Where the dead blogs are

A Disaggregated Exploration of Web Archives to Reveal Extinct Online Collectives

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

ABSTRACT

The Web is an unsteady environment. As Web sites emerge and expand every days, whole communities may fade away over time by leaving too few or incomplete traces on the living Web. Worldwide volumes of Web archives preserve the history of the Web and reduce the loss of this digital heritage. Web archives remain essential to the comprehension of the lifecycles of extinct online collectives. In this paper, we propose a framework to follow the intern dynamics of vanished Web communities, based on the exploration of corpora of Web archives. To achieve this goal, we define a new unit of analysis called *Web fragment*: a semantic and syntactic subset of a given Web page, designed to increase historical accuracy. This contribution has practical value for those who conduct large-scale archive exploration (in terms of time range and volume) or are interested in computational approach to Web history and social science. By applying our framework to the Moroccan archives of the e-Diasporas Atlas, we first witness the collapsing of an established community of Moroccan migrant blogs. We show its progressive mutation towards rising social platforms, between 2008 and 2018. Then, we study the sudden creation of an ephemeral collective of forum members gathered by the wave of the Arab Spring in the early 2011. We finally yield new insights into historical Web studies by suggesting the concept of *pivot moment of the Web*.

Keywords: web archives, exploratory analysis, online migrant collectives, computational social science

1. INTRODUCTION

At the end of the 90's, the development of the Information and Communication Technologies (ICT) reshaped the notion of time, space, and border. The rises of Internet, electronic messaging, and mobile phones provided new remote tools of communication and organisation to worldwide migrant collectives.

By the same time, scientific interests for online representations of diasporas grew rapidly [13]. Digital communities of migrants, such as networks of blogs or online discussion forums were created on top of common cultural values [40] likely to gather people in diasporic cyberspaces. The Web became an environment favourable to the establishment of hubs for migrants to connect with each other [32], to preserve pieces of a scattered collective memory [52], or to become tools for militant purposes[1].

The will to understand the diffusion of the Web throughout diasporic communities finally resulted in the completion of the e-Diasporas Atlas [14] directed by D. Diminescu. It revealed diasporic collectives that organize first and foremost on the Web, as networks of migrant websites, connected to each other through hypertext links. The atlas led to the observation of 10,000 migrant Web sites distributed along 30 diasporic networks (Moroccan, Tunisian, Egyptian, etc.) called *e-Diasporas*[2]. But facing, month after month, the partial or total disappearance of some of the observed migrant Web sites, it was decided to start archiving them weekly or monthly to ensure the preservation of their digital history and to allow forthcoming researches.

Web archiving. Since the creation of the Web at the CERN in the early 90's [10], the loss of the digital content that constitutes the Web itself has been considered a major issue. Started as a volunteer initiative with the creation of Internet Archive by Brewster Kahle in 1996 [20], Web archiving was gradually assumed by various states. Throughout the early 2000's, different strategies of harvesting were implemented: the Australian project Pandora[3] chose to focus on a short selection of Web sites, the Swedish project Kulturaw3[4] went for a wider crawl approach, etc.

With the recognition of the *Charter on the Preservation of the Digital Heritage* by UNESCO [48], the prerogative to archive their own Web domains was finally given to many national libraries (such as .uk, .jp, or .fr). Thereby, terabytes of Web pages were saved worldwide by archiving the genesis of the Web. But surprisingly, after 20 years of Web archiving, it must be said that there is an asymmetry between works focused on upstream archive acquisition [12, 30, 37] and analysis of existing Web archives corpora [44, 45]. This may be the result of: 1) a set of limited explorations tools, 2) the lack of an established framework of exploration.

In practice, most national libraries allow local and limited consultation points with no remote access. The online portal of the WayBack Machine[5] only provides a restrictive search-by-URL system without any full-text search facility. Existing archive exploration tools are designed and effective for refining past versions of a known URL, not for proceeding to a large-scale exploration (in terms of time range and volume). Thus, existing research based on Web archive explorations chooses to manually track the evolution of a given set of URLs [2, 35, 39] or to focus on the visual aspects of an archived Web page [4].

Extinct online collectives. The analysis of Web archives opens new research perspectives for sociologists and historians by offering a longitudinal vision of a corpus of Web sites. In this paper, we study the problematic of extinct online migrant collectives for

[1] The Pajol website retraces the history of the occupation of the Saint-Ambroise church by migrants in 1996 in France: http://www.bok.net/pajol/index2.html

[2] http://www.e-diasporas.fr/

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

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

[5] https://archive.org/web/

which too few or incomplete traces remain on the living Web. We hypothesize that their structure is permeable to the impact of exogenous events or shocks. Our aim is then to search for correlations between a given political and social context and the topographic evolutions of a vanished community. To fulfill this goal, we will face technical challenges: How to query a wide corpus of Web archives other than through a given URL? How to handle large-scale time ranges and volumes of data? How to conduct an exploration of Web archives from a methodological point of view? How to visualize and discuss the results?

Summary of main contributions. We propose a framework based on the exploration of corpora of Web archives, to follow the internal dynamics of extinct online collectives. We use this framework to study two online Moroccan migrant communities extracted from the e-Diasporas Atlas archives: an established collective of blogs and an ephemeral group of forum members. The first one is built on hypertext citations between sites, the second one is a network of users writing forum posts about a shared set of threads. Both communities vanished from the Web at some point before 2018. Moreover, we know from the outcomes of e-Diasporas Atlas that they were both susceptible to be impacted by exogenous political and social events.

Our framework is designed to conduct an exploration of a corpus of Web archives as described in Sections 2 and 3. Thus, we introduce in Section 2 an entity called *Web fragment*: a new unit to query and explore archived corpora that results from the segmentation of a Web page. We show in Section 3 from a theoretical and from an experimental perspective how we can gain benefits of using our framework in comparison to existing Web archive exploration systems (such as the WayBack Machine). We then briefly discuss its technical implementation.

By applying this framework to our corpus of Web archives, we propose a computational approach to Web history and social science. We highlight in Sections 4 and 5 two characteristic examples of explorations that use the features previously introduced. The first one is focused on a defined set of entities (i.e., known blogs) and conducted on a large range of time. The second one is focused on a targeted time period and conducted on an unknown but supposedly wide set of entities (i.e., forum posts). Indeed, in Section 4 we first witness the collapsing of an established Moroccan blogosphere. We show its progressive mutation towards rising social platforms, between 2008 and 2018. Then, we highlight how blogs embedded their own communities of readers and discuss the key role of the Arab Spring in this mutation. The Arab Spring was a revolutionary wave of demonstrations, protests, and civil wars in North Africa and the Middle East that began on December 2010 in Tunisia with the Tunisian Revolution and ended on December 2012[6]. In Morocco the protests started in February 20th 2011 and ended shortly after political concessions made by King Mohammed VI[7].

In Section 5, following the insight revealed in the previous part, we study the creation of an ephemeral collective of members of the forum *yabiladi.com* gathered by the wave of Arab Spring in the early 2011. In particular, we show how some old users converged suddenly around the online organisation of the demonstration of February 20th 2011, without any old-established preparations. This ephemeral collective persisted during a few months before totally disappearing from the forum.

Finally, we discuss in Section 6 the limitations of Web archives as a source of information that can only reflect a partial vision of the Web and witness, in the bulk of this paper, a particular time called *pivot moment of the Web*. A moment of transition between two systems when new Web usages fork from established habits. Here we analyse the specific moment when Web actors chose to move from the Web 2.0 to the Web of social networks in order to support their growing will of expressing themselves.

