Introducing Web Fragments

An exploration of web archives beyond the webpages

Quentin Lobbé (LTCI, Télécom ParisTech & Inria Paris)

Medialab's research seminar – October 17, 2017

> Let's start with a first leading question :

How can we go through an exploration of web archives over time ?

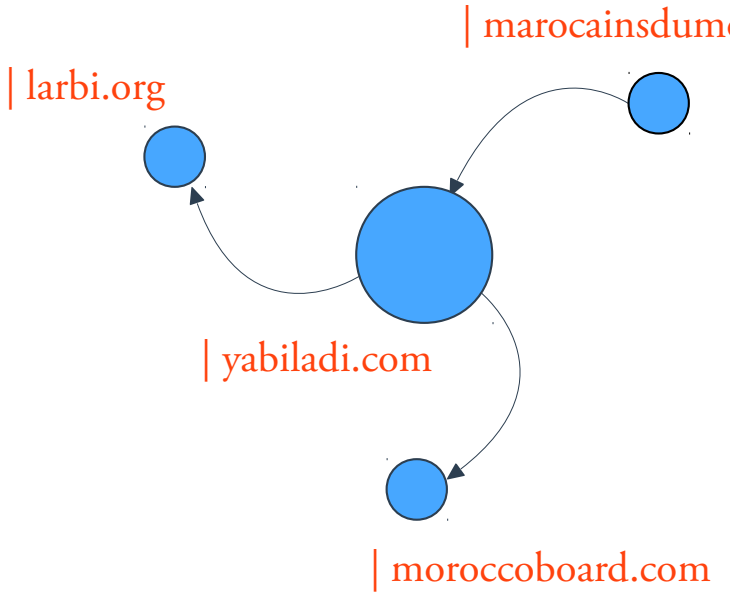
> To understand if and how :

The structure and content of the web can be permeable to the effects of shocks and external events such as political and social mobilizations?

Summary

The e-Diasporas Atlas

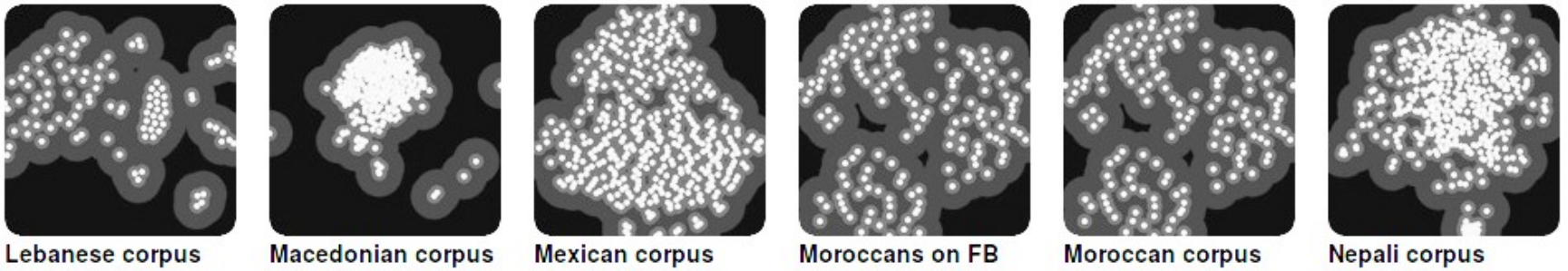
> A collection of online migrant collectives



A migrant web site is a website created or managed by migrants and/or that deals with them

