Where the dead blogs are. A Disaggregated Exploration of Web Archives to Reveal Extinct Online Collectives

Quentin Lobbé

Where the dead blogs are

A Disaggregated Exploration of Web Archives to Reveal Extinct Online Collectives

Quentin Lobbé[0000−0003−2691−5615]

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, digital heritage, online migrant collectives

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. The Web then became an environment favourable to the establishment of online hubs for migrants to connect with each other or to preserve pieces of a scattered collective memory. In 2012, the e-Diasporas Atlas [7], directed by D. Diminescu, 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*[1]. 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 to ensure the preservation of their digital history and to allow forthcoming researche.

Web archiving. Since the creation of the Web in the early 90's [5], 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 [9], Web archiving was gradually assumed by various states. Shortly after the recognition of the *Charter on the Preservation of the Digital Heritage* by UNESCO [18], terabytes of Web pages were saved worldwide by archiving the genesis of the Web. But after 20 years of Web archiving, it must be said that there is an asymmetry between works focused on upstream archive acquisition [13] and analysis of existing Web archives corpora [17]. In practice, most national libraries allow limited consultation points with no remote access to their archived corpora. The online portal of the WayBack Machine[2] only provides a restrictive search-by-URL system without any full-text search facility. Existing tools are designed and effective for refining past versions of a known URL, not for proceeding to a large-scale exploration. Thus, related research based on Web archives chooses to manually track a set of URLs [16] or to focus on the visual aspects of an archived Web page [1]. As a new insight, N. Brügger introduces the notion of *analytical Web strata* [3]. He then suggests the possibility of building a dynamic system to resize historical studies from an archived Web page to its individual Web elements.

Summary of main contributions. In this paper, we propose a framework based on the exploration of corpora of Web archives, to follow the internal dynamics of extinct online migrant collectives: communities for 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 to search for correlations between a given political and social context and the topographic evolutions of a vanished community. We will study two online migrant collectives extracted from the Moroccan section of the e-Diasporas Atlas archives[3]: an established blogosphere (Section 4) and an ephemeral group of forum members (Section 5). They both vanished from the Web at some point before 2018. The Moroccan *e-Diaspora* is a network of 254 Web sites, built on hypertext citations, created or managed by Moroccan migrants or that deal with them, and initially mapped in 2008 (Figure 1). It can be divided in 3 clusters: 1) Institutional Web sites managed by the Moroccan government, 2) Associations and NGOs, 3) The blogosphere edited by citizens. The forum *yabiladi.com* can be seen as a hub between the 3 clusters, *larbi.org* is depicted as one of the leading blogs. The whole network was then weekly archived from March 2010 to September 2014: we count 2,683,928 archived pages for *yabiladi.com* and 78,311 for *larbi.org*.

[1] http://www.e-diasporas.fr/
[2] https://archive.org/web/
[3] Available at http://www.e-diasporas.fr/wp/moroccan.html

We start by introducing, in Section 2, an entity called *Web fragment*: a new unit to explore archived corpora that results from the segmentation of a Web page. We then discuss the benefits of using it in comparison to existing approaches. We show, in Section 3, the way we can extract and shape Web fragments out of archived Web pages. We then depict the technical implementation of the whole framework. In Section 4, we witness the collapsing of the Moroccan blogosphere. We show its progressive mutation towards rising social platforms, between 2008 and 2018. In Section 5, we study the creation of an ephemeral collective of members of the forum *yabiladi.com* gathered by the wave of Arab Spring[4]. In particular, we show how some old users converged suddenly around the online organisation of the demonstration of February 20th 2011. Finally, we discuss in Section 6 the limitations of Web archives as a source of information and introduce the notion of *pivot moment of the Web*.

