# Revealing Historical Events out of Web Archives

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

## ABSTRACT

Corpora of web archives are wide and sparse. As the living web expands, worldwide volumes of web archives constantly increase, making difficult to identify archived webpages relevant to a specific sociological or historical study. We propose an application for detecting historical events out of a web archives corpus and discovering pertinent archived digital contents. We introduce here the usage of a new entity called *Web Fragment* to reduce issues related to corpus quality and consistency, and effectively guide researchers through exploration of web archives. A web fragment is defined as a semantic and syntactic subset of a given webpage and has the particularity to be indexed by its edition date (the time when the web fragment was written) instead of its archiving date (the time when its parent webpage was crawled and saved). Building on top of web fragments, we show how this application can be used to study a large archived Moroccan forum and to understand how this online collective reacted to the Arab Spring at the end of 2010.

**Keywords**: web archives, information extraction, event detection, online migrant collectives

## 1. INTRODUCTION

Since the creation of the web at the CERN in the early 90's [6], 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 [12], web archiving was gradually assumed by various states, until the worldwide recognition of the *Charter on the Preservation of the Digital Heritage* by UNESCO [22] who gave to many national libraries the prerogative to archive their web domains (such as .uk, .jp or .fr). Thereby, terabytes of webpages were saved worldwide by archiving the genesis of the web.

While related works mainly focus on upstream web archive acquisition [16], we choose in this paper to perform the exploration of an existing corpus, taking into account its redundancies and deficiencies but still following the idea that the archived web can constitute a gold mine for all kind of sociological or historical researches [23]. Thus, looking at the complex ways in which the web evolves over time is essential to understand the interactions between human activities and digital technologies. For example: how, as an environment, can the web be permeable to the effects of shocks and external events like political or social mobilizations?

Unfortunately, whoever wants to go through an exploration of web archives has to first brave diverse accessibility issues. Apart from the online portal of the WayBack Machine[1], the majority of archived web corpora only allows local consultation points, inside public libraries, with no remote access or API. Even more worrying is the fact that researchers often end up alone facing a giant, wild, and sparse archived corpus, without any full-text search facility (we will see, in Section 4, that the WayBack Machine only provides a restrictive search by URL) or strategy to focus their analysis.

Archiving the web is all about losing and destroying the majority of existing webpages by choosing to save only a few representative sets of contents. As the web is in constant evolution, archiving systems investigate webpage change frequency and crawler effective scheduling [7, 9, 15] but these remain a central issue when the crawler responsible of the harvesting becomes blind for any reason and misses many page modifications [2]. This crawl legacy effect can be evaluated using two notions: the quality of a web archive corpus [20] (Is it not too sparse and temporally coherent?) and the consistency among archived webpages [21] (Is the link between those two pages still available even if they were archived many years apart from one another?). Thus, the archived web is not a high-fidelity and trustful clone of the living web.

**Summary of main contributions**. In this paper, we introduce the usage of a new entity called *web fragment* to reduce or avoid crawl legacy effects and effectively guide researchers through an exploration of web archives at retrieval time. A web fragment is a sub-division of an archived webpage with high informational contents and some metadata (such as author or edition date).

We propose an application, called *Web Archive Explorer* (WAE) for detecting historical events out of a web archive corpus. We think that most archive explorers pursue the discovery of events of some sort. Once the events have been revealed, WAE allows the user to access a full-text and facet search facility built on top of the corpus to identify related archived webpages and digital contents.

We use WAE to study a large, thematically, and temporally coherent online collective : the Moroccan e-Diasporas archive [8]. By applying WAE to this corpus, we show that we can fill in many crawler blindness issues, as the edition date of a web fragment is more accurate than the archiving date of its parent webpages. WAE lets us rethink the notion of consistency among archived webpages: what matters is consistency among fragments independently of spurious modifications in irrelevant parts of the parent webpages such as footers, navigation menus, etc. We finally show that we can use the WAE to understand how the online community of the Moroccan forum *yabiladi.com* reacted to the Arab Spring at the end of 2010.

Next, in Section 2, we describe our web archive corpus and introduce the required theoretical elements. In Section 3, we present the main components and architecture of WAE. We then detail, in Section 4, the demonstration scenario that will be presented at the conference. Please note that a guide video for this demonstration paper is available at https://youtu.be/snW4O-usyTM.

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

## 2. SETUP

In the following we first describe our experimental dataset and then introduce some theoretical elements of the methodology used in our demonstration.

