, les enquêtes terrains menées par le service Études et Recherche de la bibliothèque et l'expérience que nous commençons en acquérir, nous conduisent à revoir nos priorités d'exploration et à penser une analyse des logs **par dimensions** particulières. Comme la masse des logs générée par les sites verbeux écrase les cas singuliers, nous nous proposons de construire des sous-ensembles de données, suivant une ligne thématique dictée par nos observations *in situ*. Par exemple, les URL des sites de traduction, contiennent, en elles mêmes, les mots que l'usager a cherché à traduire :

```
http://www.linguee.fr/francais-anglais/search?source=auto&query=mémoire
```

Figure 6.8: Exemple d'url du site de traduction linguee, pour le mot mémoire, du français vers l'anglais

Ces informations, agrégées, à l'échelle d'une semaine, d'un mois ou d'une année peuvent nous renseigner indirectement sur la nationalité de certains usagers ou sur leurs préoccupations de traduction : s'agit-il de termes courants ou de termes techniques ? Aussi, comme base de nos futures recherches par dimensions, nous assemblons plusieurs sous-ensembles de données, soit en déconstruisant les URL, soit en scrapant directement les pages Web pour ainsi obtenir : les termes

traduits, les requêtes adressées aux moteurs de recherches (Google, Bing...), les titres des articles des sites de news et d'actualité, les titres et les descriptions des vidéos visionnées...

À ce niveau, nous n'en sommes qu'aux prémisses de nos recherches. Dans le cadre d'une collaboration avec N. Gaumont et les outils développés pour le projet Politoscope [24], nous pouvons par exemple déterminer la couleur politique des sites de news français visités à la BPI : plutôt neutres mais néanmoins en accord avec la moyenne des articles postés sur le Twitter politique français (Gaumont *et al.*, 2017).

24. https://politoscope.org/

Des expérimentations [25] plus ouvertes sont également menées, à la manière d'un G. Legrady qui propose de visualiser certaines dimensions singulières des données d'emprunts de la bibliothèque municipale de Seattle (Legrady, 2005).

25. Travail réalisé en collaboration avec A. Sciberras, A. Habis, T. Mokart, J. Grenier, M. Arrivat, A. Saheb, M.A. Chamli, P. Bernat, étudiants à Télécom ParisTech.

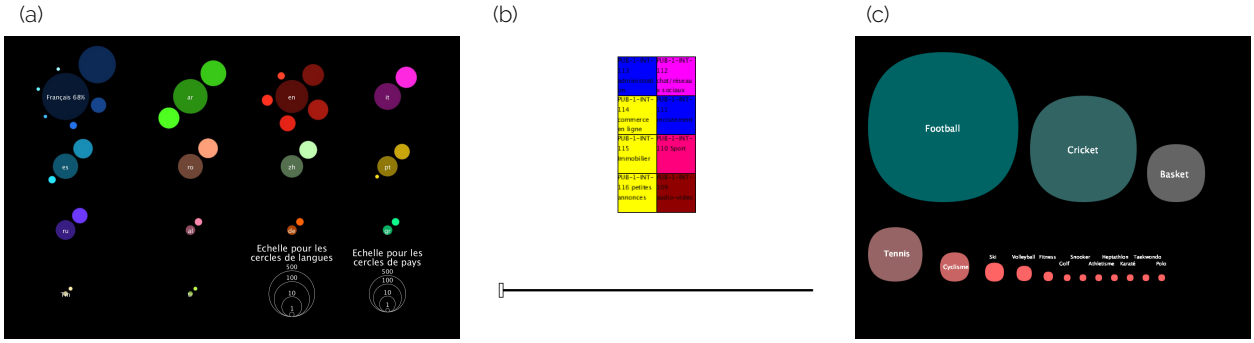


FIGURE 6.9: Prototypes de systèmes de visualisations, basés sur des dimensions particulières des logs de la BPI

Ainsi, entre autres prototypes sur lesquels nous travaillons, nous pouvons citer un système de visualisation des catégories d'URL par ilots de postes (Figure 6.9, b), une analyse des langues utilisées par les sites de news (où l'on voit la place importante de l'arabe à travers le site d'Aljazeera notamment, Figure 6.9, a) , ou encore une expérimentation sur les pages sportives, révélant un intérêt non négligeable pour les compétitions de cricket comparé au football [26] (Figure 6.9, c).

26. Données extraites des mois de mai & juin 2018, couvrant une coupe du monde de football et la finale de la ligue de cricket indienne.

Ces travaux exploratoires, finiront certainement par révéler de nouvelles pistes de recherches, pour continuer à étudier le Web de la BPI, ses visiteurs et la manière dont ils naviguent sur la toile. Une analyse à venir des sites de la CAF ou de Pôle emploi, pourront peut être déboucher sur la mise en place de services d'accompagnement dédiés.

Mais plus nous laissons le temps passer et plus les données du Web se protègent de nos explorations. En effet, le taux de sites Web sécurisés augmente chaque jour et les URL en *https* [27] empêchent le proxy de la BPI d'enregistrer l'entièreté des adresses hypertextes : sur une journée moyenne de logs [28], 88% des sites Web sont identifiés comme protégés. Par exemple, toutes les URL du site *youtube.com*, contenues dans les logs,

27. https://fr.wikipedia.org/wiki/HyperText_Transfer_Protocol_Secure

28. Données de juillet 2018

sont tronquées après le nom de domaine. Nous ne pouvons donc plus avoir accès à l'adresse exacte d'une vidéo ou d'une page personnelle. Ainsi en va-t-il du Web et de ses archives, lorsque le premier évolue les secondes deviennent momentanément obsolètes faute d'intégrer elles même ces nouvelles mutations.

6.2 Les dynamiques de l'accueil des réfugiés (2015-2017)

Certaines zones de la toile se trouvent hors de ce qu'il est coutume d'appeler **le Web indexable** (*the indexable Web*) (Lawrence et Giles, 1998), c'est-à-dire la surface du Web rendue accessible par le travail des crawlers des moteurs de recherche généralistes (Google, Yahoo...). La partie immergée du Web, dissimulée derrière les logins et les mots de passes, traduit une réalité technique de la toile : la grande majorité des formulaires, des services Web et des diverses bases de données, sur lesquels s'appuient les sites, ne sont pas adressables depuis la structure hypertexte du Web (Senellart, 2007). Ainsi, le **Web profond** (Bergman, 2001) représente, selon une étude publiée en 2001, 500 fois la taille du Web de surface [29] et constitue, de fait, un défi pour les archivistes du Web cherchant, autant que faire ce peut, à le préserver (Masanès, 2006a).

29. `http://archive.wikiwix.com/cache/?url=http://www.press.umich.edu/jep/07-01/bergman.html`

Dans cette section, nous partirons de l'archivage d'un de ces formulaires du Web profond pour tenter de comprendre, à travers l'analyse de ses évolutions successives (structure et contenu), un moment récent de l'histoire des migrations : la *crise des réfugiés* de l'été 2015 et les formes multiples d'accueil qu'elle a engendrées ou réactivées [30].

30. Cette section fait l'objet d'une publication en cours de rédaction : D. Diminescu, Q. Lobbé, *Les mots de l'accueil* (titre provisoire).

Été 2015 : organiser l'accueil

À l'été 2015, les médias et les réseaux sociaux regorgent de reportages sur les réfugiés en détresse, aux portes de l'Europe. Qualifiée de *crise migratoire* sans précédent, cette vague d'arrivées trouve son origine dans la persistance des conflits en Irak et en Syrie. Avec l'échec de la révolution citoyenne de 2011 qui se transforme, petit à petit, en une guerre civile meurtrière, ce sont près de 5 millions de Syriens qui sont forcés de franchir leurs frontières, pour fuir vers les territoires voisins : 3 millions d'entre eux s'installent dans des camps en Turquie et le Liban accueille l'équivalent humain du tiers de sa population (Blanchard et Rodier, 2016).

Par la suite, au début du printemps 2015, des milliers de personnes commencent à débarquer quotidiennement sur les iles de la mer Égée, au prix d'une traversée mortelle, partant des côtes turques et transformant ces eaux en un cimetière à ciel ouvert. Ce faisant, devant l'ampleur

d'un phénomène qu'il est incapable de gérer seul, l'État grec décide rapidement de financer le transfert des réfugiés vers Athènes (Pillant, 2016). Point à partir duquel beaucoup décident de suivre la *route des Balkans*, espérant trouver un abri en Europe de l'ouest, notamment en Allemagne ou en Angleterre. L'espace de quelques mois, les frontières balkaniques deviennent ainsi *facilement* franchissables, et ce malgré une couverture médiatique grandissante, augmentant les chances de se faire arrêter en chemin.

Mais cet état de fait ne dure pas longtemps. La Hongrie s'enferme derrière des kilomètres de barbelés et les contrôles aux frontières sont rétablis au sein même de l'espace Schengen. Plusieurs gouvernements d'Europe centrale choisissent de condamner les points de passage dont ils ont le contrôle. Ainsi, après l'échec des processus de relocalisation [31], l'Union Européenne (UE) finit par verrouiller ses frontières extérieures, avec la mise en place de hotspots, dès le mois d'octobre 2015 : des lieux de privation de liberté alors présentés comme des *centres d'identification et de réception* pour les réfugiés.

31. Transfert, depuis l'Italie et la Grèce, de personnes éligibles aux droits d'asile, et ce, vers d'autres pays de l'Union Européenne (Tervonen, 2016).

Bien que le Haut-Commissariat des Nations unies pour les réfugiés (HCR) comptabilise, pour la seule année 2015, près de 65 millions de personnes victimes de migrations forcées dans le monde [32], l'arrivée, à la même période, de ces réfugiés en Europe reste pourtant difficile à estimer. Sur ce point nous pouvons, tout de même, nous tourner vers les statistiques de l'agence Eurostats [33] portant sur le nombre de demandes d'asile, déposées dans l'UE cette année là. Soit 1 255 640 primo-demandeurs sur l'ensemble des 28 États membres (en augmentation de 123% par rapport à 2014). L'Allemagne en accueille à elle seule 35,2%, devant la Hongrie (13,9%) et la Suède (12,4%).

32. Demandeurs d'asile, réfugiés enregistrés, personnes apatrides ou déplacées dans leur propre pays.

33. Voir `https://frama.link/rb9m8yRD`

Loin derrière, mais sans pour autant rester totalement à l'écart, la France enregistre en 2015 plus de 70 000 nouvelles demandes d'asile sur son territoire, le taux de reconnaissance du statut de réfugié [34] atteignant 31,5%. Or derrière chacune de ces demandes se trouve une personne à accueillir et à prendre en charge. Ainsi, parallèlement aux services de l'État, se développe un tissu associatif qui s'oriente vers l'organisation et la mise en place de l'accueil des réfugiés en France.

34. Chiffres de l'Office français de protection des réfugiés et apatrides (Ofpra) `https://www.ofpra.gouv.fr/fr/l-ofpra/actualites/premiers-chiffres-de-l-asile-en`.