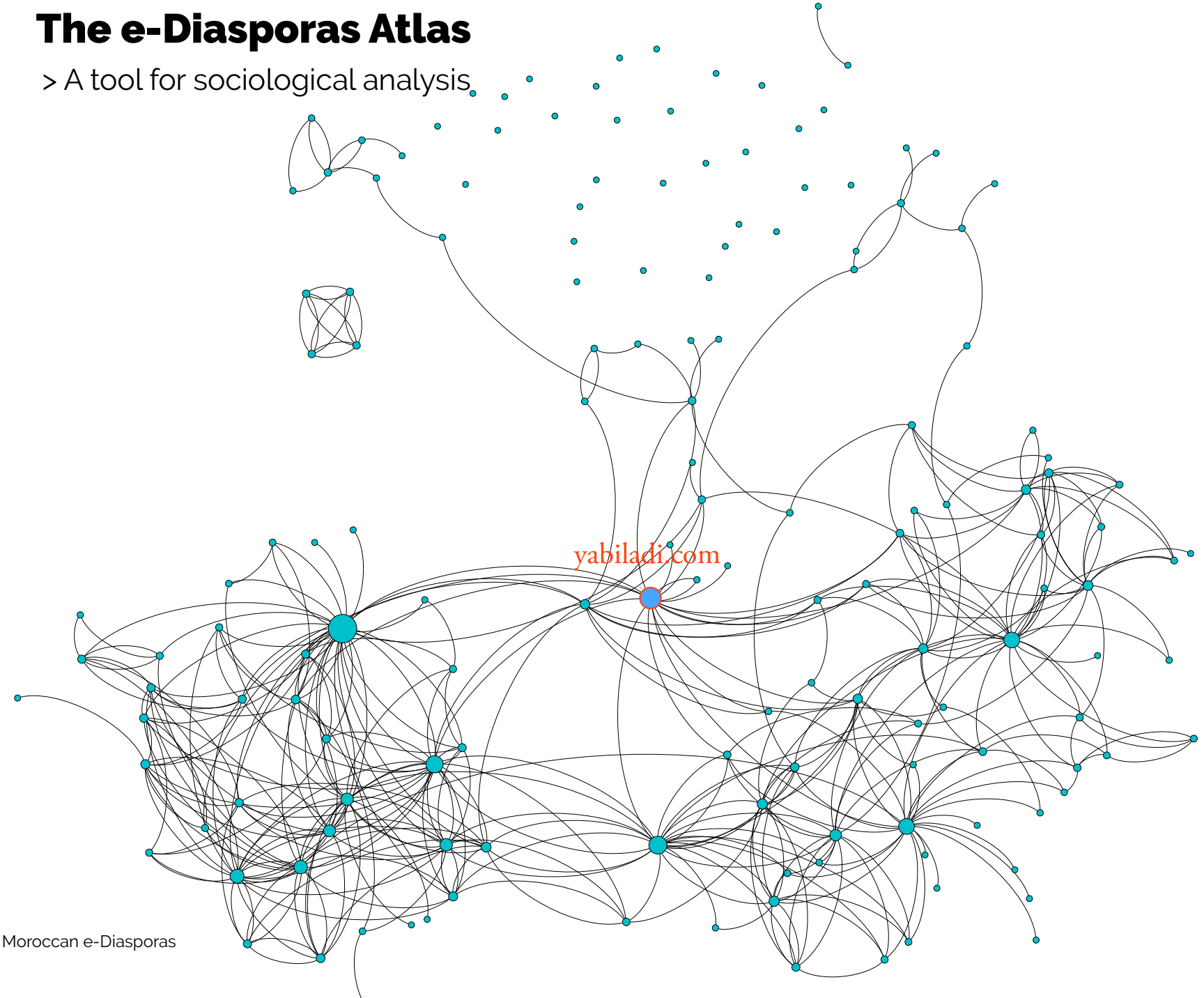
An e-Diaspora is a directed network of migrant websites linked by url (href)