2. SETUP AND INSIGHTS

In the following, we first describe our experimental dataset of Web archives. Then, we discuss the inspirations and insights of our work. In particular, we address the question of timestamping Web archives. Finally, we introduce the *Web fragment* as a key element of our framework: a new unit to query and explore archived corpora that results from the segmentation of an archived Web page.

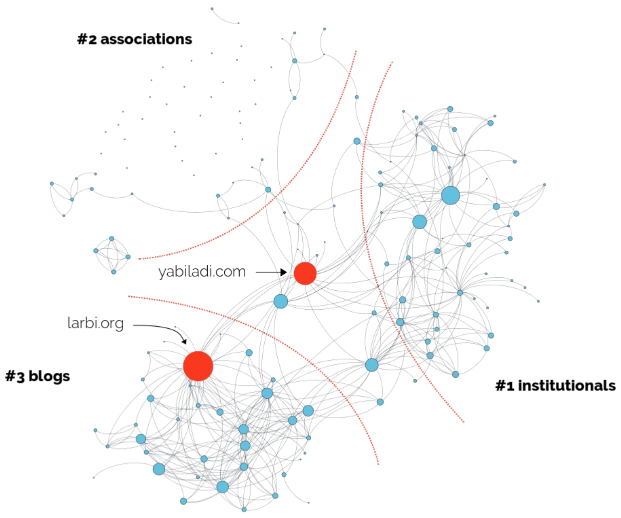

Figure 1: The Moroccan e-Diasporas (mapped by D. Diminescu, M. Renault and M. Jacomy)

Experimental corpus. The framework presented in this paper, is targeted to handle large Web archives corpora. As input data, we use the Moroccan section of the e-Diasporas Atlas corpus, a highly thematically and temporally coherent corpus, designed on purpose[8]. The archive is based on an initial network (called *e-Diaspora*) of 254 Web sites created or managed by Moroccan migrants or that deal with them. The network can be divided in 3 clusters as illustrated in Figure 1: #1 Institutional Web sites managed by the Moroccan government, #2 Associations and NGOs Web sites, #3 The blogosphere edited by citizens. The Moroccan e-Diasporas was initially mapped

[6] https://en.wikipedia.org/wiki/Arab_Spring
[7] http://www.middle-east-online.com/english/?id=44526

[8] The corpus is publicly available and explorable at http://maps.e-diasporas.fr/index.php?focus=map&map=5&section=5

	larbi.org	yabiladi.com
Archive count (e-Diasporas)	78,311	2,683,928
Archive count (Internet Archive)	24,537	887,981
First archived (e-Diasporas)	March 2010	March 2010
First archived (Internet Archive)	Oct. 2002	Feb. 2001
Last archived (e-Diasporas)	Sept. 2014	Sept. 2014
Last archived (Internet Archive)	March 2018	March 2018

Table 1: Statistics of the archives of larbi.org and yabiladi.com

in 2008 and then weekly archived from March 2010 to September 2014. It was recorded under the DAFF format[9].

Later in Section 4, we will focus on a well-framed community of 47 blogs[10]. And, in particular, we will point out the evolution of *larbi.org*: the most influential blog in 2008. Then in Section 5, we will focus on a group of members of *yabiladi.com*: an established forum considered as one of the main hubs of the Moroccan network.

Statistics of their respective archives are given in Table 1. It indicates the variation of scope between the e-Diasporas Atlas corpus and the Internet Archives corpus. In the following, we will mostly use the e-Diasporas corpus for large scale information extraction tasks and, in addition, use the WayBack Machine to perform targeted visual validations.

Timestamping Web archives. We aim to discuss here the benefits of a disaggregated exploration of Web archives in comparison to existing approaches. As a major insight, we first want to upscale the timestamping of Web archives corpora by retrieving their true edition dates.

Indeed, archive file formats (WARC or DAFF) are designed on top of Web pages. They are basically a collection of crawled HTML pages associated with a download date. In existing Web archive explorer systems, when one proceeds to a longitudinal search, the corresponding result is stamped by download date, even if it was effectively created years ago. In this paper, we propose to introduce a scale of accuracy (divided in three levels) to upscale the estimation of the true creation date of an archived Web page. We enumerate them from the least accurate to the most truthful date:

(1) the **download date**: the date when the Web page was archived
(2) the **last modification date**: the date when the Web page was modified for the last time on the living Web
(3) the **first edition date**: the first date when a content of the Web page was edited or published on the living Web

The difficulties to automatically retrieve editions date in a given set of archived Web pages have been addressed by [3], but the benefits in term of historical accuracy are impressive. For instance, in Figure 2, we compare the temporal distribution of all archived pages of *yabiladi.com* in the e-Diasporas corpus, based on download date (red line) versus their effective first edition dates (blue line). In this experiment, we proceed to the manual scraping of the HTML nodes framing the commentary date of each forum post. Here, we assume to approximate the first edition date of each of the archived pages by their corresponding first extracted commentary date. This can reproduce on various archived Web sites on condition that one can clearly identified HTML nodes framing edition dates by using their HTML or CSS class names[11]. By starting with a given dataset of archives collected from 2010, we can extend the comprehension of our corpus to consider contents effectively written up to 2003. Plus, in the distribution by download date, we can see a lack of data around 2013. This is what we call a *crawl legacy effect*. This does not mean that the forum did not produce any content at that time, but in reality it induces that the crawler was stopped during many months. Crawl legacy effects are biases that one has to keep in mind to proceed an exploration: Web archives are not direct traces of the Web, they are in fact direct traces of the crawler..

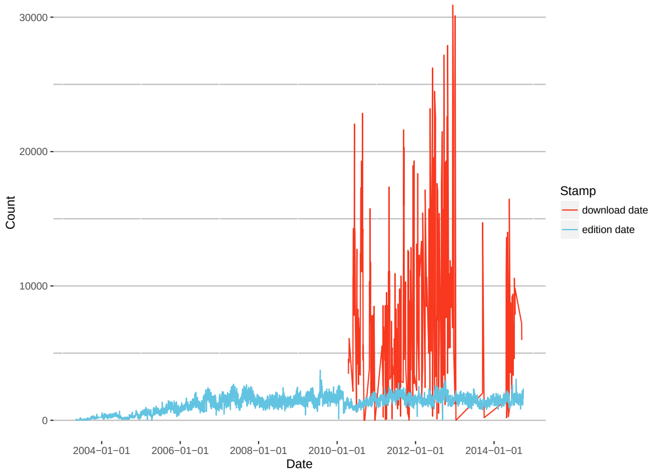

Figure 2: Distribution of the archived pages of *yabiladi.com* using download dates versus edition dates

Going under the level of a Web page. The inspiration of our framework comes from two seminal papers. In [50], Weikum et al. point out the benefits of a longitudinal search engine plugged on top of a Web archive corpus in order to retrieve past digital contents based on their edition dates. In [7], Brügger suggests the possibility of building a dynamic system to resize historical studies from an archived Web page to its individual contents. He then introduces the notion of Web strata:

> "*one can distinguish the following five analytical strata: the web as a whole; the web sphere; the individual website; the individual webpage; and an individual textual web element on a webpage, such as an image.*"

From this point, we will now assume the necessity of finding a new unit of exploration of Web archives corpora, framed within the scope of the fifth Web strata. This unit should be, as much as possible, related to an edition date to avoid all kinds of crawl

[9] Digital Archive File Format as opposed to the more popular WARC archives file format, DAFF has the particularity to separate the meta data contents (URL, download date, MIME type, charset, etc.) from the data contents (original HTML content)

[10] See the complete list here http://maps.e-diasporas.fr/index.php?focus=value&graph=2&map=5&nodeattribute=5&section=5&value=blog

[11] https://www.w3schools.com/html/html_classes.asp

legacy effects and maximise the historical accuracy. As the shape of this unit is linked to a specific exploration context and as we want that sociologists or historians could be able to seize it in their own works, the following definition will be purposely generic. A practical definition, targeted for the question of extinct online migrant collectives, will be given later in Section 3.