**A collective of online migrants**. As input data for our framework we use the Moroccan section of the e-Diasporas Atlas [8] corpus. Pioneer research program in the sociological analysis of online activities associated with migrant populations, the e-Diasporas Atlas project revealed diasporic communities and 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 websites distributed and aggregated along 30 diasporic networks (Moroccan, Tunisian, Egyptian, etc.) called *e-Diasporas*. But facing, month after month, the partial or total disappearance of some of the observed migrant websites, for instance the desertion of the Moroccan blogosphere between 2012 and 2015 [13], it was decided to start archiving them weekly or monthly to ensure the preservation of this digital history and to allow forthcoming researches. As a corpus of web archives, the e-Diasporas Atlas has the specificity to be highly thematically and temporally coherent as it has been designed on purpose. In our scenario (see Section 4), we will focus on the Moroccan forum *yabiladi.com*: an established and hybrid website (a mix between news, forum, and classifieds), considered as one of the main hubs of the network[2]. The Moroccan corpus was archived from March 2010 to September 2014, covering 254 websites. Yabiladi, on its own, represents a set of 2.8 million archived webpages.

**Web fragments**. Archive file formats are designed on top of webpages. They are basically a collection of crawled HTML pages associated with crawling dates. Our WAE application goes beyond the level of archived webpages and expands the concept of strata of the web [4]. There is wide research covering the extraction of individual components out of webpages [1, 5, 11, 14, 18, 24] mostly designed for automatic large scale processes like natural language analysis. But little is done about defining an individual web object that can be equally useful and understood by computer scientists, sociologists, or historians. This is a key issue and we want here to introduce an entity that deals with a web content and catches the way it has been written and published online :

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 [3, 17] a webpage is the result of the logical arrangement of distinct semantic components, we define a web fragment[3] as a semantic and syntactic subset of a given webpage. There is a scale relationship between a webpage and its web fragments. A web fragment is a coherent set of textual, visual, audio or animated content extracted from a web page that can be understood on its own. Within the same webpage, 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.).

As an example: a web fragment can be a meaningful object like a post inside a forum, a news article, or a comment. To extract a set of web fragments from a given web page, we extend the boilerplate detection method from [14] to clean up the webpage and then use a combination of vision-based [5] and tag-based extraction methods [11]. The web fragment object can be used in various research contexts, but for the purpose of this demonstration, we will use its *edition date* meta property to detect historical events out of archived webpages.

**Event detection model**. Yabiladi has the particularity to be both a news website and a popular forum for the Moroccan diaspora. In Section 3, will see how we extract web fragments out of it. But the bulk of this demonstration is to see how we can use a historical event detection system in order to effectively guide researchers through the web archives. Dealing with events means detecting, identifying and explaining those events. Following our logic of data exploration, we don't want to target specific events with expert knowledge, so we avoid patterns and clustering methods. We instead use a threshold-based heuristic [10] within a sliding time frame of one week. In this paper, we choose to have a fixed time window, but in future work we will give the user the possibility of selecting its own time frame or we will try to automatically suggest an adaptated one. We define an event as a detected outlier in a distribution of web fragments. Our engine divides the web fragments into sets of bi-grams. The event is then identififfied as a burst of bi-grams matching a given set of keywords. Finally, we try explaining the event by finding semantic correlation between the resulting bi-grams and a set of Morrocan news titles. As *yabiladi.com* is a combination between a news provider and a forum, we choose to divide its web fragments in two indexes: 1) extracted from the news section 2) extracted from forum section. Thus, we construct an index of potential events using the *title* and *edition date* meta properties of the web fragments index. According to our model, a historical event is the semantic encounter between a well-dated news title and a burst of forum posts.

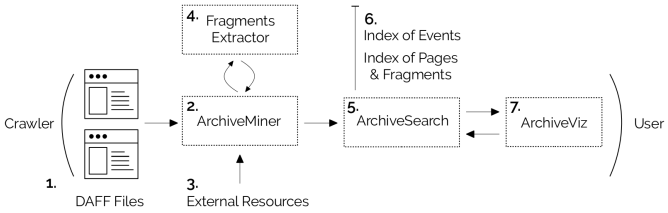


**Fig. 1: Architecture of WAE**

## 3. ARCHITECTURE

We now introduce the different components of WAE (Web Archive Explorer). We refer to Figure 1 as an illustration of the architecture of the application, with items of the following list corresponding to the numbers in the figure. WAE is a modular application where

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

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

each item can be used independently. They are all released under an open-source license and publicly available:

(1) WAE relies on a corpus of web archives. Our data set is recorded under the DAFF[4] formalism : as opposed to the more popular WARC[5] archives file format, DAFF has the particularity to separate the metadata contents (URL, download date, MIME type, charset, etc.) from the data contents (original HTML content). For example, our Moroccan corpus results in a 30GB metadata DAFF file and a 300GB data DAFF file.