10.000 migrant websites crawled, categorized and organized among 30 e-diasporas



The e-Diasporas Atlas

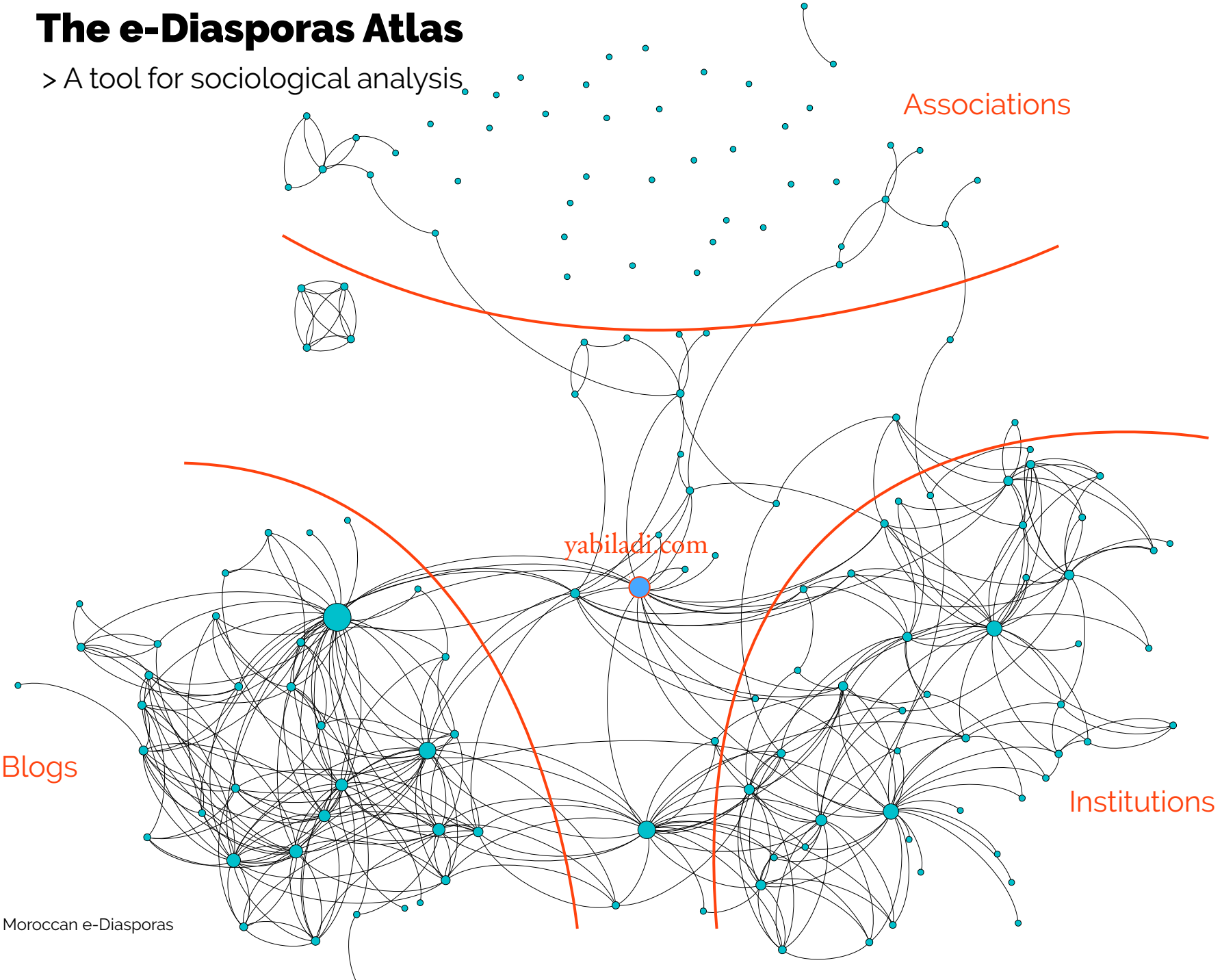
> A tool for sociological analysis