The Web Fragment. Considering the Web page as the basic unit of access to the World Wide Web, built using its own digital writing modalities, and noticing that from the point of view of human perception [5, 31] a Web page is the result of the logical arrangement of distinct semantic components, we define a Web fragment[12] as a semantic and syntactic subset of a given Web page.

There is a scale relationship between a Web page and its Web fragments. A Web fragment is a coherent set of textual, visual, audio or animated contents extracted from a Web page. The Web fragment should be comprehensible on its own. Within the same Web page, two fragments cannot overlap, even partially. A Web fragment must go with an associated set of extracted meta contents (an author, a title, an edition date, etc.) and it must also encompass all the writing and sharing elements used for publishing this content on the Web (CMS widgets, integrated text editor, hypertext links, rss feed, etc.).

3. DISAGGREGATING WEB ARCHIVES

In this section, we first discuss the benefits of upscaling the historical analysis of Web archives by using the Web fragment instead of the Web page as unit of exploration. Then we proceed by describing (from a technical perspective) our framework for disaggregating Web archives, extracting Web fragments, and conducting a temporal exploration on top of it.

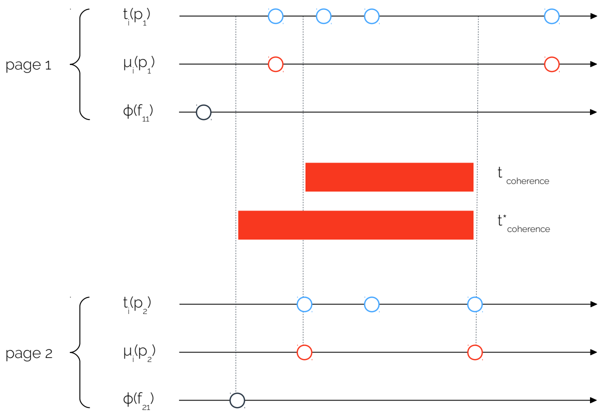

Figure 3: The successive crawls of pages p_1,p_2 and their coherence intervals

3.1 Upscaling the exploration

Existing exploration systems (such as the WayBack Machine) use the Web page as unit of analysis. In the following, we will emit the hypothesis of a framework based on a more accurate unit: the Web fragment as introduced in Section 2. We will anticipate the possibility of indexing Web fragments instead of Web pages and explain how our framework can reduce subsequent exploration issues: 1) crawl blindness, 2) observable coherence, 3) duplicate archived contents.

Assumptions and notation. We first assume that an archived Web site consists of n Web pages numbered $\{p_1,...,p_n\}$. A corpus of Web archives is the result of one or many successive crawls $\{c_1,...,c_l\}$. We call the process of downloading the Web pages $\{p_1,...,p_n\}$ of an entire Web site a crawl c_i. The time taken for downloading pages is neglected. We call $t_i(p_j)$ the download date of page p_j during crawl c_i. The first download date of page p_j is denoted as $\min_i t_i(p_j)$. We assume to know the last modified stamp of page p_j denoted $\mu_i(p_j)$ during crawl c_i (having $\mu_i(p_j) \leq t_i(p_j)$). Figure 3 gives a visual illustration of this notation for the two archived Web pages p_1,p_2.

Crawl blindness. In Section 2, we introduced the notion of *crawl legacy effect* that covers all the mechanical impacts of a crawl c. We call here *crawl blindness* the action of mis-timestamping a change on a page p_j after a crawl. A change can be the creation, the update, or the deletion of all or part of a Web page [11, 26]. Whereas it is usually hard to retrieve stamped traces of updates and deletions, we know that an act of creation (adding a post, a comment, etc.) can be approximately re-timestamped by finding edition dates [19]. In existing exploration systems (such as the WayBack Machine), the creation date of a page p_j is approximated by its first download date $\min_i t_i(p_j)$. In this paper, we argue that a Web page p_j consists of m Web fragments numbered $\{f_{j1}, \ldots, f_{jm}\}$. We also assume to know the edition date of each Web fragment $\phi(f_{j1}), \ldots, \phi(f_{jm})$, so that:

$$\forall p_j, f_{jk} \exists \phi(f_{jk}) : \phi(f_{jk}) \leq \mu_i(p_j) \leq t_i(p_j)$$
$$\text{where } c_i \text{ is a crawl in which } f_{jk} \text{ exists}$$

Figure 2 already illustrated this intuition, by comparing the distribution of a set of first download dates versus their corresponding true edition dates. As an experiment, we select now the 109,534 archived Web pages of the forum section of *yabiladi.com* stamped by first download date. We split them into 422,906 Web fragments stamped by their edition dates. We then select the first edition date $\min_k \phi(f_{jk})$ of each archived page to approximate a more accurate creation date. Then, we calculate the difference $\min_i t_i(p_j) - \min_k \phi(f_{jk})$ between the first edition dates and the first download dates in days. The corresponding quartiles are given in Table 2. With our framework, doing an exploration on top of Web fragments is more historically accurate than looking at the original Web pages stamped by download dates.

Disaggregated observable coherence. During the exploration of a corpus of Web archives, one may need to consider the coherence between two or more archived pages sharing, for instance,

[12] Not to be confused with the notion of a url fragment identifier (https://en.wikipedia.org/wiki/Fragment_identifier)

Where the dead blogs are

quartiles	difference in days
Q1	256
Q2	777
Q3	1340

Table 2: Quartiles of the difference $\min_i t_i(p_j) - \min_k \phi(f_{jk})$ in days

a hypertext link. Is the time period between the linked records a matter of hours, days, or years? In [42, 43], it has been defined that two or more pages are observable coherent if there is a single time point $t_{\text{coherence}}$ so that there is a non-empty interval spanning the respective download time $t_i(p_j)$ and last modified stamp $\mu_i(p_j)$. In the notation of this paper, using the most recent Web fragment f_{jk} of a Web page p_j, two or more pages are observable coherent if:

$$\forall p_j, f_{jk}, \exists t_{\text{coherence}} : t_{\text{coherence}} \in \bigcap_{i=j}^{n} [\phi(f_{jk}), t_i(p_j)] \neq \emptyset$$

But here, we can go further and introduce the notion of disaggregated observable coherence. For instance, if the explorer want to check the coherence between two news articles, resulting of a full text search and based on a specific query, he may want to know the nature of the fragments framing the shape of the interval $t_{\text{coherence}}$. Thus, the coherence may be irrelevant if the only invariant fragments are navigation bars instead of text articles. So we can define a discrete subset of fragments of interest $\{f^*_{j1}, ..., f^*_{jl}\}$ having $l \leq m$ so that:

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

$$t^*_{\text{coherence}} \in \bigcap_{j=1}^{n} [\phi(f^*_{jk}), t_i(p_j)] \neq \emptyset$$

Figure 3 depicts, in a graphical way, the difference between $t_{\text{coherence}}$ and $t^*_{\text{coherence}}$ of pages p_1 and p_2 based on the selected Web fragments f^*_{11} and f^*_{21}.