À cette époque, nombre d'innovations sociales et numériques (ISN) cherchent à résoudre par la technologie différents problèmes sociaux (Gossart *et al.*, 2018). Parmi elles, l'association Singa [35] propose, à travers son programme *Comme à la maison* (CALM) [36], de mettre en relation des familles françaises souhaitant accueillir chez elles une personne ayant déposé une demande d'asile. L'accueil proposé par Singa ne cherche pas à répondre à une situation d'urgence, mais à s'inscrire dans le temps long pour faciliter l'intégration des personnes, l'hébergement allant de 3 à 6 mois. La médiation entre les particuliers **accueillants** et les réfugiés **accueillis**, se fait d'abord en remplissant un formulaire

35. `https://www.singafrance.com/`

36. Programme dont la création a été accompagnée et documentée par le Diasporas Lab.

en ligne[37], c'est ensuite au personnel de Singa que revient la tâche de créer les pairs accueillants/accueillis. Ainsi, partant du cas particulier de l'association Singa, nous chercherons dans la suite de cette section, à interroger la manière avec laquelle le numérique s'inscrit dans le social et contribue potentiellement à son changement.

37. `http://calm.singa.fr/inscription/accueillant/`

Les archives du formulaire des accueillants

Dans le cadre d'un projet de recherche, monté en partenariat avec l'équipe en charge de CALM, nous avons eu accès, entre 2015 et 2017, aux enregistrements des formulaires d'inscriptions et à leurs évolutions successives[38].

38. Données anonymisées par les membres de l'association.

L'inscription se déroule comme suit : lorsque l'on se connecte sur la plateforme CALM, il convient d'abord d'indiquer si l'on est un futur accueillant ou un futur accueilli, le système nous dirige ensuite vers un formulaire dédié (Figure 6.10). L'inscription est individuelle autant pour les accueillis que pour les accueillants, même si pour ces derniers, on sous-entend plus généralement la notion de foyer (une famille, un couple, voire une personne seule...). Ces formulaires s'appuient sur la technologie gratuite Google Forms[39] et se présentent comme une suite de questions appelant des réponses ouvertes (d'expression libre) ou fermées (à choix multiples ou listés).

39. Application Web permettant de créer et de partager à la volée des formulaires en ligne `https://docs.google.com/forms/u/0/`.

Figure 6.10: Premières questions du formulaire accueillants

Dans les faits, le formulaire des accueillis n'a jamais été réellement mis en place, si ce n'est durant quelques jours, au début de l'automne 2015. Durant ce court laps de temps, le service est assailli et croule rapidement sous plusieurs centaines d'inscriptions. Aussi, faute de pouvoir gérer la demande, le formulaire est quasi-immédiatement fermé, l'équipe de CALM proposant plutôt aux futurs accueillis de se déplacer dans les locaux de l'association pour les rencontrer directement. En revanche, le formulaire des accueillants reste lui accessible et fait l'objet de mises à jour régulières par l'association. Nous dénombrons, ainsi, 4 versions successives : **1)** avant juin 2015 **2)** de juin 2015 à novembre 2016 **3)** de novembre 2016 à juin 2017 **1)** après juin 2017.

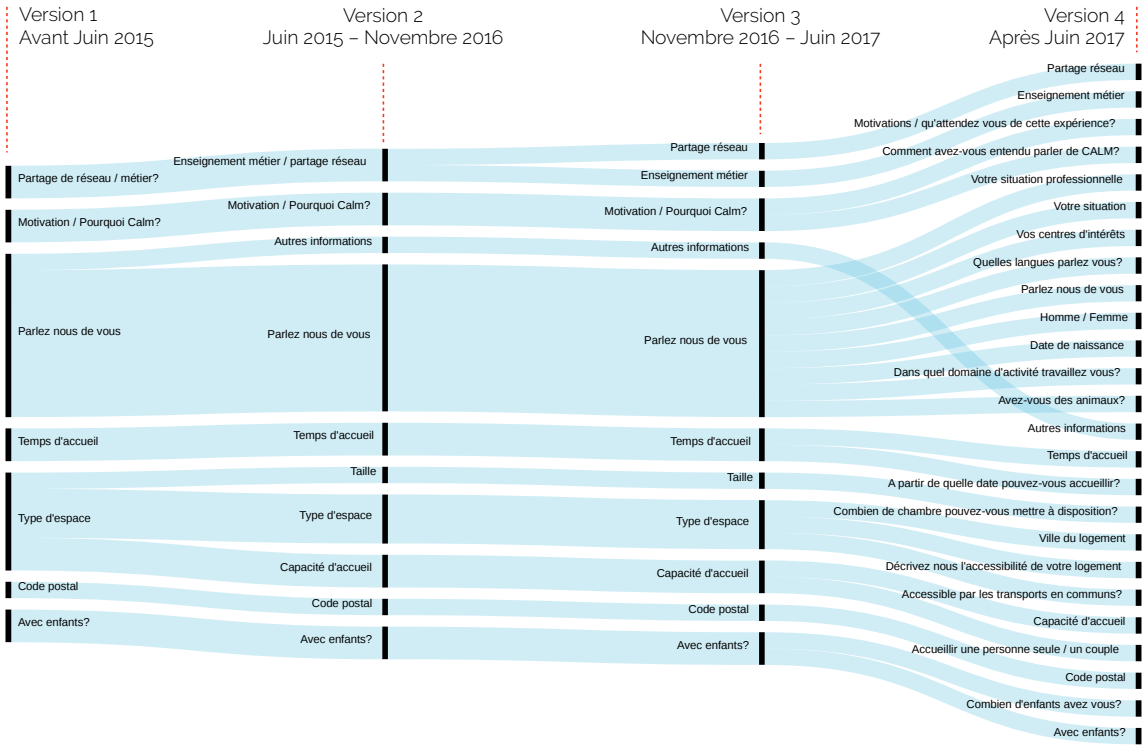


Figure 6.11: Évolution de la structure du questionnaire des accueillants

Partant de là, notre première tâche, lorsque nous entrons en possession de ces archives, consiste à les nettoyer, à les structurer et à les uniformiser [40]. En effet, il est courant que, d'une version à l'autre, la formulation d'une question puisse évoluer. Certaines apparaissent, d'autres disparaissent. Par exemple, la question de l'âge d'un accueillant est ramenée, progressivement, à sa date de naissance. Or, si l'information finale reste la même, la donnée à traiter, soit par un humain, soit par un script automatique, est différente. Il nous faut donc : ou bien transformer les âges en date, ou bien faire l'opération inverse afin de mener, comme nous l'envisageons maintenant, une analyse longitudinale de l'ensemble de

40. Travail réalisé en collaboration avec E. Barnoud, S. Dahan, M. Garnier, T. Glasser, N. Leclerc, K. Nguyen, C. Ruppli, M. Verstraete, étudiants à Télécom ParisTech.

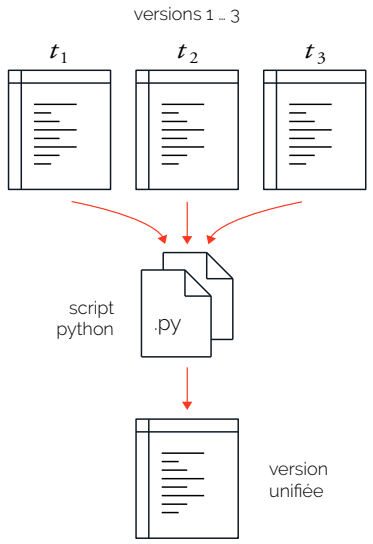


Figure 6.12: Unification des différentes versions d'une même base de données

ces inscriptions. Pour ce faire, nous développons une première chaîne semi-automatique d'unification des données (Figure 6.12).

Mais avant de plonger dans l'exploration de l'archive maintenant nettoyée, commençons par observer l'évolution de sa structure au cours du temps, telle que nous l'expose la Figure 6.11. Contrairement à l'historique des sections d'un forum (Figure 5.20) qui peut être révélateur de grandes ruptures ou de grands changements, la structure du formulaire CALM s'inscrit plutôt dans une forme de continuité. Il n'y a pas d'apparition de nouvelles questions à proprement parler, mais, une tendance générale à la fragmentation et la spécialisation de questions ouvertes (*Parlez nous de vous*, *Type d'espace*...) en une suite de champs plus restreints (*Quelles langues parlez vous ?*, *Combien de chambres pouvez vous mettre à disposition ?*...).

À travers l'évolution du formulaire, ce sont les descripteurs avec lesquels sont qualifiés les accueillants, en base de données, qui se transforment et se rationalisent : d'un point de vue d'informaticien, il est ainsi plus simple de gérer à grande échelle un paramètre clairement défini et quantifiable (l'âge, un nombre de chambre, une liste de langues...), qu'un champ d'expression libre, d'où il faudra extraire une information potentiellement mal formulée, voire inexploitable. La Figure 6.11 traduit, de fait, une évolution d'ingénierie, pensée pour améliorer la gestion et l'exploitation des données d'inscription. Aussi, cette discrétisation d'un individu en un ensemble de descripteurs normés sous-tend ce que A. Rouvroy et T. Berne nomment **gouvernementalité algorithmique**, soit :

« *un certain type de rationalité (a)normative ou (a)politique reposant sur la récolte, l'agrégation et l'analyse automatisée de données en quantité massive de manière à modéliser, anticiper et affecter par avance les comportements possibles* » — (Rouvroy et Berns, 2013, p.11)

La gouvernementalité algorithmique est une stratégie de rationalisation et de transcription du réel sous forme de données numériques exploitables par les systèmes informatiques. L'incertitude générée par les comportements humains est effacée (ici les champs d'expression libre) au profit d'une information spontanée et objective, qu'il n'est plus nécessaire de discuter. Aussi pourquoi l'association Singa a-t-elle souhaité rationaliser à ce point les modalités de gestion de ses données d'inscriptions ? Pourquoi modéliser le profil des familles d'accueillants ?

Né du hackathon[41] *Réfugiés Connectés* co-organisé par Singa et par le Diasporas Lab en février 2015, le programme CALM est, dans un premier temps, géré par des personnes qui ne se considèrent pas comme des développeurs informatiques (Figure 6.11, versions 1 & 2). Par le choix d'une infrastructure technologique simple et accessible (Google

41. Évènement de courte durée réunissant développeurs, experts industriels et associatifs pour participer au développement d'un projet collaboratif (Lobbé, 2018)

Forms et feuilles de calcul Excel), servant de base à une mise en relation manuelle des couples accueillants/accueillis[42], l'association favorise d'abord un travail humain, qualitatif et fondé sur l'expérience acquise par son équipe. Chaque médiation est l'objet d'une discussion singulière entre familles et réfugiés.