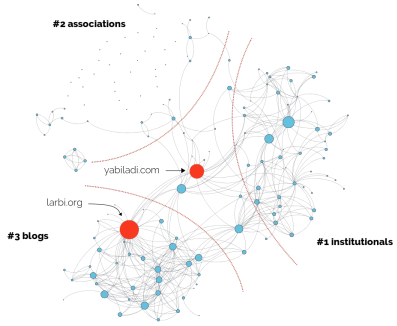

Figure 1: The Moroccan e-Diasporas (mapped by D. Diminescu, M. Renault and M. Jacomy)

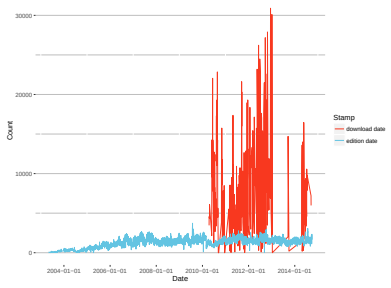

Figure 2: Distribution of the archives of *yabiladi.com* using download dates vs edition dates

2. The Web fragment

In the following, we introduce the Web fragment and discuss the benefits of upscaling the historical analysis of Web archives by using the Web fragment instead of the Web page as new unit of exploration.

2.1 Definition

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 [14] a Web page is the result of the logical arrangement of distinct semantic components, we define a Web fragment as a semantic and syntactic subset of a given Web page. There is a scale relationship between a Web

[4] The Arab Spring was a revolutionary wave of protests in North Africa and the Middle East between 2010 and 2012 (https://en.wikipedia.org/wiki/Arab_Spring)

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.).

2.2 Upscaling the exploration

Assumptions. Archive file formats[5] are basically a collection of crawled HTML pages associated with a download date. In existing Web archive explorer tools (such as the Wayback Machine), the results are stamped by download date. By contrast, a Web fragment is related to an edition date: the date when it was created or published on the living Web. The many difficulties to retrieve edition dates have allready been addressed by [8], but the benefits in term of historical accuracy are impressive. From this point, we will anticipate the possibility of extracting Web fragments (see Section 3 for a technical implementation).

Reducing crawl blindness. We call here *crawl blindness* the action of mistimestamping a change on a page after a crawl. A change can be the creation, the update, or the deletion of all or part of a Web page [6]. As a proposition, we first call the process of downloading the Web pages $\{p_1,...,p_n\}$ of an entire Web site a crawl c_i. We then assume that a corpus of Web archives is the result of one or several successive crawls $\{c_1,...,c_l\}$. An archived Web site consists of n Web pages numbered $\{p_1,...,p_n\}$. 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. We assume to know the last modified stamp of page p_j denoted $\mu_i(p_j)$ during crawl c_i (having $\mu_i(p_j) \leq t_i(p_j)$). Here, we argue that a page p_j consists of m Web fragments numbered $\{f_{j1}, ..., f_{jm}\}$. We also assume to know the edition date of each Web fragment $\phi(f_{j1}), ..., \phi(f_{jm})$, where c_i is a crawl in which f_{jk} exists, so that an edition date will always be more historically accurate than a download date:

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

Increasing the historical accuracy. We now assume that the first download date of page p_j is denoted as $\min_i t_i(p_j)$ and we approximate its creation date by its first edition date $\min_k \phi(f_{jk})$. As an experiment, we select the 109,534 archived Web pages of the forum section of *yabiladi.com* stamped by first download date. We split them into 422,906 deduplicated Web fragments stamped by edition dates. To be more specific, the remaining Web fragments can be seen as single archived forum messages associated with a publication date. In Figure 2, we compare the temporal distribution of the archived pages stamped by first download dates (red line) versus their corresponding first edition dates (blue line). First

[5] WARC (Web ARChive) or DAFF (Digital Archive File Format) file formats

of all, we witness a crawl blindness around 2013 (red line): the crawler stopped archiving during many months. This crawl legacy can be attenuated by switching to the edition dates (blue line). Then, as archived pages keep the traces of past messages, we can extend the comprehension of our corpus (archived since 2010) to consider contents written up to 2003 (blue line). If we calculate the difference $\min_i t_i(p_j) - \min_k \phi(f_{jk})$ in days, the corresponding quartiles are: Q1) 256, Q2) 777, Q3) 1340. With our framework, doing an exploration on top of Web fragments stamped by edition dates will always be more historically accurate than looking at the original Web pages stamped by download dates.

3. Disaggregating web archives

In the following, we will move to the practical implementation of our framework and discuss our method for extracting and framing Web fragments.