Duplicate archived contents. In existing exploration systems, indexing the text content of an archived Web page induces the presence of many duplicate items. For instance an unmodified comment will be indexed as many times as it has been crawled. This will cause unnecessary memory consumpting and inaccurate results. By using Web fragment as a unit of indexation, we can define an identity function *id* to compare the invariance of a fragment f_{jk} extracted from a page p_j during two consecutive crawls c_1 and c_2 at $t_1(p_j)$ and $t_2(p_j)$ so that:

$$id(t_1(f_{jk})) = t_2(f_{jk})$$

As a practical proposition, we test the equality of the outcomes of a *SHA-2* hash function applied to the text content of the unduplicated Web fragments extracted from the forum section of *yabiladi.com*. The quartiles of the distribution of the number of duplicated fragments are given in Table 3. The benefit here is less visible than previous results. In fact, it induce that each pages of *yabiladi.com* was crawled, in average, only twice. But we still want to point out that the maximum number of duplicated fragments in this example is 45. It corresponds to the front page of the forum. So, at the very least, our framework can avoid to re-index a reasonable number of archived contents.

quartiles	number of duplicates
Q1	1
Q2	1
Q3	2

Table 3: Quartiles of the numbers of duplicated Web fragments

3.2 Extracting Web fragments

In the following, we will move from theoretical perspectives to practical implementation: building a framework for exploring a corpus of Web archives disaggregated in multiple Web fragments.

Implementation. Due to the particularities of the DAFF format used to store the e-Diasporas archives, we have to build our own system from scratch, but other initiatives (well suited for WARC formalism) are also in development[13]. Our architecture is released under an open-source license[14] and follows a classical implementation model [29]: 1) DAFF files are grabed by a Java extractor and then uploaded into a Hadoop Distributed File System (HDFS[15]). 2) A Spark pipeline[16] ingests the HDFS and filters the archives by domain names or ranges of dates. 3) A dedicated library extracts the Web fragments out of the archived pages. 4) The text content of each fragment is indexed into a Solr[17] search engine. 5) Then we build different data visualizations to query and explore the archives.

Extraction method. In Section 2, we described the Web fragment as a semantic and syntactic subset of a given Web page. Thus, an extracted Web fragment should be equally understood by computer scientists and social scientists. For instance, a historian would be interested in seeing both the content and the context of a blog post (its date, author, title, comments, etc.) and not only a set of keywords. We now describe our extraction method.

We consider an archived Web page p as a finite set of m HTML nodes $\{n_1, ..., n_m\}$, organized as a DOM tree t and associated with some CSS style rules. First, we clean the DOM tree using the boilerplate method of [23] to filter out ads and navigation nodes. Knowing that the extraction of individual components out of Web pages is mostly designed for large-scale processes [1, 9, 19, 23, 34, 51], we prefer to follow user-centric strategies such as Mozilla's Readability[18] and Fathom[19] projects. As Readability was designed to find the most important part of a Web page (like an article), we extend it using the Fathom agglomerative clustering algorithm to find all the remaining coherent clusters of HTML nodes. In the Fathom algorithm, all the HTML nodes are initially stored in an $m \times m$ sparse adjacency matrix called A. An agglomerative clustering is then applied node by node, having the rows of A

[13] http://archivesunleashed.org/about-project/
[14] https://github.com/lobbeque/archive-miner
[15] http://hadoop.apache.org/
[16] https://spark.apache.org/
[17] http://lucene.apache.org/solr/
[18] https://github.com/mozilla/readability
[19] https://github.com/mozilla/fathom

incrementally going from single nodes to clusters of nodes. A pseudo-code implementation of it is given in Algorithm 1. We call d the distance function resulting of the depth difference between two nodes in the DOM tree t. We assume the existence of a function named *closestRows* that returns the two closest rows of A based on the distance between their respective nodes. The variable *minDist* is the minimal distance to allow for agglomerate two nodes.

```
while rows(A) > 1 and closestRows(A) < minDist do
    $\{r_i, r_j\} = closestRows(A)$
    $newRow = \{\}$
    for $r \in rows(A)$ do
        if $r \neq r_i$and$r \neq r_j$ then
            $newRow[r] = \min(d(r_i, r), d(r_j, r))$
        end
    end
    remove($A[r_i]$)
    remove($A[r_j]$)
    remove($A[*][r_i]$)
    remove($A[*][r_j]$)
    append($A, newRow$)
end
```

Algorithm 1: Fathom agglomerative clustering

In the scope of this paper, completion time is not the main selection criteria since the extraction of Web fragments is done at indexation time and not at retrieval time. Thus, as a contribution, we extend the distance function d of Readability with visual-based penalties introduced by [9] and tag-based penalties introduced by [19] to handle the *"human perception"* part of the definition introduced in Section 2. In practice, we initialize the variable *minDist* for each Web site after human validation. For each remaining cluster, we parse the HTML and CSS class-names of all the constitutive nodes using a set of dedicated regular expressions to identify and extract edition dates. Finally, we index the text contents as well as all the HTML id and class names[20]. To sum up, a Web fragment is a coherent cluster of HTML nodes.

4. ARCHIVED TRACES OF DIGITAL MUTATION

We now transition from the description of our framework to the second contribution of this present work: addressing the question of extinct online migrant collectives for which too few or incomplete traces remain on the living Web. In particular, we focus on analysing the collapsing of an old-established community of migrant Moroccan blogs between 2008 and 2018. We first define the space of exploration and emit the hypothesis of a digital mutation from the blogs to rising social platforms. We discuss time granularity and query specificity. The results,suggest that almost half of the blogs moved to Twitter, Facebook, etc. There, they recreated and increased their former collective and conserved their diasporic aspect. We show that a part of their respective community of readers followed them into social platforms. And we finally discuss the key role of Arab Spring in this mutation.

An old-established blogosphere. In 2008, a set of 47 blogs linked together by hypertext citations and created or managed by Moroccan migrants was discovered and mapped as illustrated by Figure 1 (full network) and 5 (close up, left). With a high density of edges within them and a lower density of edges shared with the rest of the network [41, 49], those blogs were categorized as a community. The political blog *larbi.org* possessed the highest in-degree (i.e, the number of edges incoming to a vertex) and occupied a central position in the community[21]. The 47 blogs used French as a main language and produced a bundle of political thoughts, daily moods and intimist texts. In 2015, a first report [22] induced that many of the blogs were no longer active or deleted. By updating this survey in 2018, we show that 20 blogs are now dead, 22 have not been updated since at least 2 years and only 5 are still alive (see Figure 5, right).

Exploration task. We follow the principes of Exploratory Data Analysis [46] (EDA) and its recent updates [15]. EDA is a fundamentally iterative process that is deliberately part of a logic of observation, discovery, and astonishment [47] well-suited for our leading question: What happened to the dead blogs?

For precision, we call *author* someone who created or managed one of the targeted blogs and we call *reader* someone who visited or commented them. We first emit the two following hypotheses: 1) the blogs totally disappeared, the authors of the sites deleted them and moved definitively away from the Web, 2) the blogs changed their digital skins, the authors migrated from one Web territory to another (such as Twitter or Facebook). Therefore, we define our task of exploration as finding, inside the archives of the blogs, the past traces of a digital mutation. As first clues, we know that Facebook groups were contemporary of the blogs [28] and we also know that social platforms like Twitter were inspired by blogging practices [24]. Thus, there might be some bridges between the Moroccan blogs and those social platforms. We choose to target the first traces of social media in the whole archives, such as a *"tweet"* button or a *"like"* widget. In our framework, we request for fragments containing HTML nodes related to social networks like: *<span class="Twitter"></span>*, *<button class="Facebooks-share"></button>*, etc or directly mentioning *facebook, youtube, pinterest*, etc inside their textual contents. The space of exploration is focused on the 47 blogs but is not limited in time.