> the Moroccan e-Diasporas

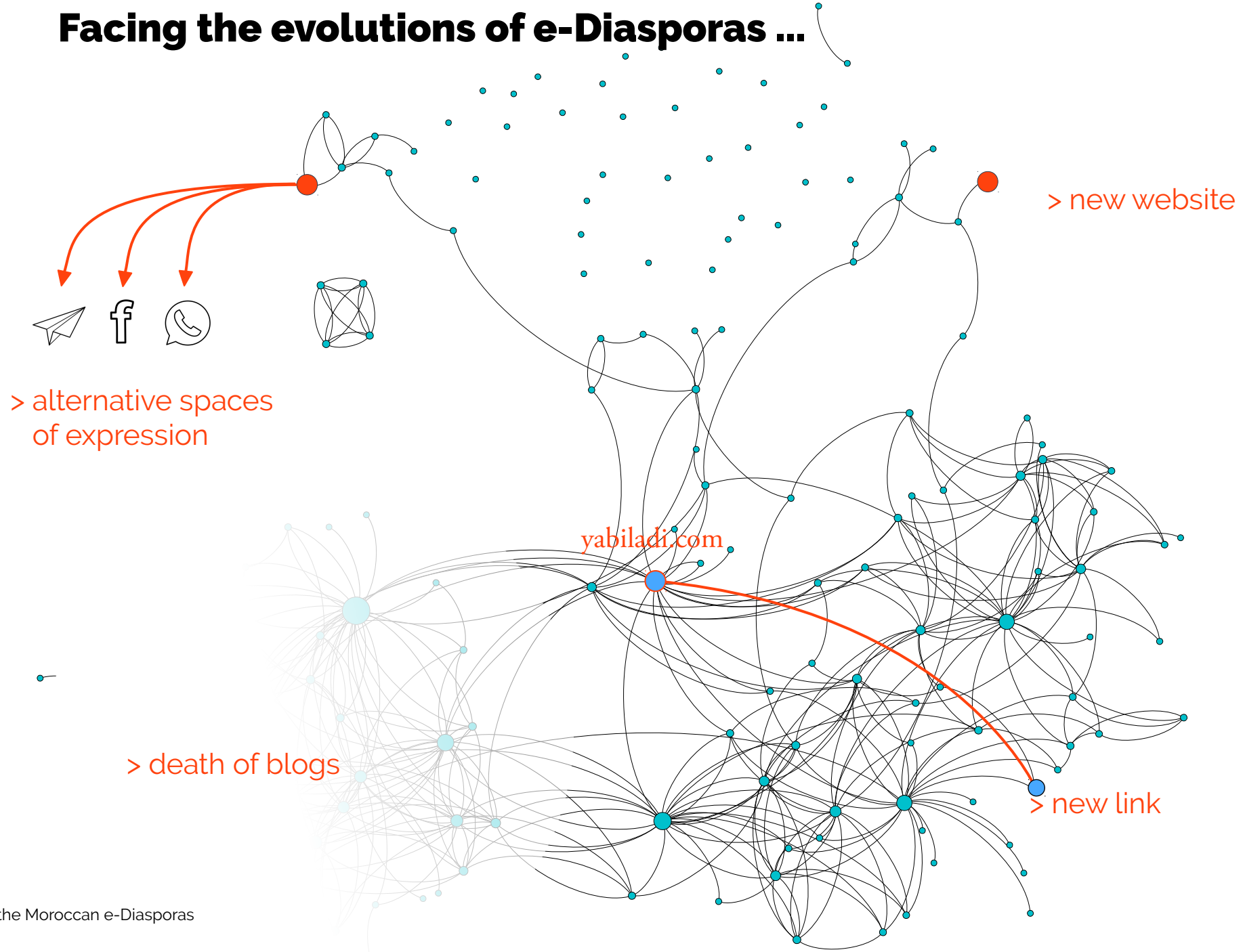
The e-Diasporas Atlas

> A tool for sociological analysis



> the Moroccan e-Diasporas

Facing the evolutions of e-Diasporas ...