Implementation. Our architecture is released under an open-source license[6] and follows a classical implementation model[7]: 1) archives files are grabed by a Java extractor and then uploaded into a Hadoop Distributed File System (HDFS). 2) A Spark pipeline ingests the HDFS and filters the archives. 3) A dedicated library extracts the Web fragments out of the archived pages. 4) Then the text content of each Web fragment is indexed into a Solr search engine.

Extraction of Web fragments. Here, we consider an archived Web page 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 t using the boilerplate method of [11] to filter out ads and navigation nodes. We then follow user-centric scraping strategies[8]: Mozilla's Readability and Fathom 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 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 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. As a contribution, we extend the distance function d of Readability with visual-based penalties introduced by [4] and tag-based penalties introduced by [8] to handle the *"human perception"* part of the Web fragment. In practice, we initialize the variable $minDist$ for

[6] https://github.com/lobbeque/archive-miner and https://github.com/lobbeque/archive-search

[7] Using Hadoop (http://hadoop.apache.org/), Spark (https://spark.apache.org/) and Solr (http://lucene.apache.org/solr/)

[8] https://github.com/mozilla/readability and https://github.com/mozilla/fathom

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[9]. To sum up, a Web fragment is a coherent cluster of HTML 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

4. Archived traces of digital mutation

We now transition to the question of extinct online migrant collectives. In particular, we focus on analysing the collapsing of an old-established community of migrant Moroccan blogs between 2008 and 2018.

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 3 (close up, left). The political blog *larbi.org* possessed the highest in-degree[10] and occupied a central position in the community. The 47 blogs used French as a main language and produced a bundle of political thoughts and daily moods. In 2015, a report [10] induced that many of the blogs were no longer active. 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 3, right).

Exploration task. We emit two hypotheses: 1) the authors of the blogs deleted them and moved away from the Web, 2) the authors migrated from one Web territory to another (such as Twitter or Facebook). As we know that Facebook groups where contemporary of this set of blogs [12], we define our task of exploration as finding the past traces of a digital mutation. We choose to target

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

[10] The number of edges incoming to a vertex

the first traces of social media in the whole archives. In our framework, we request for fragments containing HTML nodes related to social networks like **, *<button class="Fb-share"></button>*, or directly mentioning *facebook, youtube, pinterest, etc* inside their textual contents.

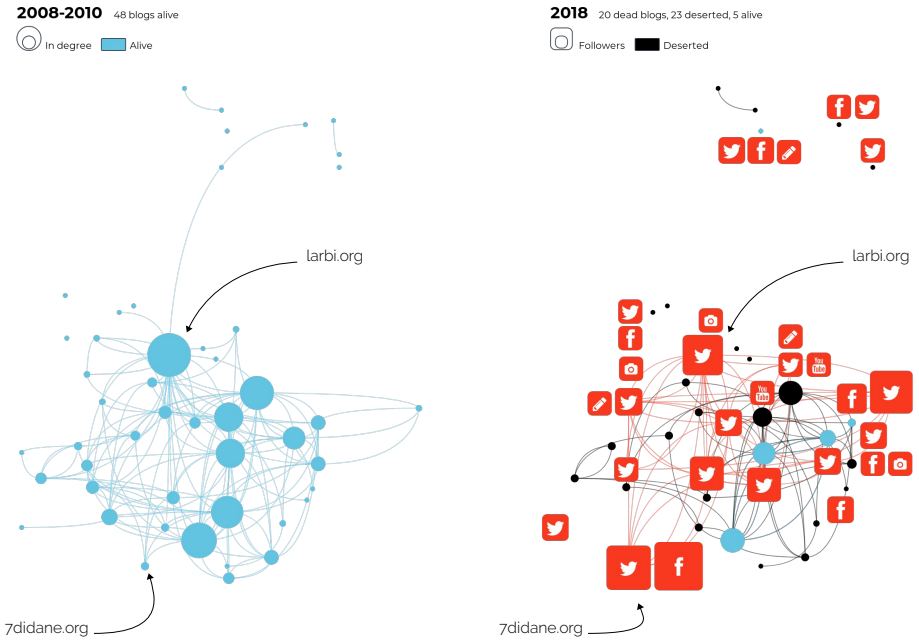

Figure 3: Evolution of the Moroccan blogosphere between 2008 (left) and 2018 (right) with a kept position