The recomposition of the community. The results consist in a filtered set of fragments timestamped and grouped by blogs. Some of them contain the URL or the account name of linked social media. We use the WayBack machine to visualy validate each URL and deliver a qualitative analysis[22]. After managing a blog of their own, 20 authors moved to a social platform: 8 have a Facebook standalone page, 16 are on Twitter and a minority of them use Youtube, Pinterest, Flickr or Medium. Created between 2005 and 2007, the blogs slowly died or felt asleep around the early 2010's. Only two authors recreated a new blog after the death of their

[20] See https://github.com/lobbeque/rivelaine for the whole implementation and https://frama.link/XYj1FNSY for the set of regular expressions

[21] Larbi.org was elected as best Moroccan blog in 2008 https://fr.wikipedia.org/wiki/Maroc_Web_Awards

[22] The results can be download here https://frama.link/FP-T6Z8_

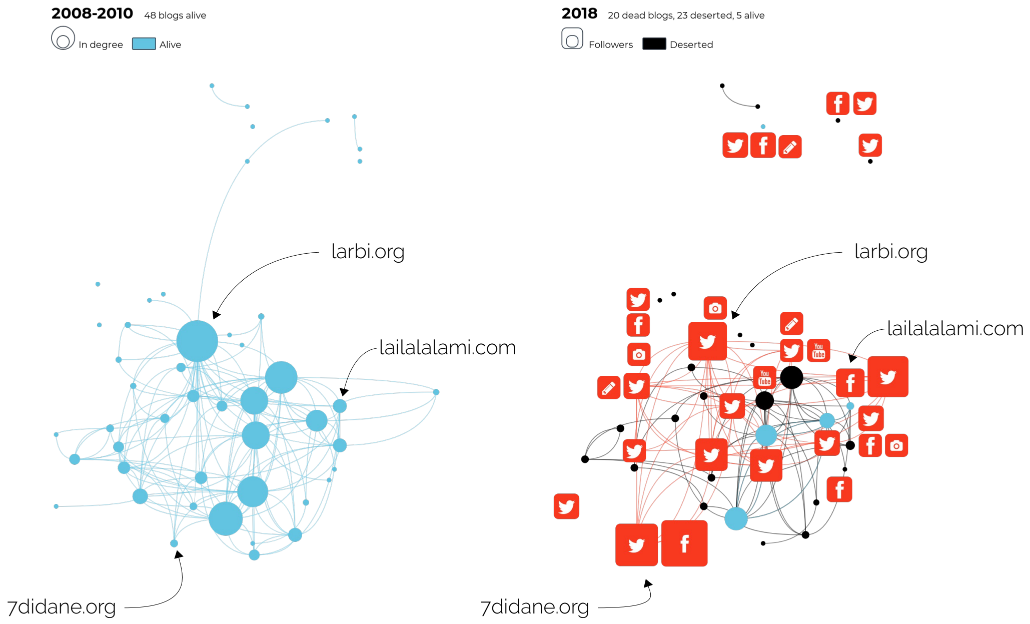

Figure 4: Evolution of the Moroccan blogosphere between 2008 (left) and 2018 (right) with a kept position

former one. Blogs as a medium don't seem to be a space of expression they chose to prolong. Keeping alive its digital identity is a shared characteristic in this community, as all the authors reused their pseudonyms (or a close variation of them) on the social media. Thus, the blogosphere mutated into a multitude of social accounts. The online expression is now fragmented and specialized by type of medium. Some choose to have both a Facebook and a Twitter account like *7didane.org*. Others use Youtube or Flicker to upload videos and photos like *larbi.org*. We can observe the dual-use of Twitter alongside Medium, where one writes a long piece of text on Medium and chooses to promote it by using Twitter like *eatbees.com*. We show that, focusing on the authors that moved on Twitter, the density of the graph of follower/following is higher than the density of the old corresponding citations graph: it goes from 0.16 in 2008 to 0.24 in 2018. The community aspect of the old blogosphere is conserved and even increased.

Followed by the readers. In Figure 4, the size of each new social node is correlated to the size of their community of followers or friends. For instance *7didane.org* is followed by 43,512 people on Twitter and has 141,947 friends on Facebook. In the new age of social platform, the influence of an author is usually linked with the volume of readers he can communicate with. So the internal dynamics of the blogosphere changed as well: *larbi.org* grew down as *7didane.org* rised up. By looking at the social profile of the followers and friends, we show that the diasporic characteristic of the community is conserved. Authors still speak from the outside of Morocco to both Moroccan residents and migrants. We use the netvizz app[23] to extract the country of origin of the followers of each Facebook page. In practice, *7didane.org* is mainly connected to Morrocan people (82%) but also French (2%) and Egyptian (2%) ones and *lailalalami.com* still speaks to Morrocan (24%), American (15%) and Pakistani folks (8%). As we explained that authors kept their online identities, we now emit the hypothesis of finding the same behavior with the readers. It is reasonable to said that in the case of a crowd-engaging medium like blogs, when a strong connection is created between an author and its readers, they may want to preserve this relation during the process of digital mutation. So, we assume that readers conserved their pseudonyms on Twitter to follow the authors they supported on blogs. As an experiment, we request for Web fragments following the template of a comment on *larbi.org*: a user name, a date and a text contents disposed bellow the main article of an archived page. We then extract the pseudonyms of 4177 past readers of *larbi.org* and compare them to its actual followers on Twitter. This results in a lower bound of 647 persistent readers that followed the author on Twitter. Those readers represented a significant part of the past audience of the blog, as they wrote 26% of all the comments posted.

The Arab Spring as a key moment. In our archived corpus, only 6 blogs wrote a clear farewell message before dying such as *7didane.org*[24]. Saying goodbye is not a discriminant characteristic

[23] https://apps.facebook.com/netvizz/
[24] https://web.archive.org/web/20120415100250/http://www.7didane.org:80/

of this community, blogs usually stay non-updated during many months before suddenly vanishing. But our Web archives contain some interesting elements. For instance, *7didane.org* wrote[25] that he discovered Twitter by following the 2009 protests in Iran[26]. We also notice that *larbi.org* first publicly mentioned Twitter by the end of 2010 during the Arab Spring and pointed out the use of Twitter as a tool to organized citizens actions for the upcoming protest of February 20th, 2011 with the hash-tag #20Fev [27]. It's hard to say out of those too few examples that political mobilisations caused the mutation of blogs into social media. But we can reasonably say that the Arab Spring may have been a key moment for the authors to discover the democratic possibilities of those social platforms. Thus, if the blogs reacted to the particular protest of February 20th, 2011 by pointing out Twitter as a promising space of expression, we might be able to find other impacts of this mobilisation in our archived, as discussed in the following.

5. AN EPHEMERAL PROTEST COLLECTIVE

We have so far concentrated on characterizing the mutations over the past 10 years of a community of blogs. The relative impact of the Arab Spring on the Moroccan migrants blogoshere may have been one of the many factors of its progressive digital transformation to social platforms. The revolutionary wave of protests that spread through the Maghreb and Middle East between 2010 and 2012 was in many ways influenced by an active use of social media as a means for collective organisation. In particular, social platforms played a key role in the empowerment of Egyptian and Tunisian populations [38]. In Morocco, the protests occurred early in 2011 and culminated on February 20th, 2011 when over 10.000 Moroccans demonstrated to demand democratic reforms[28]. In our archives, the large forum of *yabiladi.com* has remained central and stable over nearly 15 years and by extension during the Arab Spring. So, we turn now to the possibility of identifying passing groups inside archived Web pages. We define a strategy of exploration based on the daily observation of Web fragments corresponding to forum messages. We emit the hypothesis of a community of members gathered around the events of the February 20th. The results discussed next suggest that some old users converged suddenly around the online organisation of this demonstration, without any old-established preparations. This ephemeral collective persisted during a few months before disappearing in its entirety from the forum.