42. Les membres de l'association fouillent à la main les inscriptions dans leur base de données pour détecter de bons candidats à l'accueil

Cependant, quelques mois plus tard (Figure 6.11, versions 3 & 4), devant faire face à une augmentation des inscriptions au programme, Singa décide de recruter un développeur pour CALM. Au cours d'entretiens organisés avec ce dernier, nous apprenons qu'il travaille alors sur la conception d'un système de « *matching automatisé* » des accueillants et des accueillis, pour épauler, voire remplacer, à terme, le travail fastidieux des membres de l'association. En favorisant des questions précises et non bruitées dans les formulaires, il pense dégager « *une trentaine de paramètres* » décrivant finement le profil des accueillants. La base de données sera, de fait, adaptée à un futur traitement algorithmique. Son système de suggestion automatisé proposera alors « *un sous-ensemble qualifié de familles d'accueillants* qui, une fois validé par l'accueilli, seront à même de discuter ensemble « *dans un salon en ligne privé, hébergé sur la plateforme* ». Ce faisant, le développeur pousse au maximum la logique d'automatisation de la mise en relation, avec une médiation devenue purement algorithmique, supposée neutre ou juste, car rendue invisible et sans intermédiaire humain. L'incertitude disparait.

Mais la collaboration entre le développeur et l'association n'aboutit pas : la mise en production du système est annulée, incapables de faire converger leurs visions respectives de l'outil. En effet, Singa comprend rapidement que la mise en relation humaine des réfugiés et des familles doit rester au cœur de leur savoir-faire. Elle ne doit pas être déléguée à une machine, car cette expertise est vecteur de confiance et de liens entre les deux parties.

En cela, Singa et CALM font office d'exception dans le paysage des innovations sociales et solidaires, émergeant à mesure que la crise des réfugiés s'aggrave. Ils sont ainsi nombreux[43], dans la veine de l'initiative Techfugees[44] à se lancer, par altruisme ou par simple opportunisme, dans la création de toutes sortes d'applications et de services automatisés à destination des personnes réfugiées : diffusion d'informations, appels d'urgence, aide à la navigation, traduction... Pour D. Diminescu, l'année 2015 doit être considérée comme un véritable **moment technophorique**, où l'on voit exploser l'intérêt pour ce type d'innovations.

43. Dans un rapport encore non publié, D. Diminescu et L. Macias recensent l'apparition de 82 applications mobiles d'aide aux réfugiés sur cette période.

44. `https://techfugees.com/`

Mais nous tenons à souligner que cette forme de solutionisme technologique (Morozov, 2013) ouvre une brèche potentiellement dangereuse, qui nous oblige à interroger les modalités de gestion, de traitement et de protection des données numériques personnelles des populations réfugiées. En effet, ces multiples applications ont accès à des informa-

tions extrêmement sensibles (des noms, des adresses, des localisations GPS...), qu'elles stockent et accumulent sans qu'aucun consensus technique ou éthique n'ait encore été établi et partagé à grande échelle. Aussi, comment mettre en garde et sensibiliser les personnes réfugiées vis-à-vis d'une utilisation malveillante ou maladroite de leurs données personnelles par un tiers ? Comment concevoir et penser des protocoles de recherches, basés sur ces données et qui seraient à la fois sécurisés et sécurisants ? Ces questions restent ici ouvertes, mais nous espérons à l'avenir pouvoir mener plus avant nos réflexions.

Les mots de l'accueil

Nettoyé et unifié, le contenu de l'archive peut maintenant faire l'objet d'une exploration longitudinale. De juin 2015 à juin 2017, Singa met en place l'accueil de près de 300 personnes réfugiées et enregistre, dans le même temps, 11 892 inscriptions d'accueillants sur sa plateforme. Mais la distribution temporelle de ces inscriptions est loin d'être uniforme, le mois de septembre 2015 comptabilisant quasiment 10 000 propositions d'accueil. La Figure 6.13 donne le détail des inscriptions par jour, du 1er au 15 Septembre 2015.

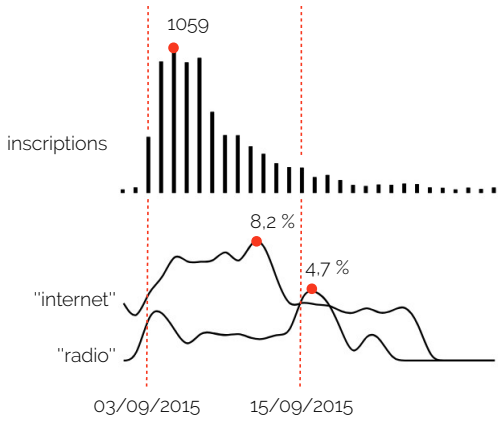


Figure 6.13: Inscriptions par jour au programme CALM (Septembre 2015) et pourcentage de d'utilisation des mots *internet* et *radio* dans le champ *Motivations ?*

En aout 2015, peu de personnes s'intéressent au programme CALM, le plus souvent cantonné à un milieu militant et sensible à la cause des réfugiés. Pourtant, les inscriptions d'accueillants explosent le 3 septembre 2015, passant de quelques dizaines à plusieurs milliers en l'espace d'une semaine. La diffusion mondiale, par la presse et la télévision, de la photographie du corps du jeune A. Kurdi [45], retrouvé noyé sur les côtes turques, suscite une grande émotion et génère une vague d'inscriptions sans précédent sur CALM. Ce faisant, nous tirons de la base de données plusieurs informations démographiques basiques que nous confrontons avec des entretiens d'accueillants et enrichissons

45. `https://fr.wikipedia.org/wiki/Alan_Kurdi`

par la lecture des champs d'expression libre *Motivations ?* et *Parlez nous de vous*.

Les nouveaux accueillants vivent, ainsi, dans les grands foyers urbains français, proposant de simples pièces dans de petits appartements, voire uniquement un canapé. Ceux qui résident à la campagne mettent parfois à disposition des maisons entières ou des dépendances demeurées vacantes, pouvant accueillir bien plus d'un réfugié à la fois. Jeunes trentenaires ou parents se retrouvant seuls après le départ de leurs enfants (Figure 6.15), les accueillants sont, pour 50% d'entre eux (Figure 6.14), des cadres, des retraités ou des personnes exerçant une profession intellectuelle supérieure (des enseignants, des journalistes, des artistes ...).

Si, pour certaines familles, l'accueil de réfugiés s'inscrit dans le prolongement d'un engagement philosophique ou politique, voire fait écho à une histoire familiale passée (pour les descendants d'immigrés), il nous semble pourtant, à la lecture du champ *Motivations ?*, qu'une majorité d'entre elles pense répondre, d'abord et avant tout, à une situation d'urgence. Les reportages sur la situation en mer Égée, diffusés quotidiennement, poussent un public non militant à se tourner vers Singa, à s'engager auprès de l'association.

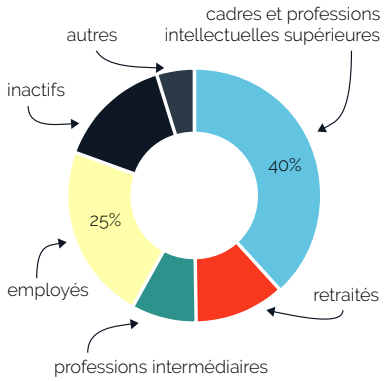


FIGURE 6.14: Répartition des accueillants par classes professionnelles

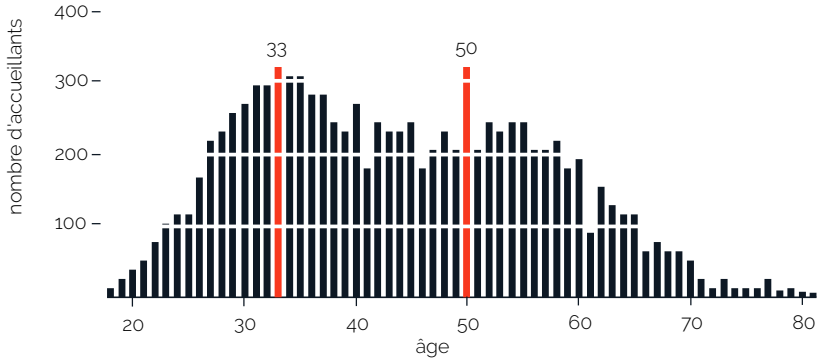


FIGURE 6.15: Distribution de l'âge des accueillants

Afin de vérifier cette impression, nous développons plusieurs scripts de tokenization et de stemming (Section 3.3), que nous appliquons à l'ensemble des déclarations de motivations des accueillants. Ces déclarations peuvent se cantonner à une ou deux lignes de texte seulement, mais également s'étirer sur plusieurs paragraphes argumentés. Nous construisons ainsi un **vocabulaire** représentatif des mots les plus fréquemment utilisés par les familles. Pour chacun de ces mots, nous traçons une courbe de fréquence trimestrielle et normalisée par le nombre de nouveaux inscrits sur la même période. Nous pouvons ainsi suivre l'évolution, au cours du temps, de la fréquence d'utilisation de

l'ensemble de ces termes (Figure 6.16).

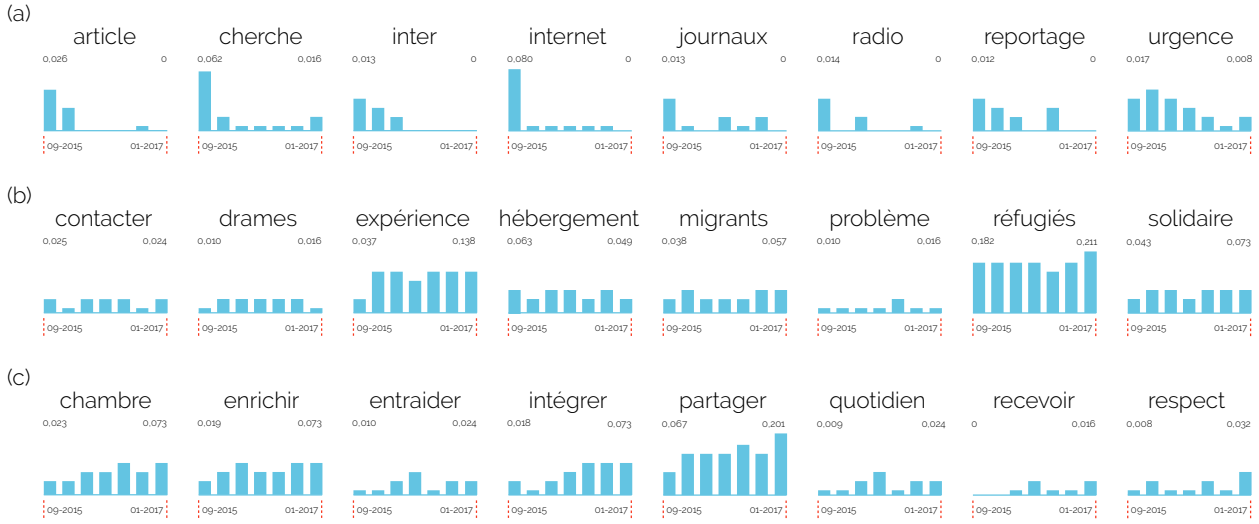


Figure 6.16: Évolution de la fréquence d'utilisation de certains termes du champ *Motivations ?*

Visuellement, ces mots peuvent être classés suivant quatre grandes tendances : **1)** les mots fréquemment utilisés à la fin de l'été 2015 et dont la proportion décroit par la suite (Figure 6.16, a), **2)** les mots utilisés de manière stable de juin 2015 à juin 2017 (Figure 6.16, b), **3)** les mots prenant de l'importance à mesure que la couverture médiatique diminue (Figure 6.16, c), **4)** les mots utilisés de manière non caractéristique. De fait, un vocabulaire propre à chaque tendance peut être dégagé :

Si la volonté de rester *solidaire* vis-à-vis du *drame* et de la détresse des personnes *réfugiées* en leur proposant un *hébergement* reste stable dans le temps, c'est en réponse au choc médiatique (*télévision, radio, reportages*. . .) que nombre d'accueillants viennent dans un premier temps s'inscrire au programme Calm, comme un réflexe face à l'*urgence* de la situation.