(2) The ArchiveMiner[6] component grabs the two DAFF files using a pure Java extractor which uploads the data and the metadata contents into a more convenient Hadoop Distributed File System (HDFS[7]). Then a distributed Spark[8] pipeline ingests the HDFS. There, the ArchiveMiner component groups the metadata contents by time-stable versions of archived pages (unchanged ones) and joins them to the data contents. One can apply here a set of filters based on specific ranges of crawling dates or domain names. For example, the resulting structure is filtered by URL to select only webpages belonging to a given website.

(3) The ArchiveMiner component also enriches the original corpus of web archives by adding qualitative informations obtained from the outcomes of the e-Diasporas Atlas project such as the main language (French, Moroccan, Spanish, etc.), possible location (Paris, New York, Montreal, etc) or category (forum, blog, media, etc.) of each website.

(4) After having cleaned and filtered the archived contents, WAE sends them to the FragmentsExtractor[9] component. There, each archived webpage is divided into web fragments following the definition given in Section 2. Every edition date is translated from a natural language format into a normalized date format. Additionally, the component extracts fields from the header of the webpages such as its description, its publisher name, etc. As a result, a web fragment is basically a JSON object containing some text contents, meta contents (author, title, edition date, etc.), offset of the fragment inside the original webpage, ratio of this webpage represented by the fragment, etc. At the end of this process, we obtain a set of web fragments joined with the information inherited from their parent webpages like the URL or the archiving date. WAE can now access enriched contents effectively searchable in time.

(5) The ArchiveSearch[10] component indexes the web fragments into a Solr[11] search engine in order to make them searchable using the full-text properties and the handy query language of Lucene[12]. A natural language lemmatizer is then applied to increase the accuracy of the search engine. Custom requestHandlers[13] are built in order to provide different time query strategies for requesting web archives. For instance, regarding a given date in a request, one can choose to select the closest forthcoming or past archived version of a webpage.

(6) The ArchiveSearch component provides two different inverted indexes as introduced in Section 2: an index of web fragments extracted from the forum section of yabiladi.com and an index of events extracted from its news section.

(7) The ArchiveViz[14] component provides to the user an interface to request the archives. One can write a set of keywords and choose the granularity of the request : webpages or web fragments. The results are displayed as a list of documents, illustrated with histograms and an bi-grams viewer linked to the events detection system. We will describe ArchiveViz in more detail in Section 4.

## 4. DEMONSTRATION SCENARIO

We now describe a set of use cases where WAE helps to effectively reveal historical events out of archived webpages and identify pertinent sets of documents. We will showcase some potentially temporal and atemporal queries, observe the difference between webpages and web fragment time accuracy, and finally focus on queries related to the Arab Spring. We know from [19] that the web performed a predominant part in spreading the waves of 2010-2011's Arab revolutions. We aim, through this demonstration, to underline how the forum of yabiladi reacted to these historical events. As Morroco is a monarchy, the king *Mohammed VI* may also trigger interesting conversations. The keywords are written in French because many users of yabiladi.com are bilingual. See the accompanying video (https://youtu.be/snW4O-usyTM) and Figure 2 for a peek at the GUI.

(1) For comparison purpose, the user first tries to query the Wayback Machine. But, as it is built on top of a search-by-URL system, the keywords *morroco king* do not match the real content of the archived webpage : they can only match a strict URL or the HTML title of a webpage. So the result is not relevant to our investigation.

(2) Our system allows the user to perform a full-text search on top of the archived pages. So, the user enters in the search box *roi* (meaning king) and selects *pages* for granularity. The user has to pick up a range of dates to filter the archives. The top ten resulting archived webpages are ordered by default Lucene similarity, the results can be reordered by ascendant or descendant download date.

(3) The user now specializes his query by focusing on one of the main author contributions. The 5 most prolific authors, extracted from the web fragments, are displayed in the facets section of the GUI.

(4) The user can use the first histogram on top of ArchiveViz to see the number of matching webpages by week. Here, one can zoom in or out and choose a more focused time range using the brush selector. By zooming in or out, the ArchiveViz will automatically rescale the subservient charts and request the ArchiveSearch component for the next results. Using this count-by-week histogram, the user can see that the archives have started in 2010. Below, there is a line chart displaying the ratio of matching bi-grams by weeks. There the user can follow the evolution of the word *king* in the corpus. The event detection system does not find any matching event because the user chose to use webpages as a scale and pages are timestamped by download date without regard for any historical correlation.