… and all kinds of web site changes …

> structural changes
  move, copy, delete, inserte,
  update …

> attribute changes
  css, font …

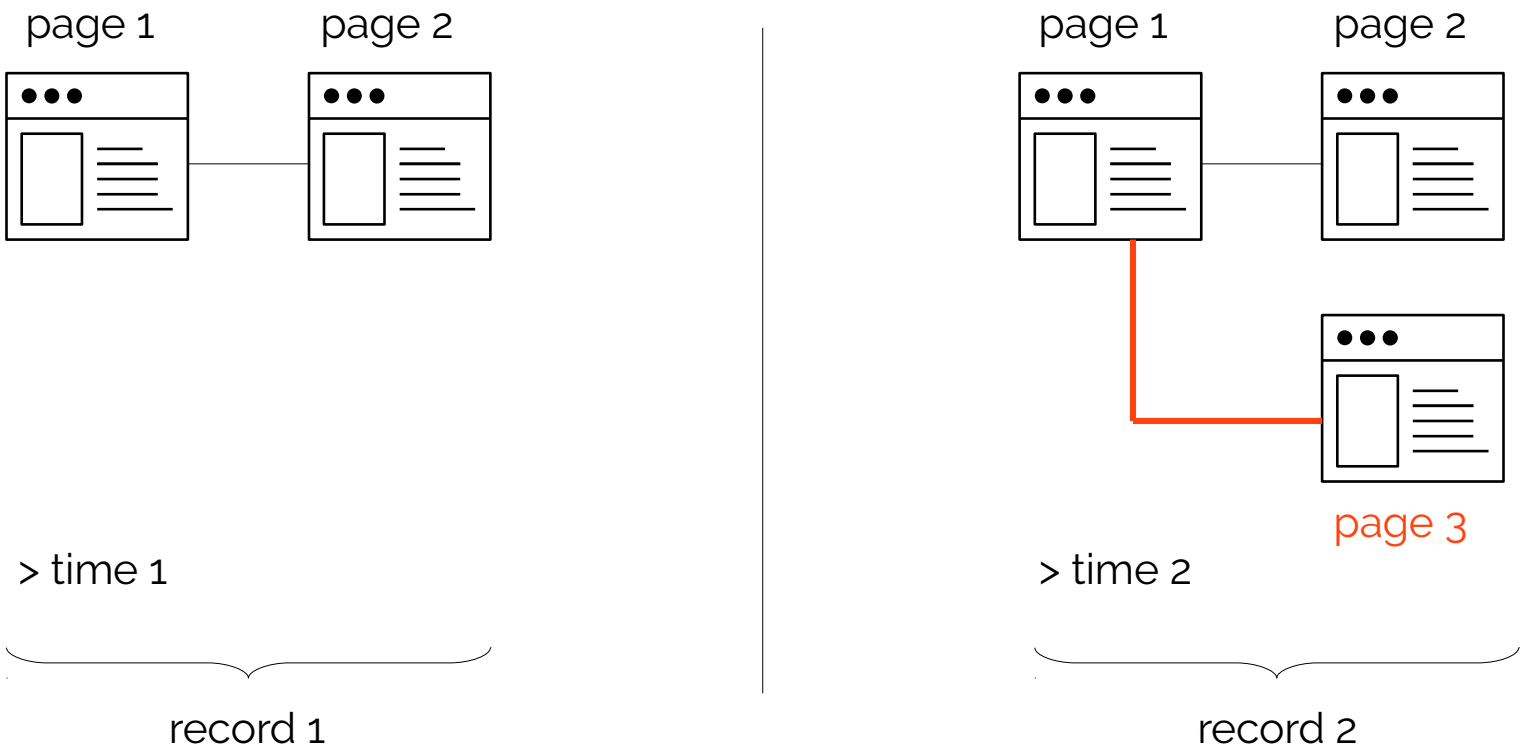
> type changes
   to <p>

> semantic changes

For example : http://www.medialab.sciences-po.fr/fr/