En revanche, plus nous nous éloignons chronologiquement de l'épicentre de la crise et plus l'accueil se conçoit comme une expérience réfléchie et mûrie de longue date : les accueillants disent vouloir *intégrer* les accueillis et *partager* avec eux, les *recevoir* dans le *confort* et dans le *respect*, *s'entraider* et *s'enrichir*. Le vocabulaire devient alors plus posé, rassurant (*chaleur, heureuse, belle, cœur*. . .) et se déploie comme autant de façons d'exprimer le soin que l'on souhaite apporter aux personnes réfugiées. Ainsi, la *chambre* que l'on partage en 2017 prend le relais du coin de *canapé* que l'on proposait en 2015.

Parmi l'ensemble des médias qui ont influencé les familles d'accueillants, France Inter [46] semble avoir joué un rôle très particulier. En effet, les 3 et 15 Septembre 2015, Singa bénéficie [47] de plusieurs passages sur

46. `https://www.franceinter.fr/`

47. Voir `https://www.franceinter.fr/emissions/grand-angle/grand-angle-03-septembre-2015` `https://www.franceinter.fr/emissions/l-esprit-d-initiative/l-esprit-d-initiative-15-septembre-2015` `https://www.franceinter.fr/emissions/un-jour-en-france/un-jour-en-france-15-septembre-2015`

les ondes de la radio nationale : un reportage, une interview et un débat. Dans le champ *Motivations ?*, les accueillants désignent alors France Inter comme vecteur de découverte principal de l'association. L'expression « *J'en ai entendu parler à France Inter* »[48] devient le dénominateur commun de plus de 500 inscrits via le formulaire en ligne.

48. Une expérience de classification par k-means nous permet de détecter un groupe de personnes utilisant le vocabulaire : *entendu, parler, france, inter, radio, association, médias...*

Aussi, en comparant la fréquence d'utilisation des mots *internet* et *radio* avec les dates des reportages de France Inter (voir la Figure 6.13, courbes du bas et tirets rouges), il est possible de voir une corrélation nette entre ces événements et le choix, momentané, du second médium au détriment du premier. L'écoute de France Inter devient alors, à nos yeux, un marqueur sociologique au moins aussi important que l'âge ou la profession des accueillants. Mais, aidé par une mise en scène sonore très immersive (captation dans les locaux de Singa en pleine agitation et rencontre avec une accueillante), le vocabulaire utilisé lors de ces passages radio a-t-il pu influencer les futurs accueillants ? Peut on retrouver la trace de ces mots dans les archives du formulaire ?

Pour ce faire, nous construisons enfin une interface d'exploration interactive[49]. Basée sur la découverte mot à mot de liens entre le texte de chaque intervention à France Inter (Figure 6.17, partie de gauche) et les déclarations de motivations des accueillants issues de la base de données (Figure 6.17, partie de droite), cette interface nous permet d'identifier l'esquisse d'un vocabulaire commun.

49. Téléchargeable ici `https://github.com/lobbeque/calm-miner/tree/master/viz`

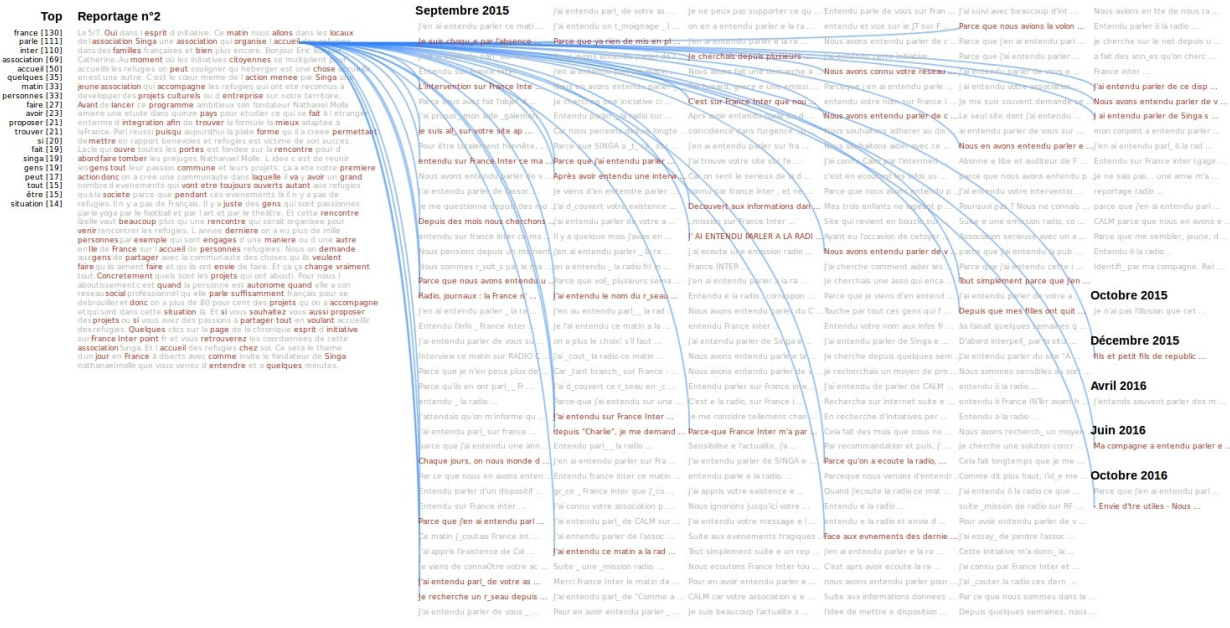


FIGURE 6.17: Utilisation d'un même vocabulaire (bleu) entre les reportages de France Inter (gauche) et les déclarations de motivation des accueillants (droite)

À ce niveau et passé cette première phase exploratoire, de futurs travaux devront permettre de quantifier, si possible, l'influence lexicale

de la radio sur ces déclarations d'intentions. Aussi et afin de compléter notre étude, il pourrait être intéressant d'élargir le panorama de recherche à d'autres types d'innovations sociales et solidaires. Chaque nouvelle structure apporte une manière différente de concevoir l'accueil et de penser la gestion des données accueillants/accueillis. Par exemple, les membres du *Jesuit Refugee Service*[50] (JRS) utilisent un système de prise de contact et d'inscription des familles d'accueillants par email. De même, en Belgique, les membres de la plateforme citoyenne *Bxlrefugeess*[51] s'organisent de manière autonome et décentralisée via des sondages Facebook[52]. Ils se rencontrent ensuite chaque soir à Bruxelles, au cœur du parc Maximilien, pour mettre en place un hébergement d'urgence et quotidien des réfugiés.

50. http://www.jrsfrance.org/

51. http://www.bxlrefugees.be/

52. https://www.facebook.com/plateformerefugiesbxl?fref=ts

***

Après avoir consacré la quasi-totalité de cette thèse à l'analyse d'archives Web *classiques*[53], ce dernier chapitre aura été l'occasion d'élargir nos horizons de recherche. Il existe, ainsi, une multitude de traces numériques à explorer pour interroger les diverses formes de représentations en ligne des populations migrantes.

53. Les collectages de sites Web par des crawlers étant, en volume, le type d'archive Web le plus représenté et le plus utilisé.

Ce faisant, les archives de navigation Web de la BPI se sont, d'abord, présentées à nous comme le moyen d'aborder, plus directement, les usages du Web de personnes en situation souvent précaire (Section 6.1). Migrants, sans domicile fixe ou séniors isolés se retrouvent à la bibliothèque du Centre Pompidou et s'installent jour après jour devant les postes Internet Libre. Ils viennent y chercher du repos, consulter des vidéos, voir simplement du monde ou effectuer des recherches documentaires approfondies.

Mais, si au final, nous avons cherché à reconstruire une chaine d'exploitation des logs assez similaire à ce que nous avions bâti pour les collectages de l'Atlas e-Diasporas (Section 3.3), le travail réalisé avec la BPI nous aura surtout permis de réaffirmer l'idée suivante : les liens entre enquêtes de terrain et explorations à grande échelle de données archivées doivent être maintenus ou encouragés.

Le Web de la BPI s'intègre dans l'écosystème global de la bibliothèque : le lieu, le personnel, les visiteurs, ce que l'on y voit, ce que l'on y entend, ce que les autres consultent... Aussi, nous ne pouvons nous satisfaire de la seule perspective que nous donnent à voir les logs de navigation Web. Les observations *in situ* préfigurent toute hypothèse de travail et nous permettent de cheminer vers la compréhension des usages du Web de la BPI, dans un dialogue constant entre explorations semi-automatiques et validations par le terrain. Mais ces traces se dégradent déjà et perdent de leur richesse, le Web se protégeant mécaniquement de nos méthodes de capture, dissimulé derrière des URL en https. L'archive Web, quelle que soit sa nature et son contenu,

n'est pas une matière immortelle ou immuable. Ses défauts et ses points de fragilité présents ou à venir doivent être pris en compte tout au long de nos analyses.

De son côté, l'étude des archives du formulaire d'inscription au programme CALM nous a permis d'évoquer, sous certains aspects, la *crise des réfugiés* de l'été 2015 (Section 6.2). En réponse à une couverture médiatique inédite, une vague soudaine de citoyens et d'anonymes se sont proposés d'accueillir les victimes de ce drame. Par l'étude automatisée de leurs déclarations de motivations et des mots qu'ils employaient pour se décrire, nous avons pu entrapercevoir les dynamiques de cet forme d'accueil, de personnes réfugiées chez des particuliers, entre 2015 et 2017. Ou comprendre comment un geste d'urgence s'est transformé progressivement en une démarche réfléchie.