[4]Digital Archive File Format
[5]ISO 28500:2009 – WARC file format
[6]Open source and available at https://github.com/lobbeque/archive-miner
[7]http://hadoop.apache.org/
[8]http://spark.apache.org/
[9]Open source and available at https://github.com/lobbeque/rivelaine
[10]Open source and available at https://github.com/lobbeque/archive-search
[11]http://lucene.apache.org/solr/
[12]http://lucene.apache.org/index.html
[13]Open source and available at https://github.com/lobbeque/archive-search-tools
[14]Open source and available at https://github.com/lobbeque/peastee

(5) Now the user switches to the web fragment level and enters the same query. ArchiveViz displays the top ten web fragments returned by WAE. The user see that she has the possibility to study web fragments written up to 2003. By using web fragments as a scale, the user can access archived pieces of information with a greater time accuracy. The event detection system now understands that around the late 2004 an event concerning the king may have focused the conversations on yabiladi.com. The system identifies it as an official visit of the Morrocan king to Mexico in November 2004.

(6) Now the user enters in the search box *Mohammed VI* and *Ben Ali* (the former Tunisian president). The WAE supports multiple queries (using comma as a separator) for comparison purpose. As in Google Trends, it displays a line in the n-gram viewer for each query and a union of the resulting fragments in the list below. The colors, used in the n-gram viewer chart, also help the user detect the matching words inside each resulting web fragments. The user can clearly see a growing percentage of the phrase *Ben Ali* in the n-gram viewer during the late 2010 that we may correlate to the beginning of the Arab Spring. This assumption is reinforced by a triggered event about the destitution of Ben Ali in January 2011.

(7) Finally, the user enters a query less specific such as *allah*. The ArchiveViz helps understand that this word does not have any temporal dimension. It is used equally throughout the whole archived period, in a religious way or as a language feature. More seasonal keywords can by entered in the search box such as the muslim month of fasting *ramadan*. Here the user observes a temporal pattern in the archives that can be explained by the cultural specificity of the Moroccan corpus.

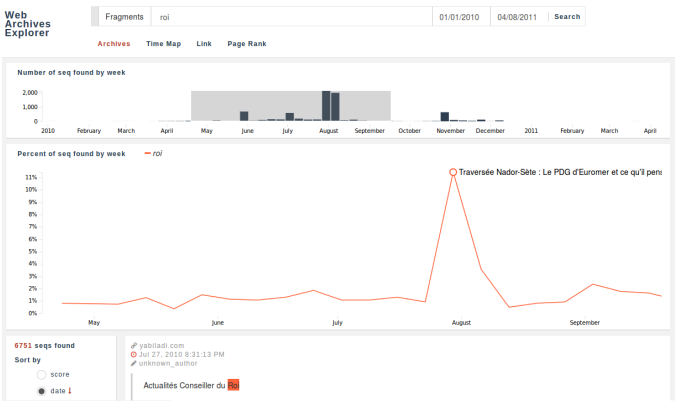


**Fig. 2: WAE interface**

## 5. CONCLUSION

In this paper, we proposed an application to reveal historical events and phenomenons out of web archives. By applying it to the Moroccan archive of the e-Diasporas Atlas, we showed from a computer science perspective that focusing on web fragments instead of webpages is a promising way to improve on crawler quality and consistency issues.

In the future, we will feed WAE with more diverse sets of web archives (social media streams, blogging platforms, etc.) and work in close collaboration with sociologists and historians to investigate multidisciplinary research questions based on web archive analysis.

Finally, upcoming research papers will go deeper on web fragment applications and address problems focused on the quality of the process of building web archives. There are many ways to use web fragments, like following the spread of a given political topic through a relevant network of web fragments or understanding the evolution of the facilities of publication and edition of webpages over time.

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

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

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

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

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

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

[8] 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.

[9] Fred Douglis, Thomas Ball, Yih-Farn Chen, and Eleftherios Koutsofios. 1998. The AT&T Internet Difference Engine: Tracking and viewing changes on the web. *World Wide Web* 1, 1 (1998), 27–44.

[10] Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Philip S Yu, and Hongjun Lu. 2005. Parameter free bursty events detection in text streams. In *Proceedings of the 31st international conference on Very large data bases*. VLDB Endowment, 181–192.

[11] 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.

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

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

[14] 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.

[15] 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.

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

[17] 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

[18] 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.

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

[20] 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.

[21] 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.

[22] 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

[23] 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).

[24] 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.