A hub for Moroccan migrants. Figure 1 illustrates the key role of *yabiladi.com* in the Moroccan e-Diasporas. As a bridge between the blogosphere and institutional Web sites *yabiladi.com* is an old-established and hybrid place. Created in the late 2001 by a young programmer living in France[29], the site bears witness to the early times of diaporic Web portals. They appeared to be hubs for spreading multi-support informations, meeting places for the migrants living abroad, or cultural showcases [25]. In 2002, *yabiladi.com* opened a forum section, organised in categories and threads that quickly expanded to receive thousands of daily contributions. The conversations, there, were characterized as a mix between reactions to Moroccan and international actualities and daily life considerations: cooking, family, religion, etc.

Features of the exploration. We define our task of exploration as finding inside the archives of *yabiladi.com* a community of users who wrote at least one message in a thread related to the protest of February 20th. Thus, we choose to target forum posts as textual unit of exploration: they are timestamped messages, written by a single (or assume to be single) user and linked together as topic labelled threads. In our framework, we request for fragments following the given ordered template: 1) a user name 2) a date 3) a text content. Then we filter the results set by selecting the fragments associated to an URL that contains the path *"/forum/"*. By analysing the archived URLs, we can group the fragments by category. Indeed, their pattern of URL follows:

*/forum/**thread_title-category_id-thread_id**.html*

We then assume the remaining Web fragments to be only composed of archived forum posts. We restrict the space of exploration to the two categories: *General* and *Moroccan and Worldwide Actuality* as explained in Table 4.

	General	Actuality
Number of threads	17,025	20,553
Number of posts	352,231	328,965
Number of users	19,745	10,819
First archived post	Janv. 2002	Janv. 2002
Max. posts per user	5,019	6,041
Max. threads per user	1,316	2,057
Avg. posts per thread	21	16

Table 4: Statistics of the archived forum of *yabiladi.com*

Revealing a collective. Our aim here is to reveal a collective of members. By revealing, we mean finding good filters and insights able to mine relevant informations. To this end, we start by targeting some specific threads as entry points. We query our system for a set of thread titles matching the French keywords: *"#20Fev"*, *"20 fevrier"* and *"manifestation fevrier"*. We manually validate 12 threads of messages out of them, whether they directly deal with the organisation of the protest or react afterwards to it. We call V_0 this initial group of 12 threads, consisting of 196 messages written by a set of 94 unique users named E_0. We then use E_0 as a new layer of selection in order to reveal a structure of community. We select all the threads where at least 2 users of E_0 wrote a message. This new group of 343 threads is called V_1 and we can now define the graph $G = (V_1, E_0)$ as a network of threads linked by co-contributors[30]. With Figure 5, we visualize G on a timeline (in

[25] https://web.archive.org/web/20090627012354/http://www.7didane.org:80/2009/06/16/1453/

[26] https://en.wikipedia.org/wiki/2009_Iranian_presidential_election_protests

[27] https://web.archive.org/web/20110319191709/http://www.larbi.org:80/post/2011/02/Morocco-Feb20-Maroc-20Fev

[28] https://en.wikipedia.org/wiki/Moroccan_constitutional_referendum,_2011

[29] http://lavieeco.com/news/portraits/mohamed-ezzouak-le-mre-qui-a-lance-yabiladi-com-6596.html

[30] Downloadable results V 1 : https://frama.link/_eModem_, E 0 : https://frama.link/hcxacx89

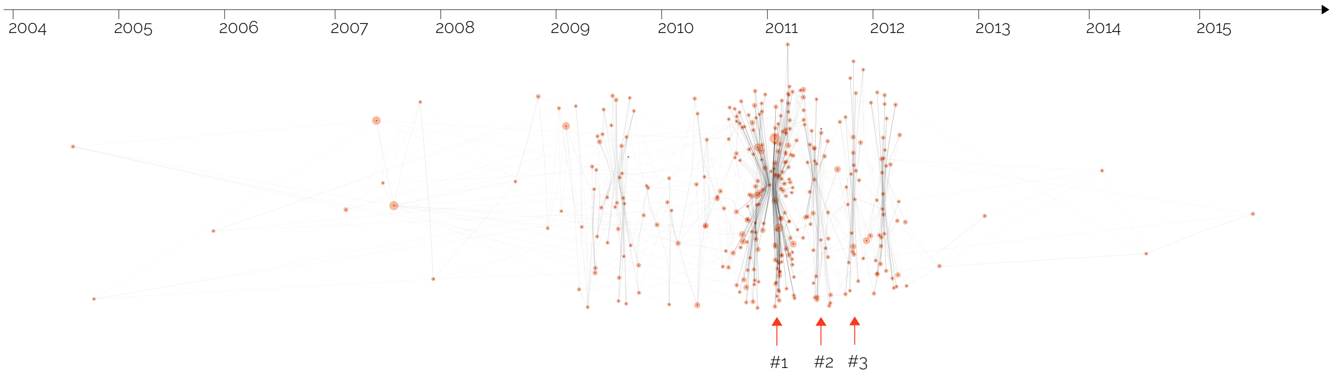

Figure 5: Time distribution of *yabiladi.com*'s threads related to the February 20th

abscissas). Red dots refer to the threads V_1 stamped by the date of their first contribution and sized by numbers of posts. Black links represent the users E_0 writing messages from one thread to another. Links are stacked using the edge bundling method [17] to improve the visual comprehension of the Figure 5. The vertical position of each thread (in ordinate) is a fixed and arbitrary value chosen to clarify the reading of this visualisation, it will not be taken in account in the upcoming analysis. We find in Figure 5 a specific moment (label #1) where the threads of G are very densely aggregated. This moment could be seen as a *fixation point* when the whole graph seems to gather at once. Between January and February 2011, 25% of V_1 were created. This obviously makes sense regarding our original set V_0 but also indicates that the protest of February 20th aggregated an old established community of users that were already using the forum. If we split the Figure 5 in two parts, separated by label #1 and called respectively *pre-protest* and *post-protest* parts, we see that the pre-protest part represent a wide and sparse subgraph spread over a long period of time (from the early 2004 to 2011). In fact, 62% of E_0 users wrote their first message on the forum during the *pre-protest*, and in particular 20% of E_0 registered to *yabiladi.com* in 2007-2008 following a huge wave of new members. They suddenly aggregate each other around label #1 and subsequent fixation points (labels #2 and #3) of the *post-protest* part. We know that the remaining 38% of E_0 contributed first and foremost to label #1 and to the rest of the *post-protest* threads.

To sum up, we have two different patterns: 1) old established users converging as a group by the time of label #1, 2) new members arriving directly on *yabiladi.com* to contribute to the conversation of label #1 and taking part to the *post-protest* debates. But both parts similarly and suddenly disappeared in the early 2012.

Refine the results. To better understand the dynamics of convergence of old and new members around the protest of February 20th, we refine our comprehension of G by conducting a clustering analysis out of it[31]. In Figure 6, threads are spatialized using a force atlas algorithm [18] and clustered by modularity class [6]. In the same way as Figure 5, we find that threads are distributed along a longitudinal axis, confirming the temporal aspect of G. Members contribute from threads at t_i to threads at t_{i+1} with only very few steps backwards: a forum is seen as a continuous flow of upcoming new threads. Clusters of threads can be interpreted as subsequent moments of the evolution of G. We can follow its own dynamics by analysing titles and messages of each of the 8 clusters highlighted in Figure 6. Cluster #1 deals with internal debates about the functioning of the forum. Cluster #2 and #3 bundle daily-life considerations. Then cluster #4 focus on thoughts about the Moroccan identity and comparisons between Morocco and other Maghreb countries. Suddenly, cluster #5 witnesses the rise of a majority of threads related to the protest of February 20th after having questioned the legitimacy of the Moroccan monarchy. Cluster #6 aggregates post-protest messages and diverse political contributions. Cluster #7 deals with the political legacy of the protest, by debating about the new Moroccan constitution announced in March 2011. And finally, cluster #8 goes back to daily life conversations.