The recomposition of the community. The results consist in a filtered set of Web fragments timestamped and grouped by blogs. Some of them contain the URL or the account name of a linked social media. We use the WayBack Machine to visually validate each URL and deliver a qualitative analysis[11]. After managing a blog of their own, 20 authors moved to a social platform: 8 have a Facebook standalone page, 16 are on Twitter (Figure 3). Created between 2005 and 2007, the blogs slowly died around the early 2010's. 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. 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

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

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 3 (right), the size of each 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. 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[12] to extract the country of origin of the followers of each Facebook page. As an example, *lailalalami.com* still speaks to Morrocan (24%), American (15%) and Pakistani folks (8%). 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. 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. 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. They also represented a significant part of the past audience of the blog, as they wrote 26% of all the archived comments.

The Arab Spring as a key moment. Only 6 blogs wrote a clear farewell message before dying. But the author of *7didane.org* indicated that they discovered Twitter by following the 2009 protests in Iran[13]. We also notice that *larbi.org* first publicly mentioned Twitter by the end of 2010 during the Arab Spring and pointed out its use as a tool to organized citizens actions for the upcoming protest of February 20th, 2011 with the hash-tag #20Fev [14]. 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.

5. An ephemeral protest collective

The Arab Spring was in many ways influenced by an active use of social media as a mean for collective organisation [15]. In Morocco, the protests occurred early

[12] https://apps.facebook.com/netvizz/

[13] The message: https://frama.link/DUo84Yhx, Iran protests: https://frama.link/nZmQD_Y1

[14] https://frama.link/-Gd44Pq3

in 2011 and culminated on February 20th, 2011 when over 10.000 Moroccans demonstrated to demand democratic reforms[15]. So, we now emit the hypothesis of a community of forum members of *yabiladi.com* gathered around the events of the February 20th.

A hub for Moroccan migrants. Figure 1 illustrates the key role of *yabiladi.com* as an old-established place in the Moroccan e-Diasporas. Created in the late 2001, the site appeared to spread multi-support informations and to be a meeting places for the migrants living abroad. In 2002, *yabiladi.com* opened a forum section, organised in categories and threads. 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 February 20th. Thus, we request for Web fragments following the given ordered template: 1) a user name 2) a publication date 3) a text content. They also have to be associated to an URL containing the path *"/forum/"*. Then, we can group the fragments by category and thread, as their pattern of URL follows *"/forum/**thread_ title-category_ id-thread_ id**.html"*. We assume the remaining Web fragments to be only composed of archived forum posts. We restrict our space of exploration to the two categories: *General* and *Moroccan and Worldwide Actuality*.

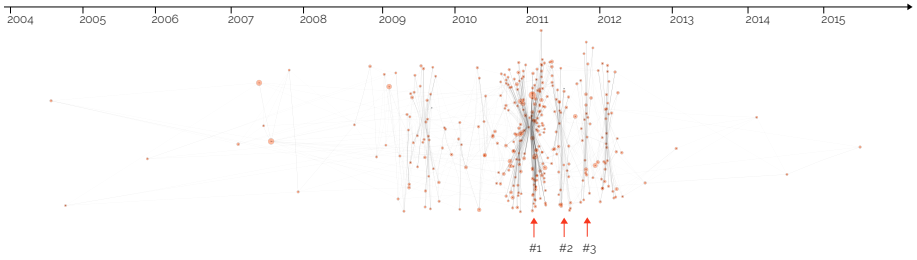

Figure 4: Time distribution of *yabiladi.com*'s threads related to the February 20th

Revealing a collective. We start by querying our system for a set of thread titles matching the French keywords: *"#20Fev"*, *"20 fevrier"*, etc. We manually validate 12 threads 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 select all the threads where at least 2 users of E_0 wrote a message. This new

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

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[16]. With Figure 4, we visualize G on a timeline (in 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. The vertical position of each thread (in ordinate) is a fixed and arbitrary value chosen to clarify the reading of this visualisation. We find in Figure 4 a specific moment (label #1) where the threads of G are very densely distributed. Between January and February 2011, 25% of V_1 were created. This indicates that the protest of February 20th aggregated an old established community of users that were already using the forum. We see that the pre-protest part of G (before label #1) represents 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 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 of G. 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 around the protest of February 20th, we refine our comprehension of G by conducting a clustering analysis out of it using the modularity class method [2]. The 8 resulting clusters of threads[17] can be interpreted as subsequent moments of the evolution of G. 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. 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 Moroccan constitution. 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[18].