... it was decided to start archiving the corpora

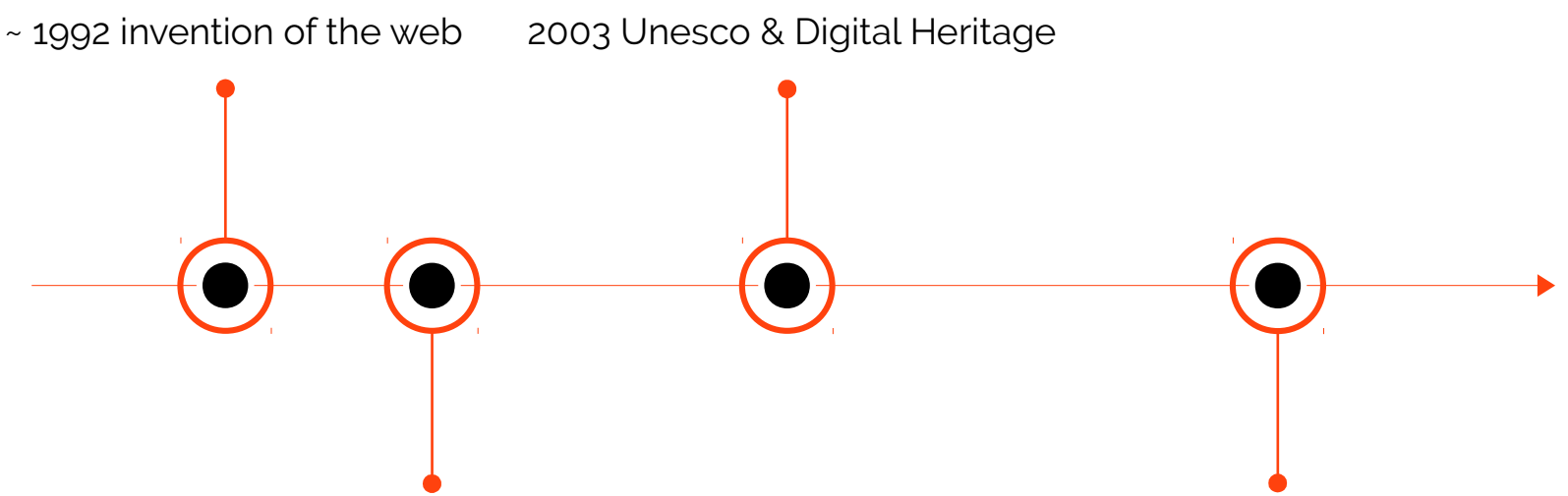
> To keep a trace of the evolutions of every website

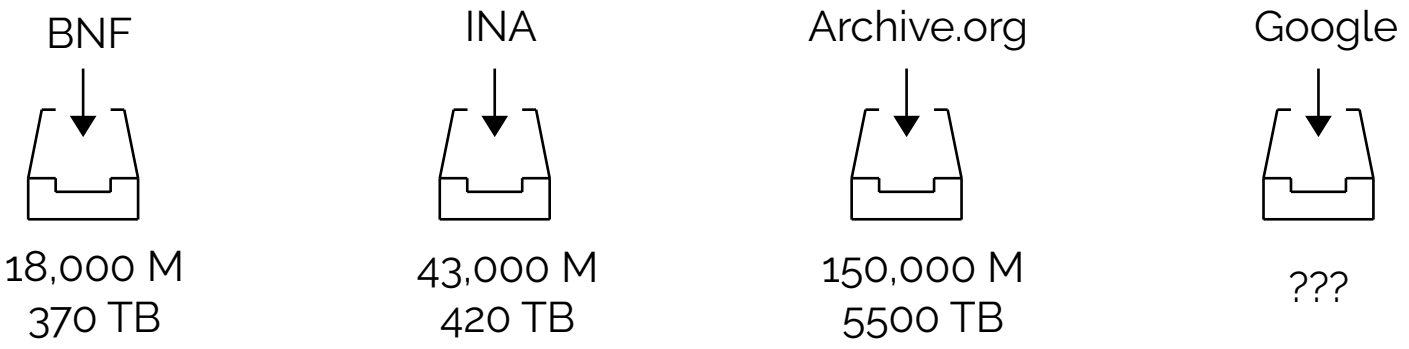


As related works mainly focus on upstream web archive acquisition
we choose here to perform the exploration of an existing corpus