To sum up, this exploration indicates that the protest was not really prepared online. A sudden spark fired a minor part of *yabiladi.com*: 94 active users out of a total of 30,564. This wave aggregated old-established members and new comers by breaking daily talks habits. The mobilization did not last in time and stopped with the reform of the constitution conducted by the Moroccan government later on. The community of users that participated to the protest vanished in the early 2012 just a few months before Twitter's first appearance on the front page of *yabiladi.com* in June 2012[32]. Out of the 94 users of E_0, we find that at least 26 of them created a Twitter account using the same user names[33]. This last result finally induces that, by using corpora of web archives, we may have been missing a major aspect of our problematic as explained in Section 6.

6. IMPLICATION FOR HISTORICAL WEB STUDIES

The development of the framework discussed in Section 2 and 3 was guided by the idea that a Web site should become the object of historical studies [4, 8, 39, 50]. To this end, we show that analysing Web archives on large time scales (Section 4) or on large spaces

[31] Downloadable results (as a GEXF graph file) G: https://frama.link/BZdU8CW8

[32] https://web.archive.org/web/20120627021244/www.yabiladi.com

[33] Manually counted and validated in April 2018

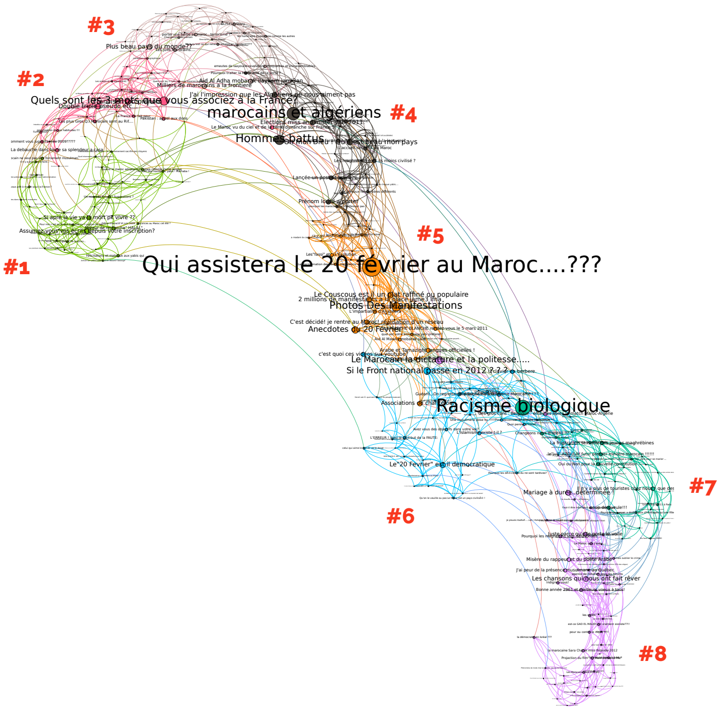
#3
Plus beau pays du monde??
#2
Quels sont les 3 mots que vous associez à la France?
marocains et algériens
#4
Hommes battus
#5
Qui assistera le 20 février au Maroc....???
#1
Le Couscous est-il un plat raffiné ou populaire
Photos Des Manifestations
Anecdotes du 20 Février
Le Marocain, la dictature et la politesse....
Si le Front national passe en 2012 ? ? ?
Racisme biologique
#7
#6
Mariage à durée déterminée
Les chansons qui vous ont fait rêver
#8

Figure 6: The network and communities of February 20th, 2011

of exploration (Section 5) can lead to a better understanding of extinct online uses. We find clues of a mutation of Web authors and readers (blog owners, commentators, forum contributors, etc.) from classical blogs and Web portals to social platforms like Facebook or Twitter. Indeed, by looking through Web archives, one can only find sparse traces of this transition (Section 4) or rather a sudden absence of traces (Section 5). Here, we reach the limits of Web archive corpora and should consider the idea that Web archives may be intrinsically incomplete and somehow partial. Mostly created and designed during the early 2000's [30], Web archiving systems followed the subsequent evolutions of the Web as a medium [11, 33, 36] but still fail to convey the web as an ecosystem [7]. The living Web is a flow of informations, creations, mutations, etc, where various actors are organically interrelated. The archived Web is a fixed set of discrete snapshots where records are stored apart from each other. While we were looking at the archived consequences of the Arab Spring in Section 4 and 5, Web actors were already moving away from forums and blogs [21, 27]. The problematic of extinct online collective of migrants, is less a question of disappearance than a question of transition and Web archives corpora only witness the first leap of this transition.

Pivot moment of the Web. In the same way as the long history of writing that was punctuated by key moments (oral to written expression, invention of printing press, etc), the Web and the Internet in general already possesse their own micro-history. We call *pivot moment of the Web* a period of transition between two systems, a moment when new Web uses fork from established habits and create gaps. A pivot moment arise from three factors: the convergence at a specific time (1) between a technological leap (2) and users sieving it (3). This leads the Web in a new direction of development. As examples, we can highlight the expansion of the MsgGroup [16] during the early 1980's, the democratization of DSL in the late 1990's or the development of smartphones and mobile Web in the 2010's. Sections 4 and 5 of the present work witness, through the lens of Moroccan migrant sites, the specific moment when Web actors chose to move from the Web 2.0 to the Web of social network in order to support their growing will of sharing informations and expressing themselves. Our Web archives (by their contents and shapes) mirrored the very last stand of a dying Web that did not reflect the reality of its time.

7. CONCLUSION

In this paper, we proposed a framework to follow the internal dynamics of extinct online communities and conduct large scale Web archives exploration. We introduced an entity called *Web fragment*: a semantic and syntactic subset of a given archived Web page. By applying this framework to the Moroccan Web archives of the e-Diasporas Atlas, we studied the interactions between online groups, exogenous historical events and technological leap on the archived Web. In the continuity of this analysis, we will support further researches to improve the Web fragment and its multiple uses as a unit of exploration. At the border between computer sciences and digital sociology, our work opens promising questions in terms of historical Web studies. In particular, it would be interesting to consider corpora of Web archives as records of a past ecosystem. We should address the question of mutations and transitions of Web uses regarding nearby *pivot moments*.

REFERENCES

[1] Eytan Adar, Mira Dontcheva, James Fogarty, and Daniel S Weld. 2008. Zoetrope: interacting with the ephemeral web. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*. ACM, 239–248.

[2] Yasmin AlNoamany, Michele C. Weigle, and Michael L. Nelson. 2016. Detecting off-topic pages within TimeMaps in Web archives. *International Journal on Digital Libraries* 17, 3 (2016), 203–221.

[3] Mohamed Aturban, Michael L Nelson, and Michele C Weigle. 2017. Difficulties of Timestamping Archived Web Pages. *arXiv preprint arXiv:1712.03140* (2017).

[4] Anat Ben-David, Adam Amram, and Ron Bekkerman. 2018. The colors of the national Web: visual data analysis of the historical Yugoslav Web domain. *International Journal on Digital Libraries* 19, 1 (01 Mar 2018), 95–106.

[5] Michael Bernard. 2003. Criteria for optimal web design (designing for usability). *Retrieved on April* 13 (2003), 2005.