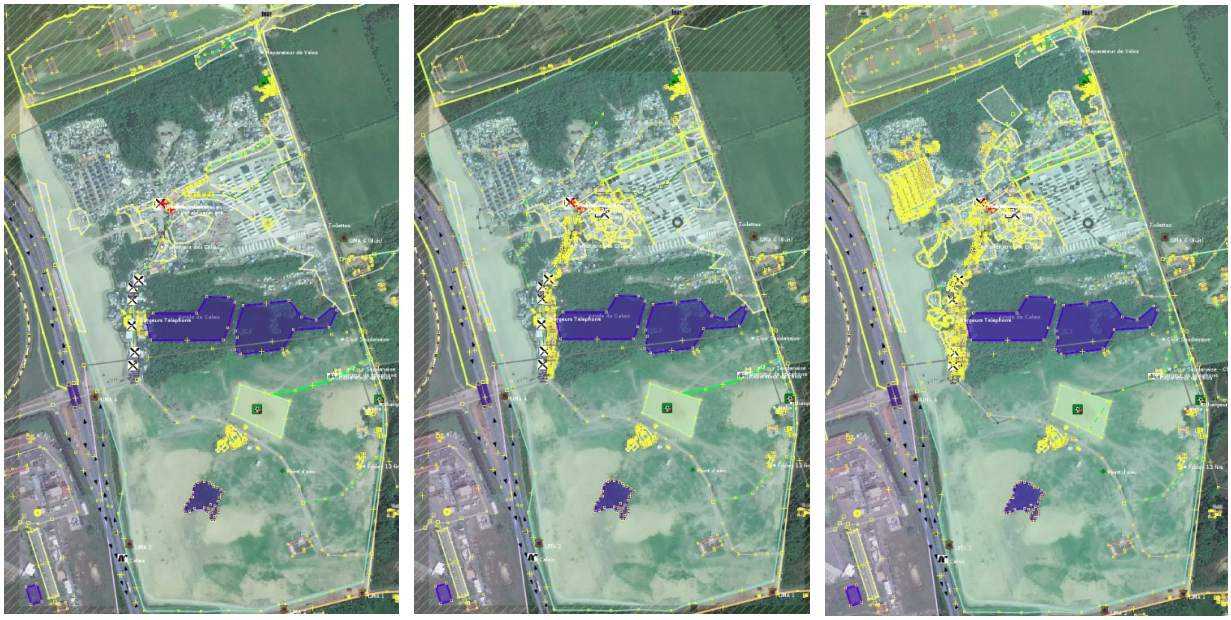
Le contenu d'une archive n'est cependant pas la seule et unique source d'information à interroger. La forme de cette dernière et la compréhension du contexte historique dans lequel elle a été définie, puis potentiellement modifiée, sont autant d'éléments susceptibles d'enrichir nos explorations. Le programme CALM a, ainsi, été confronté à la tentation de l'automatisation et du tout algorithme là où, justement, le savoir-faire manuel des équipes de l'association incarnait sa singularité, dans un moment de pure technophorie. Aussi, et de manière plus générale, les modalités de gestion, d'analyse et de traitement des données liées aux migrations doivent faire l'objet de recherches dédiées et approfondies, ne serait-ce que pour garantir la sécurité des personnes concernées.

Enfin, au cours de cette thèse, d'autres corpus d'archives auront été constitués au grès des rencontres et de l'évolution de nos recherches. Conversations Whatsapp, groupes Facebook de réfugiés syriens ou encore vidéos Youtube de traversées de la mer Égée sont autant de fragments extraits de la masse quotidienne de données générées par les migrations. Diverses et éparses, la valeur historique et sociologique de ces traces grandit à mesure que le temps passe et le devenir de certaines d'entre elles nous frappe encore aujourd'hui. Ainsi, lorsqu'à l'été 2016, nous accompagnions les équipes d'OpenstreetMap [54] sur une campagne de cartographie [55] des camps de Calais et de Grande-Synthe, il nous aurait été impossible d'imaginer que, seulement quelques mois plus tard, ces cartes en ligne deviendraient archives des *jungles* telles qu'elles avaient été (Figure 6.18).

54. Carte du monde en ligne, open source et collaborative `https://www.openstreetmap.org`.

55. Ce travail a fait l'objet d'une publication : Q. Lobbé, (2016), *Cartographier les jungles*, Plein Droit n°110, 2016

(a)



(b)

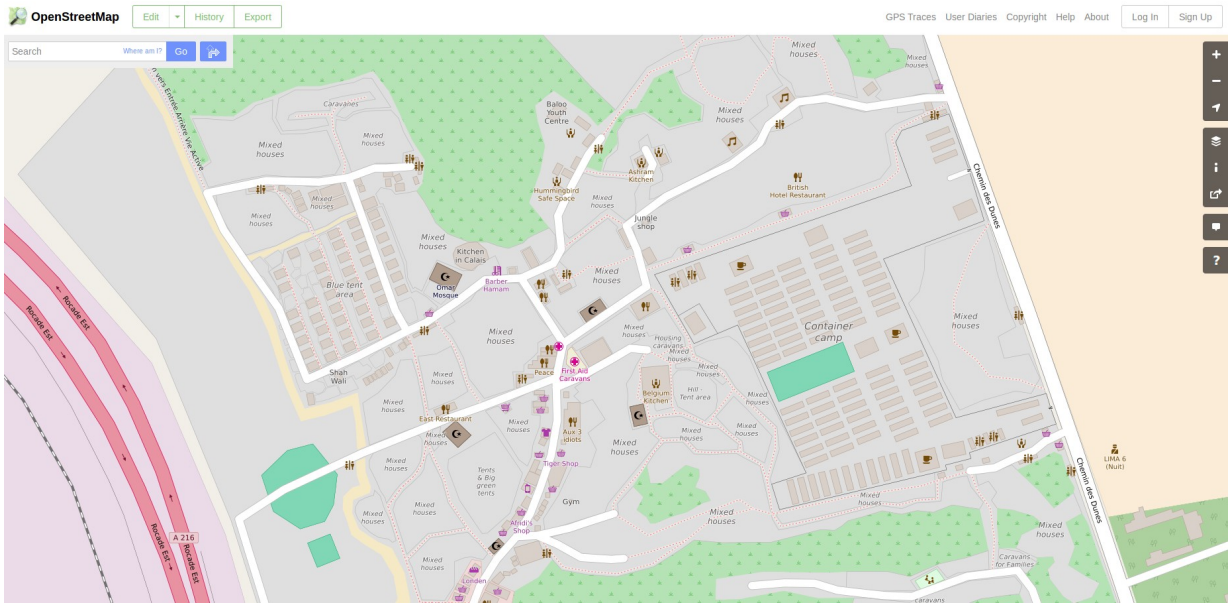


Figure 6.18: Évolution de la création de la carte de la Jungle de Calais en juillet 2016 (a) et carte finale (b)

Conclusion

Cette thèse a été menée avec la volonté assumée de revenir au sens social et historique premier des archives Web manipulées. Quitte à laisser en cours de route des pistes à peine défrichées, nous nous étions donné pour objectif de proposer l'illustration concrète d'une analyse du Web passé (Chapitre 5) afin d'entrevoir les possibilités offertes par le fragment Web (Chapitre 4). Ce faisant, nous souhaitons maintenant revenir sur quelques éléments restés à la marge.

Il serait intéressant, en premier lieu, de poursuivre la discussion engagée sur les artéfacts de crawl (Section 3.2 et 4.4). Nous pourrions profiter de la nature de nos corpus pour proposer un modèle plus ambitieux de gestion de la cohérence entre pages et fragments archivés. Passant d'un crawl unique à l'intégration de plusieurs campagnes de collecte, il s'agirait alors de penser la cohérence non plus du point de vue du crawler, mais en considérant cette fois le regard singulier d'un observateur.

Cette réflexion déboucherait sur la mise en place d'une palette d'outils destinés à juger de la qualité générale ou relative des corpus étudiés. Des indicateurs dynamiques de cohérence (structurelle, temporelle, thématique...) permettraient au chercheur de s'orienter tout au long de son exploration. Il pourrait aussi s'agir de généraliser l'usage des échelles de datation (Table 4.1) en calculant, à la volée, l'écart entre les diverses dates d'une page Web consultée et la période ciblée par une requête donnée.

La méthode d'extraction des fragments Web (Section 4.3) gagnerait, pour sa part, à être comparée à d'autres approches. Même si nous soutenons l'idée qu'une page archivée soit segmentée relativement à chaque question de recherche, cela ne doit pas nous empêcher de faire l'inventaire des défauts et des qualités de notre stratégie. D'autres chercheurs pourraient alors s'en saisir et proposer des implémentations concurrentes ou complémentaires.

Malgré nos efforts, la capacité du fragment Web à capturer les dispositifs d'écriture et de partage d'un contenu sur le Web (Section 4.2, point n°5) n'a pas été exploitée. Cette dimension aurait pourtant pu

enrichir notre analyse de la migration des blogs marocains vers les plateformes de réseaux sociaux (Section 5.2), et ce, en questionnant les gains cognitifs et sémiologiques d'une telle mutation. C'est peut-être là que se niche la clé de compréhension du passage d'un territoire du Web à un autre.

Enfin, fidèles à notre approche exploratoire (Section 5.1), il nous semble évident que l'étude des conversations postées sur *yabiladi.com* pourrait être poussée plus avant. Par exemple, nous n'avons pas observé dans le détail les prises de position des pro- et des anti- mouvement du 20 février. Une analyse fine du vocabulaire employé pourrait nous aider à entrevoir les conflits internes au forum. De nouvelles catégories apparaitraient alors, nous offrant une lecture approfondie de ces événements.

À la manière d'un inventaire, le Chapitre 6 nous invitait à porter notre réflexion au-delà des archives Web et des formes classiques de collecte. Constitués à différentes époques et sur la base de technologies ou de standards parfois datés, les corpus archivés se laissent aujourd'hui découvrir dans toute leur diversité. Ainsi, chaque plongée dans la mémoire du Web passé nous pousse, tout autant, à repenser notre rapport au Web tel qu'il a été qu'à questionner, au présent, les mutations du Web vivant.

Assez paradoxalement et malgré la multiplication des moyens de connexion, force est de constater que le Web d'aujourd'hui s'appauvrit. En effet, la toile subit les assauts de stratégies industrielles visant à lui imposer une forme de gouvernance marchande. Les GAFA[56] se nourrissent de la singularité de nos passages sur le Web en réduisant les traces que nous y laissons à un ensemble fini de particularités calculables. Penser le Web, d'abord et avant tout, pour y faire prospérer des algorithmes de recommandation ou de profilage finira inéluctablement par impacter le devenir de l'archivage. L'altérité que l'on pouvait éprouver face à la multitude des ressources et des personnes évoluant sur le Web, disparait au profit d'une normalisation basée sur l'historique de nos propres navigations. De fait, avec la personnalisation généralisée de nos manières d'être sur la toile (Section 1.1), c'est toute l'architecture commune du Web qui tend à s'effacer.

56. Acronyme de Google, Apple, Facebook et Amazon désignant de manière plus générale l'ensemble des géants du Web.

Et si l'avenir du Web vivant appelle une reprise en main collective de son développement, pour nous, explorateurs d'archives, l'enjeu est de veiller à ce que les corpus du Web passé restent des communs de l'humanité. Nous avons la chance d'être contemporains de la matière sur laquelle nous menons nos recherches. Aussi, prenons garde à ce que les archives ne deviennent pas des denrées à monétiser et favorisons plutôt un modèle où elles contribueraient à produire des savoirs.

Ce faisant, l'une des dimensions les plus importantes du fragment Web

est peut-être de vouloir recentrer l'exploration sur le chercheur et sur ses choix, en préférant au consommateur passif la figure du praticien actif.