20 years of web archiving

> Saving the micro-history of the web

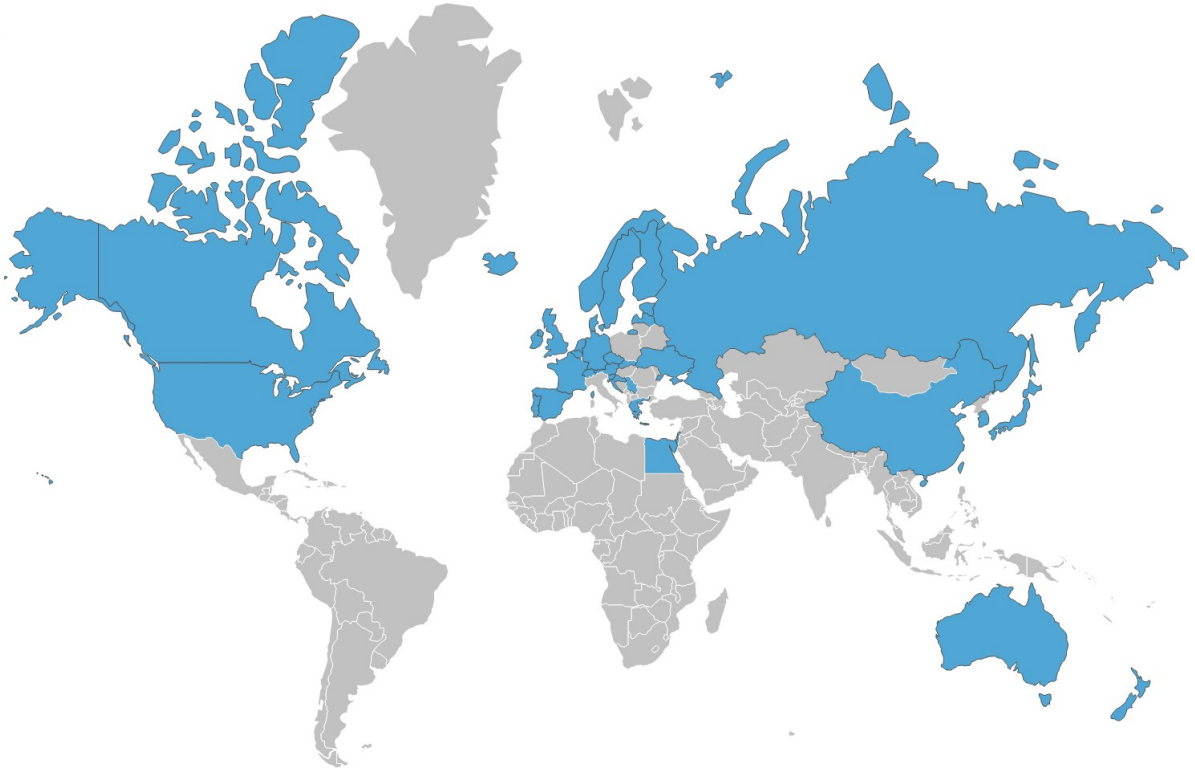




Unfortunately, whoever wants to go through an exploration of web archives has to first brave diverse accessibility issues...

Consulting a corpus of web archives

> Accessibility & issues



> Map of Web archiving initiatives worldwide (2017)

> Online access :

INTERNET ARCHIVE
WayBackMachine

memento

The National Archives

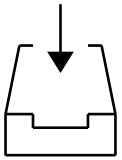
But : - restricted search
- no strategy to focus analysis

> Local access :

{ BnF

A 1h30 long journey to access the local consultation point

The e-Diasporas Atlas Archive



1030 M of webpages
70 TB
Crawled weekly or monthly, from April 2010 to September 2014
Hosted and performed by the INA

> The Moroccan archive

153 websites

From 09/2017 :