[6] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

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

[8] Niels Brügger and Ralph Schroeder. 2017. *The web as history: Using web archives to understand the past and the present.*

[9] Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma. 2003. Vips: a vision-based page segmentation algorithm. (2003).

[10] CERN. 1993. The document that officially put the World Wide Web into the public domain. (1993). http://cds.cern.ch/record/1164399

[11] Junghoo Cho and Hector Garcia-Molina. 1999. *The evolution of the web and implications for an incremental crawler.* Technical Report. Stanford.

[12] Dimitar Denev, Arturas Mazeika, Marc Spaniol, and Gerhard Weikum. 2011. The SHARC Framework for Data Quality in Web Archiving. *The VLDB Journal* 20, 2 (April 2011), 183–207.

[13] D. Diminescu (dir.). 2005. Les documents numériques, Méthodologie d'archivage et perspectives de recherche sur les migrations. *Migrance* 23 (2005).

[14] Dana Diminescu (dir.). 2012. *e-Diasporas Atlas. Explorations and Cartography of Diasporas on Digital Networks.* Ed. de la Maison des Sciences de l'Homme, Paris.

[15] Benjamin Jotham Fry. 2004. *Computational information design.* Ph.D. Dissertation. Massachusetts Institute of Technology.

[16] Katie Hafner and Matthew Lyon. 1998. *Where wizards stay up late: The origins of the Internet.* Simon and Schuster.

[17] Danny Holten. 2006. Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Transactions on visualization and computer graphics* 12, 5 (2006), 741–748.

[18] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. 2014. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PloS one* 9, 6 (2014), e98679.

[19] Adam Jatowt, Yukiko Kawai, and Katsumi Tanaka. 2007. Detecting age of page content. In *Proceedings of the 9th annual ACM international workshop on Web information and data management.* ACM, 137–144.

[20] B. Kahle. 1997. Preserving the Internet. *Scientific American* 276 (March 1997), 82–83.

[21] Habibul Haque Khondker. 2011. Role of the new media in the Arab Spring. *Globalizations* 8, 5 (2011), 675–679.

[22] Jihane Khouzaimi. 2015. e-Diasporas : Réalisation et Interprétation du corpus marocain. (2015).

[23] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate Detection Using Shallow Text Features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM '10).* ACM, New York, NY, USA, 441–450.

[24] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *Proceedings of the 19th international conference on World wide web.* ACM, 591–600.

[25] Eric Leclerc. 2012. Le cyberespace de la diaspora indienne. (2012).

[26] Seung-Jin Lim and Yiu-Kai Ng. 2001. An automated change-detection algorithm for HTML documents based on semantic hierarchies. In *Data Engineering, 2001. Proceedings. 17th International Conference on.* IEEE, 303–312.

[27] Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, and others. 2011. The Arab Spring| the revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International journal of communication* 5 (2011), 31.

[28] Sabrina Marchandise. 2014. Le Facebook des étudiants marocains. Territoire relationnel et territoire des possibles. *Revue européenne des migrations internationales* 30, 3-4 (2014).

[29] Nathan Marz and James Warren. 2015. *Big Data: Principles and best practices of scalable realtime data systems.* Manning Publications Co.

[30] J. Masanès. 2006. *Web Archiving.* Springer, New York.

[31] Eleni Michailidou, Simon Harper, and Sean Bechhofer. 2008. Visual Complexity and Aesthetic Perception of Web Pages. In *Proceedings of the 26th Annual ACM International Conference on Design of Communication (SIGDOC '08).* ACM, New York, NY, USA, 215–224. DOI:http://dx.doi.org/10.1145/1456536.1456581

[32] Mihaela Nedelcu. 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* (2003), 325–339.

[33] Marilena Oita and Pierre Senellart. 2010. Archiving data objects using Web feeds. In *International Workshop on Web Archiving.*

[34] Marilena Oita and Pierre Senellart. 2015. FOREST: Focused object retrieval by exploiting significant tag paths. In *Proceedings of the 18th International Workshop on Web and Databases.* ACM, 55–61.

[35] Liladhar R. Pendse. 2016. Collecting and preserving the Ukraine conflict (2014-2015): A web archive at university of California, Berkeley. *Collection Building* 35, 3 (2016), 64–72.

[36] Radu Pop, Gabriel Vasile, and Julien Masanes. Archiving web video.

[37] Thomas Risse, Elena Demidova, Stefan Dietze, Wim Peters, Nikolaos Papailiou, Katerina Doka, Yannis Stavrakas, Vassilis Plachouras, Pierre Senellart, Florent Carpentier, Amin Mantrach, Bogdan Cautis, Patrick Siehndel, and Dimitris Spiliotopoulos. 2014. The ARCOMEM Architecture for Social- and Semantic-Driven Web Archiving. *Future Internet* 6, 4 (2014), 688.

[38] Jean-Marc Salmon. 2016. *29 jours de révolution. Histoire du soulèvement tunisien, 17 décembre 2010 - 14 janvier 2011.* Les Petits matins.

[39] Valérie Schafer and Benjamin G. Thierry. 2016. The "Web of pros" in the 1990s: The professional acclimation of the World Wide Web in France. *New Media & Society* 18, 7 (2016), 1143–1158.

[40] Claire Scopsi. 2009. Les sites web diasporiques : un nouveau genre médiatique ? *tic&société [En ligne]* 3, 1-2 (2009).

[41] John Scott. 2017. *Social network analysis.* Sage.

[42] Marc Spaniol, Dimitar Denev, Arturas Mazeika, Gerhard Weikum, and Pierre Senellart. 2009. Data Quality in Web Archiving. In *Proceedings of the 3rd Workshop on Information Credibility on the Web (WICOW '09).* 19–26.

[43] Marc Spaniol, Arturas Mazeika, Dimitar Denev, and Gerhard Weikum. 2009. "Catch me if you can": Visual Analysis of Coherence Defects in Web Archiving. In *The 9 th International Web Archiving Workshop (IWAW 2009) Corfu, Greece, September/October, 2009 Workshop Proceedings.* 1.

[44] Masashi Toyoda and Masaru Kitsuregawa. 2003. Extracting Evolution of Web Communities from a Series of Web Archives. In *Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia (HYPERTEXT '03).* 28–37.

[45] Masashi Toyoda and Masaru Kitsuregawa. 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 (HYPERTEXT '05).* 151–160.

[46] John W. Tukey. 1977. *Exploratory Data Analysis.* Addison-Wesley.

[47] John W. Tukey. 1986. *The Collected Works of John W. Tukey: Philosophy and Principles of Data Analysis 1965-1986.* Number vol. 4. Taylor & Francis.

[48] UNESCO. 2003. Charter on the Preservation of Digital Heritage. (2003). http://portal.unesco.org/en/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html

[49] Stanley Wasserman and Katherine Faust. 1994. *Social network analysis: Methods and applications.* Vol. 8. Cambridge university press.

[50] G. Weikum, N. Ntarmos, M. Spaniol, P. Triantafillou, A. Benczur, S. Kirkpatrick, J. Masanes, and M. Williamson. 2011. Longitudinal analytics on web archive data: it's about time! (2011).

[51] Tim Weninger and William H Hsu. 2008. Text extraction from the web via text-to-tag ratio. In *Database and Expert Systems Application, 2008. DEXA'08. 19th International Workshop on.* IEEE, 23–28.

[52] Yann Scioldo Zürcher. 2012. Mémoires et pressions sur la toile? Étude des Français rapatriés coloniaux de la seconde moitié du vingtième siècle à nos jours. *Mémoires et pressions sur la toile* 2 (2012), 34.