Le Web passé s'éprouve dans la durée et dans une forme de rapport qualitatif à ce qu'il renferme. Les savoir-faire que nous développons incorporent alors l'expérience de situations vécues au contact direct des archives. De fait, nous recommandons des méthodes d'analyses mixtes, alliant traitements automatisés à grande échelle et validations ad hoc et manuelles. Ces stratégies doivent pouvoir s'adapter à la singularité des corpus et à la variété des époques représentées. Pour rendre compte, au présent, du Web tel qu'il a été, il faut ainsi parfois savoir se rendre anachronique à son propre temps.

Mais il ne s'agit pas, pour autant, de faire l'économie d'outils déjà existants. La Wayback Machine, au même titre que les systèmes développés par l'INA, restent indispensables à toute analyse. Aussi, lors d'une exploration, rien ne doit être jeté ou laissé de côté. Ne sachant pas l'étendue de ce qui a été collecté et de ce qui n'a pas été préservé, tout élément, même anodin, pourrait être le point de départ des recherches de demain[57]. En creux, l'exploration des archives Web nous invite à la traque de ce qui, justement, est absent des corpus. Il nous faut repérer les légers décalages entre les collectages et identifier les fantômes qui hantent les pages.

57. Voir l'articulation entre les explorations des Sections 5.2 et 5.3

Mosaïque dispersée, il est illusoire de vouloir préserver la mémoire du Web au sein d'un seul et unique corpus. Aussi, pourquoi ne pas chercher à fédérer les sources éparses de collecte pour y établir des ponts ou des jointures ? C'est en conjuguant les vastes ressources d'Internet Archive avec le plus petit des fichiers HTML égaré que nous pourrons tenter de reconstituer une image fidèle du Web passé.

Cependant, la nature et la valeur que l'on accorde à nos traces évoluent avec le temps. Ce qui, autrefois, a été déposé sur le Web est peut-être, aujourd'hui, devenu trop sensible. Certaines données doivent ainsi être anonymisées, protégées ou aggrégées suivant les cas. Mais gardons-nous bien de détruire ou d'endommager les archives. Travaillons toujours sur des copies et laissons des zones inexplorées pour les générations à venir, car la plus grande tâche n'a pas encore commencé.

En effet, il nous faut maintenant poursuivre les recherches entamées sur l'Atlas e-Diaspora en les généralisant à l'ensemble des corpus migrants. Cela sera une manière de valider nos outils en mobilisant, à nouveau, l'ensemble des contributeurs de l'Atlas. Confrontés aux évolutions des sites qu'ils avaient cartographiés, ces chercheurs seront alors à même d'analyser les dynamiques révélées.

Arrivés au terme de cette thèse, nous espérons, finalement, avoir su

créer des possibles et ouvrir quelques brèches dans les archives du Web passé. Encore balbutiantes, nos explorations finiront peut-être par s'effacer au profit d'une discipline nouvelle, plus mature et qu'il nous faudra entièrement construire. Viendra alors le temps d'établir, enfin, les bases d'une véritable archéologie du Web.

| Bibliographie

Abiteboul, S., Cobena, G., Masanes, J. et Sedrati, G. (2002). A first experience in archiving the French Web. *In International Conference on Theory and Practice of Digital Libraries*, pages 1–15. Springer.

Adamic, L. A. et Glance, N. (2005). The political blogosphere and the 2004 US election : divided they blog. *In Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM.

Adar, E., Teevan, J., Dumais, S. T. et Elsas, J. L. (2009). The web changes everything : understanding the dynamics of web content. *In Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 282–291. ACM.

Adar, E., Zhang, L., Adamic, L. A. et Lukose, R. M. (2004). Implicit structure and the dynamics of blogspace. *In Workshop on the weblogging ecosystem*, volume 13, pages 16989–16995.

AlNoamany, Y. A., Weigle, M. C. et Nelson, M. L. (2013). Access patterns for robots and humans in web archives. *In Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pages 339–348. ACM.

Amar, M. (2018). Analyse des logs de consultation d'Internet en accès libre à la Bpi : qu'apporte le Big Data ? Rapport technique, Bibliothèque Publique d'Information.

Amar, M. et Béguet, B. (2008). Les consultations " libres " d'Internet à la Bpi : enquête exploratoire. Rapport de recherche.

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

Amoore, L. (2006). Biometric borders : Governing mobilities in the war on terror. *Political geography*, 25(3):336–351.

Anand, A., Bedathur, S., Berberich, K., Schenkel, R. et Tryfonopoulos, C. (2009). EverLast : a distributed architecture for preserving the

web. *In Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pages 331–340. ACM.

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

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

Augustin, S. (421). *Confessions*.

Bachimont, B. (2009). Archivage audiovisuel et numérique : les enjeux de la longue durée. *Archivage et stockage pérennes. Paris : Hermès Science Publications : Lavoisier*, pages 195–222.

Bachimont, B., Drugeon, T. et Piéjut, G. (2005). Documenter et partitionner une archive du Web : vers le dépot légal d'un domaine media. *Archives & Museum Informatics*, page 2.

Baeza-Yates, R., Castillo, C., Marin, M. et Rodriguez, A. (2005). Crawling a country : better strategies than breadth-first for web page ordering. *In Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 864–872. ACM.

Barabási, A.-L., Albert, R. et Jeong, H. (2000). Scale-free characteristics of random networks : the topology of the world-wide web. *Physica A : statistical mechanics and its applications*, 281(1-4):69–77.

Barats, C. (2016). *Manuel d'analyse du web-2e éd. : En Sciences Humaines et Sociales*. Armand Colin.

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

Beaudouin, V., Garron, I. et Rollet, N. (2016). Je pars d'un sujet, je rebondis sur un autre. Rapport de recherche, Télécom ParisTech ; Bibliothèque nationale de France ; Labex Obvil.

Ben-David, A. (2012). The Palestinian diaspora on the Web : Between de-territorialization and re-territorialization. *Social Science Information*, 51(4):459–474.

Ben-David, A. et Amram, A. (2018). The Internet Archive and the socio-technical construction of historical facts. *Internet Histories*, pages 1–23.

Ben-David, A., Amram, A. et Bekkerman, R. (2018). The colors of the national Web : visual data analysis of the historical Yugoslav Web domain. *International Journal on Digital Libraries*, 19(1):95–106.

Bennafla, K. et Seniguer, H. (2011). Le Maroc à l'épreuve du printemps arabe : une contestation désamorcée ? *Outre-terre*, (3):143–158.

Bennani-Chraïbi, M. et Jeghllaly, M. (2012). La dynamique protestataire du Mouvement du 20 février à Casablanca. *Revue française de science politique*, 62(5):867–894.

Bergman, M. K. (2001). White paper : the deep web : surfacing hidden value. *Journal of electronic publishing*, 7(1).

Bernal, V. (2006). Diaspora, cyberspace and political imagination : the Eritrean diaspora online. *Global networks*, 6(2):161–179.

Bernard, M. (2003). Criteria for optimal web design (designing for usability). *Retrieved on April*, 13:2005.

Berthomière, W. (2013). A French what ? À la recherche d'une diaspora française. Premiers éléments d'enquête au sein de l'espace internet.

Bertin, J., Barbut, M. et Bonin, S. (1967). Sémiologie graphique : les diagrammes, les réseaux, les cartes. *Paris/La Haye, Éd. Gauthier-Villars/Mouton*.

Blanchard, E. et Rodier, C. (2016). «Crise migratoire» : ce que cachent les mots. *Plein droit*, (4):3–6.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. et Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics : theory and experiment*, 2008(10):P10008.

Bon, F. (2014). *Après le livre*. Tiers Livre Éditeur.

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

Bounie, D., Diminescu, D. et François, A. (2010). Une analyse socio-économique des transferts d'argent des migrants par téléphone. *Réseaux*, (1):91–109.

Brin, S. et Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. et Wiener, J. (2000). Graph structure in the web. *Computer networks*, 33(1):309–320.

Bruslé, T. (2012). Les sites diasporiques népalais, signes et conditions d'une diaspora en formation ?

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

Brügger, N. et Schroeder, R. (2017). *The Web as History : Using Web Archives to Understand the Past and the Present*. UCL Press.

Cadez, I., Heckerman, D., Meek, C., Smyth, P. et White, S. (2003). Model-based clustering and visualization of navigation patterns on a web site. *Data mining and knowledge discovery*, 7(4):399–424.

Cai, D., Yu, S., Wen, J.-R. et Ma, W.-Y. (2003). Extracting content structure for web pages based on visual representation. *In Asia-Pacific Web Conference*, pages 406–417. Springer.

Canfora, L. (1990). *The Vanished Library : A Wonder of the Ancient World*, volume 7. Univ of California Press.

Cardon, D. (2015). *A quoi rêvent les algorithmes. Nos vies à l'heure : Nos vies à l'heure des big data*. Le Seuil.

Castillo, C., Marin, M., Rodriguez, A. et Baeza-Yates, R. (2004). Scheduling algorithms for Web crawling. *In WebMedia and LA-Web, 2004. Proceedings*, pages 10–17. IEEE.

CERN (1993). The document that officially put the World Wide Web into the public domain.

Chakrabarti, S. (2002). *Mining the Web : Discovering knowledge from hypertext data*. Elsevier.

Chaney, A. J., Wallach, H. et Blei, D. M. (2015). Who, What, When, Where, and Why ? A Computational Approach to Understanding Historical Events Using State Department Cables.

Chaney, A. J.-B., Wallach, H. M., Connelly, M. et Blei, D. M. (2016). Detecting and Characterizing Events. *In EMNLP*, pages 1142–1152.

Chawathe, S. S. et Garcia-Molina, H. (1997). Meaningful change detection in structured data. *In ACM SIGMOD Record*, volume 26, pages 26–37. ACM.

Chen, J. (2011). Cerner la notion de temps. *Rue Descartes*, (2):30–51.

Chen, W. (2010). Internet-usage patterns of immigrants in the process of intercultural adaptation. *Cyberpsychology, Behavior, and Social Networking*, 13(4):387–399.

Cho, J. et Garcia-Molina, H. (1999). The evolution of the web and implications for an incremental crawler. Rapport technique, Stanford.

Cho, J., Garcia-Molina, H. et Page, L. (1998). Efficient crawling through URL ordering. *Computer Networks and ISDN Systems*, 30(1-7):161–172.

Coriat, B. (2015). *Le retour des communs : & la crise de l'idéologie propriétaire*. Éditions Les Liens qui libèrent.

Costa, M., Gomes, D., Couto, F. et Silva, M. (2013). A survey of web archive search architectures. *In Proceedings of the 22nd International Conference on World Wide Web*, pages 1045–1050. ACM.

Costa, M. et Silva, M. J. (2011). Characterizing Search Behavior in Web Archives. *In TWAW*, pages 33–40.

Costa, M. et Silva, M. J. (2012). Evaluating web archive search systems. *In International Conference on Web Information Systems Engineering*, pages 440–454. Springer.

Damian, E. et Van Ingen, E. (2014). Social network site usage and personal relations of migrants. *Societies*, 4(4):640–653.

De Jong, F., Rode, H. et 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.

De Kosnik, A. (2016). *Rogue archives : Digital cultural memory and media fandom*. MIT Press.