[16] Downloadable results V_1: https://frama.link/_eModem_, E_0: https://frama.link/hcxacx89
[17] Downloadable results (as a GEXF graph file) G: https://frama.link/BZdU8CW8
[18] Manually counted and validated in April 2018

6. Implication for historical web studies

The development of our framework was guided by the idea that a Web site should become the object of historical studies [3].But here, we may have reached the limits of Web archive corpora by missing a major aspect of our problematic: Web archives are intrinsically incomplete. Mostly created and designed during the early 2000's [13], Web archiving systems followed the subsequent evolutions of the Web as a medium but still fail to convey the Web as an ecosystem. The living Web is a flow of informations where various actors are organically interrelated. By contrast, 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, Web actors were already moving away from forums and blogs. The problem of extinct online collective, is less a question of disappearance than a question of transition and Web archives corpora only witness the first leap of what we call a *pivot moment of the Web*.

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, etc), the Web already possesses its 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 3 factors: (1) the convergence at a specific time (2) between a technological leap and (3) some users sieving it. This leads the Web in new directions of development such as during the democratization of DSL in the late 1990's, the advent of smartphones and mobile Web in the 2010's or the transition from the Web 2.0 to the Web of social network as illustrated in Section 4 and 5.

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. Ben-David, A., Amram, A., Bekkerman, R.: The colors of the national web: visual data analysis of the historical yugoslav web domain. International Journal on Digital Libraries (19, 1), 95–106 (Mar 2018)

2. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment (2008,10), P10008 (2008)
3. Brügger, N.: Website history and the website as an object of study. New Media & Society (11,1-2), 115–132 (2009)
4. Cai, D., Yu, S., Wen, J.R., Ma, W.Y.: Vips: a vision-based page segmentation algorithm (2003)
5. CERN: The document that officially put the world wide web into the public domain (1993), http://cds.cern.ch/record/1164399
6. Cho, J., Garcia-Molina, H.: The evolution of the web and implications for an incremental crawler. Tech. rep., Stanford (1999)
7. Diminescu, D.: e-Diasporas Atlas. Explorations and Cartography of Diasporas on Digital Networks. Ed. de la Maison des Sciences de l'Homme, Paris (2012)
8. Jatowt, A., Kawai, Y., Tanaka, K.: Detecting age of page content. In: Proceedings of the 9th annual ACM international workshop on Web information and data management. pp. 137–144. ACM (2007)
9. Kahle, B.: Preserving the internet. Scientific American pp. 276, 82–83 (Mar 1997)
10. Khouzaimi, J.: e-diasporas : Réalisation et interprétation du corpus marocain (2015)
11. Kohlschütter, C., Fankhauser, P., Nejdl, W.: Boilerplate detection using shallow text features. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining. pp. 441–450. WSDM '10, ACM, New York, NY, USA (2010)
12. Marchandise, S.: Le facebook des étudiants marocains. territoire relationnel et territoire des possibles. Revue européenne des migrations internationales (30, 3-4) (2014)
13. Masanès, J.: Web Archiving. Springer, New York (2006)
14. Michailidou, E., Harper, S., Bechhofer, S.: Visual complexity and aesthetic perception of web pages. In: Proceedings of the 26th Annual ACM International Conference on Design of Communication. pp. 215–224. SIGDOC '08, ACM, New York, NY, USA (2008)
15. Salmon, J.M.: 29 jours de révolution. Histoire du soulèvement tunisien, 17 décembre 2010 - 14 janvier 2011. Les Petits matins (2016)
16. Schafer, V., Thierry, B.G.: The "web of pros" in the 1990s: The professional acclimation of the world wide web in france. New Media & Society (18, 7), 1143–1158 (2016)
17. Toyoda, M., Kitsuregawa, M.: Extracting evolution of web communities from a series of web archives. In: Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia. pp. 28–37. HYPERTEXT '03 (2003)
18. UNESCO: Charter on the preservation of digital heritage (2003)