Dekker, R., Engbersen, G., Klaver, J. et Vonk, H. (2018). Smart Refugees : How Syrian Asylum Migrants Use Social Media Information in Migration Decision-Making. *Social Media+ Society*, 4(1).

Denev, D., Mazeika, A., Spaniol, M. et Weikum, G. (2009). SHARC : framework for quality-conscious web archiving. *Proceedings of the VLDB Endowment*, 2(1):586–597.

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

Desrues, T. (2012). Le mouvement du 20 février et le régime marocain : contestation, révision constitutionnelle et élections. *L'Année du Maghreb*, (VIII):359–389.

Diminescu, D. (2002). Les migrations à l'âge des nouvelles technologies. *Hommes & Migrations*, 1240(1240):6–9.

Diminescu, D. (2008). The connected migrant : an epistemological manifesto. *Social Science Information*, 47(4):565–579.

Diminescu, D. (2012a). *E-Diasporas Atlas : Exploration and Cartography of Diasporas on Digital Networks*. Ed. de la Maison des sciences de l'homme.

Diminescu, D. (2012b). *Introduction : Digital methods for the exploration, analysis and mapping of e-diasporas*. Sage Publications Sage UK : London, England.

Diminescu, D. (2016). Traces numériques. *Plein droit*, (3):3–6.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A. et Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network : Hedonometrics and Twitter. *PloS one*, 6(12).

Douglis, F., Ball, T., Chen, Y.-F. et Koutsofios, E. (1998). The AT&T Internet Difference Engine : Tracking and viewing changes on the web. *World Wide Web*, 1(1):27–44.

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

Driscoll, K. et Paloque-Berges, C. (2017). Searching for missing "net histories". *Internet Histories*, 1(1-2):47–59.

Drugeon, T. (2005). A technical approach for the French web legal deposit. *In 5th International Web Archiving Workshop (IWAW05), Viena, Austria*. Citeseer.

Dufoix, S. (2003). *Les diasporas*. Que sais-je ? Presses Universitaires de France, Paris.

Dumais, S., Cutrell, E., Cadiz, J. J., Jancke, G., Sarin, R. et Robbins, D. C. (2016). Stuff I've seen : a system for personal information retrieval and re-use. *In ACM SIGIR Forum*, volume 49, pages 28–35. ACM.

Ellison, N. B. (2007). Social network sites : Definition, history, and scholarship. *Journal of computer-mediated Communication*, 13(1):210–230.

Febvre, L. et Martin, H.-J. (2013). *L'apparition du livre*. Albin Michel.

Feng, G., Ma, G.-S. et Hu, J. (2006). Web navigation patterns mining based on clustering of paths and pages content. *In Asia-Pacific Web Conference*, pages 857–860. Springer.

Fetterly, D., Manasse, M., Najork, M. et Wiener, J. (2003). A large-scale study of the evolution of web pages. *In Proceedings of the 12th international conference on World Wide Web*, pages 669–678. ACM.

Fitch, K. (2003). Web site archiving-an approach to recording every materially different response produced by a website.

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

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

Fouetillou, G. (2008). Le web et le traité constitutionnel européen. *Réseaux*, (1):229–257.

Fry, B. J. (2000). *Organic information design*. Mémoire de master, Massachusetts Institute of Technology.

Fry, B. J. (2004). *Computational information design*. Thèse, Massachusetts Institute of Technology.

Fung, G. P. C., Yu, J. X., Yu, P. S. et Lu, H. (2005). Parameter free bursty events detection in text streams. *In Proceedings of the 31st international conference on Very large data bases*, pages 181–192. VLDB Endowment.

Gaumont, N., Panahi, M. et Chavalarias, D. (2017). Methods for the reconstruction of the socio-semantic dynamics of political activist Twitter networks.

Gebeil, S. (2016a). Les mémoires de l'immigration maghrébine sur le web français de 1999 à 2014. *Les Cahiers du numérique*, 12(3):115–138.

Gebeil, S. (2016b). Quand l'historien rencontre les archives du Web. *Revue de la BNF*, (2):185–191.

Giesler, M. (2006). Consumer gift systems. *Journal of consumer research*, 33(2):283–290.

Gillani, N., Yuan, A., Saveski, M., Vosoughi, S. et Roy, D. (2018). Me, My Echo Chamber, and I : Introspection on Social Media Polarization. *arXiv preprint arXiv :1803.01731*.

Ginzburg, C. (1980). Signes, traces, pistes. *Le débat*, (6):3–44.

Ginzburg, C. (2012). *Mythes, emblèmes, traces : morphologie et histoire*. Verdier/Poche.

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

Gomes, D., Nogueira, A., Miranda, J. et Costa, M. (2009). Introducing the Portuguese web archive initiative. *In 8th International Web Archiving Workshop*. Springer.

Goody, J., Bazin, J. et Bensa, A. (1979). *La raison graphique : la domestication de la pensée sauvage*. Collection le sens commun. Editions de Minuit.

Gossart, C., Jullien, N., Massé, D. et Özman, M. (2018). Panorama des innovations sociales numériques. *Terminal. Technologie de l'information, culture & société*, (122).

Grainger, T., Potter, T. et Seeley, Y. (2014). *Solr in action*. Manning Cherry Hill.

Hafner, K. et Lyon, M. (1998). *Where wizards stay up late : The origins of the Internet*. Simon and Schuster.

Hallgrinsson, B. et 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.

Heller, C. et Pezzani, L. (2014). Traces liquides : enquête sur la mort de migrants dans la zone-frontière maritime de l'Union européenne. *Revue européenne des migrations internationales*, 30(3):71–107.

Helmond, A. (2015). The platformization of the web : Making web data platform ready. *Social Media+ Society*, 1(2).

Helmond, A. (2017). Historical Website Ecology. Analyzing Past States of the Web Using Archived Source Code. *Web 25 : Histories from the First 25 Years of the World Wide Web*.

Holten, D. (2006). Hierarchical edge bundles : Visualization of adjacency relations in hierarchical data. *IEEE Transactions on visualization and computer graphics*, 12(5):741–748.

Holzmann, H. et Anand, A. (2016). Tempas : Temporal Archive Search Based on Tags. *In Proceedings of the 25th International Conference Companion on World Wide Web*, pages 207–210. International World Wide Web Conferences Steering Committee.

Hölscher, C. et Strube, G. (2000). Web search behavior of Internet experts and newbies. *Computer networks*, 33(1-6):337–346.

Ingold, T. et Renaut, S. (2013). *Une brève histoire des lignes*. Zones sensibles.

Jacomy, M., Venturini, T., Heymann, S. et Bastian, M. (2014). ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one*, 9(6).

Jatowt, A., Kawai, Y. et 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. et Nørvaag, 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.

Khondker, H. H. (2011). Role of the new media in the Arab Spring. *Globalizations*, 8(5):675–679.

Khoury, I., El-Mawas, R. M., El-Rawas, O., Mounayar, E. F. et Artail, H. (2007). An Efficient Web Page Change Detection System Based on an Optimized Hungarian Algorithm. *IEEE Transactions on Knowledge and Data Engineering*, 19(5):599–613.

Khouzaimi, J. (2015). e-Diasporas : Réalisation et Interprétation du corpus marocain.

Kimpton, M. et Ubois, J. (2006). Year-by-year : from an archive of the Internet to an archive on the Internet. *In Web archiving*, pages 201–212. Springer.

Koehler, W. (1999). An analysis of web page and web site constancy and permanence. *Journal of the Association for Information Science and Technology*, 50(2):162.

Koehler, W. (2004). A longitudinal study of Web pages continued : a consideration of document persistence. *Information Research*, 9(2):9–2.

Kohlschütter, C., Fankhauser, P. et Nejdl, W. (2010). Boilerplate detection using shallow text features. *In Proceedings of the third ACM international conference on Web search and data mining*, pages 441–450. ACM.

Koukoutsaki-Monnier, A. (2012). Deterritorialising the nation ? Internet and the politics of the Greek-American diaspora. *Nations and Nationalism*, 18(4):663–683.

Kumar, P. (2018). Rerouting the narrative : Mapping the online identity politics of the Tamil and Palestinian diaspora. *Social Media+ Society*, 4(1).

Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A. et Upfal, E. (2000). Stochastic models for the web graph. *In Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 57–65. IEEE.

Kwak, H., Lee, C., Park, H. et Moon, S. (2010). What is Twitter, a social network or a news media ? *In Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM.

Ladiray, D. (1997). L'AED, analyse exploratoire des données. *Courrier des statistique, 83*, 84:3–6.

Laflaquière, J., Gangloff, S., Scopsi, C., Guignard, T., Soultanova, R., Salzbrunn, M., Beaujouan, V. et Diminescu, D. (2005). Archiver le Web sur les migrations : quelles approches techniques et scientifiques ? Rapport d'étape. *Migrance*, (23):72–93.

Lawrence, S. et Giles, C. L. (1998). Searching the world wide web. *Science*, 280(5360):98–100.

Lawrence, S. et Giles, C. L. (2000). Accessibility of information on the web. *intelligence*, 11(1):32–39.

Lazard, E. et Mounier-Kuhn, P.-E. (2016). *Histoire illustrée de l'informatique : Histoire illustrée de l'informatique*. EDP sciences.

Leclerc, E. (2012). Le cyberespace de la diaspora indienne.

Legrady, G. (2005). Making visible the invisible. *Seattle Library Data Flow Visualization. In : Digital Culture and Heritage. Proceedings of ICHIM05 Sept*, pages 21–23.

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.

Leurs, K. et Ponzanesi, S. (2018). Connected migrants : Encapsulation and cosmopolitanization. *Popular Communication*, 16(1):4–20.

Licklider, J. C. et Taylor, R. W. (1968). The computer as a communication device. *Science and technology*, 76(2):1–3.

LIM, S.-J. et NG, Y.-K. (2001). An automated change-detection algorithm for HTML documents based on semantic hierarchies. *In Data Engineering, 2001. Proceedings. 17th International Conference on*, pages 303–312. IEEE.

LIU, L., PU, C. et TANG, W. (2000). WebCQ-detecting and delivering information changes on the web. *In Proceedings of the ninth international conference on Information and knowledge management*, pages 512–519. ACM.

LOBBÉ, J. (2018). *Concevoir des produits pour tous et par tous, co-créer la situation de vie*. Thèse, Université de Technologie de Belfort Montbéliard.

LOTAN, G., GRAEFF, E., ANANNY, M., GAFFNEY, D., PEARCE, I. et OTHERS (2011). The Arab Spring| the revolutions were tweeted : Information flows during the 2011 Tunisian and Egyptian revolutions. *International journal of communication*, 5:31.

LOVELUCK, B. (2015). *Réseaux, libertés et contrôle : Une généalogie politique d'internet*. Armand Colin.

MARCHANDISE, S. (2014). Le Facebook des étudiants marocains. Territoire relationnel et territoire des possibles. *Revue européenne des migrations internationales*, 30(3-4).

MARZ, N. et WARREN, J. (2015). *Big Data : Principles and best practices of scalable realtime data systems*. Manning Publications Co.

MASANÈS, J. (2006a). Archiving the hidden web. *In Web Archiving*, pages 115–129. Springer.

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

MCCANDLESS, M., HATCHER, E. et GOSPODNETIC, O. (2010). *Lucene in action*. Manning Publications Co.

MCDONNELL, J. P., KOEHLER JR, W. C. et CARROLL, B. C. (1999). Cataloging Challenges in an Area Studies Virtual Library Catalog (ASVLC) Results of a Case Study. *Journal of Internet Cataloging*, 2(2):15–42.

MICHAILIDOU, E., HARPER, S. et BECHHOFER, S. (2008). Visual Complexity and Aesthetic Perception of Web Pages. *In Proceedings of the 26th Annual ACM International Conference on Design of Communication*, SIGDOC '08, pages 215–224, New York, NY, USA. ACM.

MICHEL, J.-B., SHEN, Y. K., AIDEN, A. P., VERES, A., GRAY, M. K., PICKETT, J. P., HOIBERG, D., CLANCY, D., NORVIG, P., ORWANT, J. et OTHERS

(2011). Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Mitchell, R. (2015). *Web scraping with Python : collecting data from the modern web*. " O'Reilly Media, Inc.".

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

Morozov, E. (2013). *To save everything, click here : Technology, solutionism, and the urge to fix problems that don't exist*. Penguin UK.

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

Mussou, C. (2012). Et le Web devint archive : enjeux et défis. *Le Temps des médias*, (2):259–266.

Nedelcu, M. (2003). E-communautarisme ou l'impact de l'internet sur le quotidien des migrants. Les nouvelles migrations des professionnels roumains au Canada. *Visibles mais peu nombreux. Les circulations migratoires roumaines*, pages 325–339.

Nelson, T. H. (1967). Getting it out of our system. *Information retrieval : A critical review*, pages 191–210.

Nunes, S., Ribeiro, C. et 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.

Oita, M. et Senellart, P. (2010). Archiving data objects using Web feeds. *In International Workshop on Web Archiving*.

Oita, M. et Senellart, P. (2015). FOREST : Focused object retrieval by exploiting significant tag paths. *In Proceedings of the 18th International Workshop on Web and Databases*, pages 55–61. ACM.

Page, L., Brin, S., Motwani, R. et Winograd, T. (1999). The PageRank Citation Ranking : Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab.

Paloque-Bergès, C. (2018). *Qu'est-ce qu'un forum internet ? : Une généalogie historique au prisme des cultures savantes numériques*. OpenEdition Press.

Pandey, S. et Olston, C. (2005). User-centric web crawling. *In Proceedings of the 14th international conference on World Wide Web*, pages 401–411. ACM.

PANT, G., SRINIVASAN, P. et MENCZER, F. (2004). Crawling the web. *In Web Dynamics*, pages 153–177. Springer.

PARK, H. W. et THELWALL, M. (2008). Developing network indicators for ideological landscapes from the political blogosphere in South Korea. *Journal of computer-mediated communication*, 13(4):856–879.

PASQUIER, D. (2018). Classes populaires en ligne : des «oubliés» de la recherche ? *Réseaux*, (2):9–23.

PAUGAM, S. et GIORGETTI, C. (2013). Des pauvres à la bibliothèque. *Enquête au Centre Pompidou. Paris*.

PILLANT, L. (2016). En Grèce, une crise migratoire chronique. *Plein droit*, (4):31–34.

POP, R., VASILE, G. et MASANES, J. (2010). Archiving web video. *In International Web Archiving Workshop IWAW 2010*.

RISAM, R. (2018). Now you see them : Self-representation and the refugee selfie. *Popular Communication*, 16(1):58–71.

RISSE, T., DEMIDOVA, E., DIETZE, S., PETERS, W., PAPAILIOU, N., DOKA, K., STAVRAKAS, Y., PLACHOURAS, V., SENELLART, P., CARPENTIER, F. et OTHERS (2014). The ARCOMEM architecture for social-and semantic-driven web archiving. *future internet*, 6(4):688–716.

ROBERTSON, R. E., LAZER, D. et WILSON, C. (2018). Auditing the Personalization and Composition of Politically-Related Search Engine Results Pages. *In Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 955–965. International World Wide Web Conferences Steering Committee.

ROCCO, D., BUTTLER, D. et LIU, L. (2003). Page digest for large-scale web services. *In E-Commerce, 2003. CEC 2003. IEEE International Conference on*, pages 381–390. IEEE.

ROGERS, R. (2009). *The end of the virtual : Digital methods*, volume 339. Amsterdam University Press.

ROGERS, R., WELTEVREDE, E., BORRA, E. et NIEDERER, S. (2013). National Web Studies. *A companion to new media dynamics*, pages 142–166.

ROUVROY, A. et BERNS, T. (2013). Gouvernementalité algorithmique et perspectives d'émancipation. *Réseaux*, (1):163–196.

RUSSELL, A. L. et SCHAFER, V. (2014). In the Shadow of ARPANET and Internet : Louis Pouzin and the Cyclades Network in the 1970s. *Technology and Culture*, 55(4):880–907.

Saad, M. B., Pehlivan, Z. et Gançarski, S. (2011). Coherence-oriented crawling and navigation using patterns for web archives. *In International Conference on Theory and Practice of Digital Libraries*, pages 421–433. Springer.

Sabancioglu, M. (2011). New Custom for the Old Village Interpreting History through Turkish Village Web-Sites.

Salmon, J.-M. (2016). *29 jours de révolution. Histoire du soulèvement tunisien, 17 décembre 2010 - 14 janvier 2011*. Les Petits matins.

Sayad, A. (1985). Du message oral au message sur cassette, la communication avec l'absent. *Actes de la recherche en sciences sociales*, 59(1):61–72.

Sayad, A. (2000). *La double absence : des illusions de l'émigré aux souffrances de l'immigré*. Seuil.

Schafer, V. (2015). Part of a whole : RENATER, a twenty-year-old network within the Internet. *Information & Culture*, 50(2):217–235.

Schafer, V. et Thierry, B. G. (2016). The "Web of pros" in the 1990s : The professional acclimation of the World Wide Web in France. *New Media & Society*, 18(7):1143–1158.

Schneider, S. M., Foot, K., Kimpton, M. et Jones, G. (2003). Building thematic web collections : challenges and experiences from the September 11 Web Archive and the Election 2002 Web Archive. *Digital Libraries, ECDL*, pages 77–94.

Scopsi, C. (2009). Les sites web diasporiques : un nouveau genre médiatique ? *tic&société [En ligne]*, 3(1-2).

Scott, J. (2017). *Social network analysis*. Sage.

Senellart, P. (2007). *Comprendre le Web caché. Understanding the hidden Web*. Thèse.

Spaniol, M., Denev, D., Mazeika, A., Weikum, G. et Senellart, P. (2009). Data quality in web archiving. *In Proceedings of the 3rd workshop on Information credibility on the web*, pages 19–26. ACM.

Spaniol, M. et Weikum, G. (2012). Tracking entities in web archives : the LAWA project. *In Proceedings of the 21st International Conference on World Wide Web*, pages 287–290. ACM.

Spinellis, D. (2003). The decay and failures of web references. *Communications of the ACM*, 46(1):71–77.

Spitz, A., Strötgen, J. et 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.

Stack, M. (2006). Full text search of web archive collections. *Proc. of IWAW*.

Stevanovic, D., Vlajic, N. et An, A. (2011). Unsupervised clustering of Web sessions to detect malicious and non-malicious website users. *Procedia Computer Science*, 5:123–131.

Stevenson, M. (2016). Slashdot, open news and informated media : exploring the intersection of imagined futures and web publishing technology. *New Media, Old Media : a History and Theory Reader*, pages 616–630.

Stevenson, M. (2018). From hypertext to hype and back again : exploring the roots of social media in the early web. *The SAGE Handbook of Social Media*.

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.

Sánchez-Querubín, N. et Rogers, R. (2018). Connected routes : Migration studies with digital devices and platforms. *Social Media+ Society*, 4(1).

Tervonen, T. (2016). Finlande : le droit d'asile menacé ? *Plein droit*, (4):7–10.

Tofel, B. (2007). Wayback'for accessing web archives. *In Proceedings of the 7th International Web Archiving Workshop*, pages 27–37.

Toyoda, M. et Kitsuregawa, M. (2003). Extracting evolution of web communities from a series of web archives. *In Proceedings of the fourteenth ACM conference on Hypertext and hypermedia*, pages 28–37. ACM.

Toyoda, M. et Kitsuregawa, M. (2005). A system for visualizing and analyzing the evolution of the web with a time series of graphs. *In Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*, pages 151–160. ACM.

Toyoda, M. et 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.

Tréguer, F. (2017). *Pouvoir et résistance dans l'espace public : une contre-histoire d'Internet (XVe-XXIe siècle)*. Thèse, EHESS-Paris.

Tukey, J. W. (1962). The future of data analysis. *The annals of mathematical statistics*, 33(1):1–67.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Behavioral Science : Quantitative Methods. Addison-Wesley.

Tyner, J. A. et Kuhlke, O. (2000). Pan-national identities : representations of the Philippine diaspora on the world wide web. *Asia Pacific Viewpoint*, 41(3):231–252.

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

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

Van den Bos, M. et Nell, L. (2006). Territorial bounds to virtual space : transnational online and offline networks of Iranian and Turkish–Kurdish immigrants in the Netherlands. *Global Networks*, 6(2):201–220.

Viard, T. (2016). *Link streams for the modelling of interactions over time and application to the analysis of IP traffic*. PhD Thesis, Université Pierre et Marie Curie.

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

Wasserman, S. et Faust, K. (1994). *Social network analysis : Methods and applications*, volume 8. Cambridge university press.

Weikum, G., Ntarmos, N., Spaniol, M., Triantafillou, P., Benczúr, A. A., Kirkpatrick, S., Rigaux, P. et Williamson, M. (2011). Longitudinal analytics on web archive data : it's about time ! *In CIDR*, pages 199–202.

Weltevrede, E. et Helmond, A. (2012). Where do bloggers blog ? Platform transitions within the historical Dutch blogosphere. *First Monday*, 17(2).

Weninger, T. et Hsu, W. H. (2008). Text extraction from the web via text-to-tag ratio. *In Database and Expert Systems Application, 2008. DEXA'08. 19th International Workshop on*, pages 23–28. IEEE.

Whitaker, M. P. (2004). Tamilnet. com : Some reflections on popular anthropology, nationalism, and the Internet. *Anthropological Quarterly*, 77(3):469–498.

Yadav, D., Sharma, A. K. et Gupta, J. P. (2007). Change Detection in Web Pages. pages 265–270. IEEE.

Zijlstra, J. et Liempt, I. V. (2017). Smart (phone) travelling : Understanding the use and impact of mobile technology on irregular migration journeys. *International Journal of Migration and Border Studies*, 3(2-3):174–191.

Couverture : *De la Terre à la Lune* (détail), G. Méliès (1930)
Archives, fragments Web et diasporas, thèse rédigée par Quentin Lobbé entre les mois de mai et aout 2018