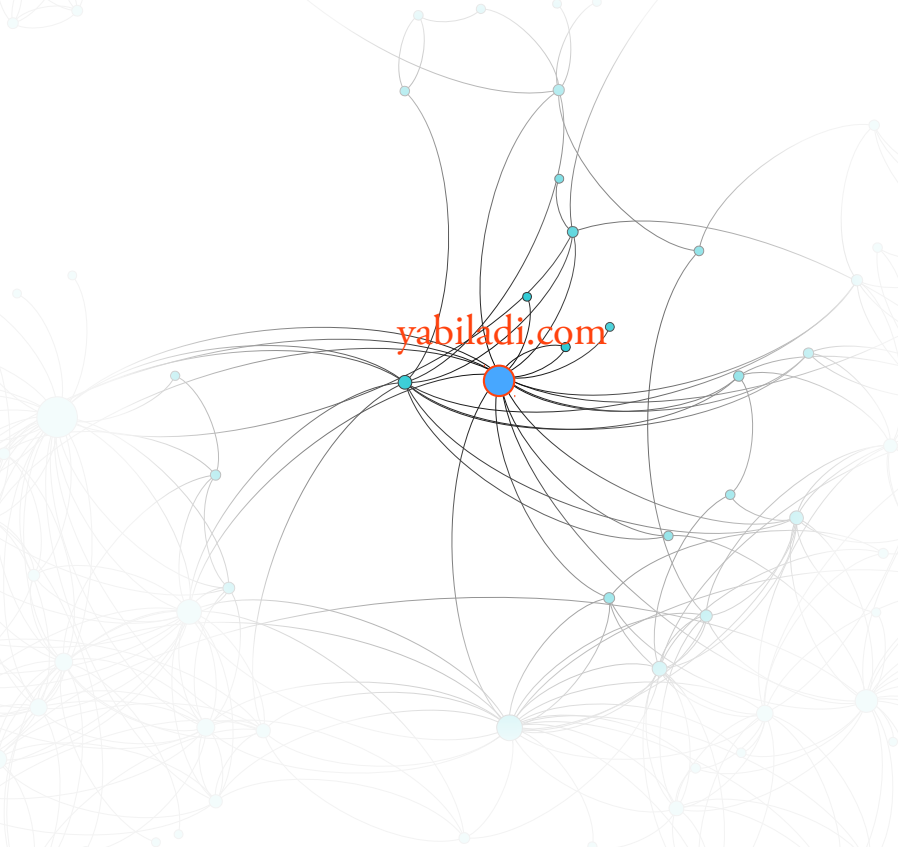
53 still alive > 34,6 %
38 no update > 25,8 %
62 dead > 40,5 %

) blogs are the most impacted

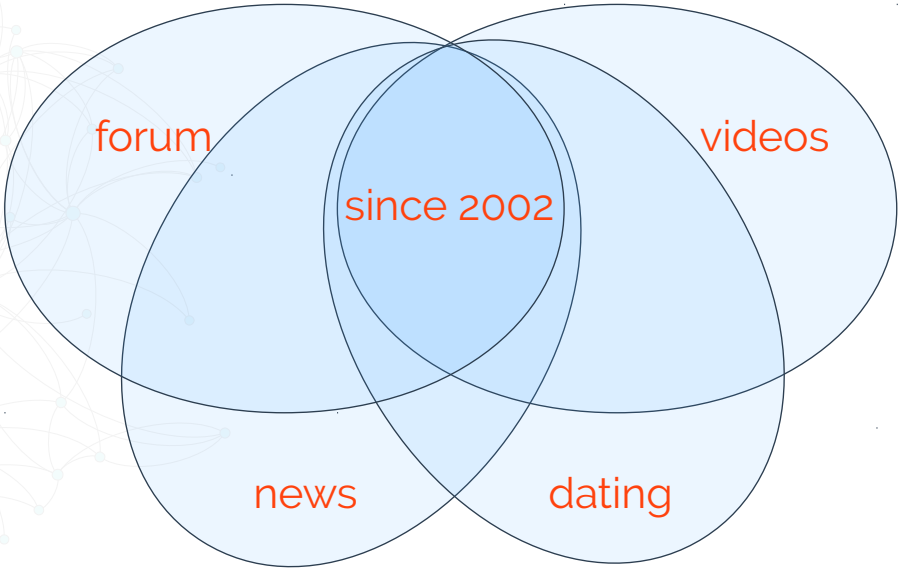
> Look at the maps for a comparison between Archive.org, BNF and e-Diasporas corpora

Focusing on the particular case of yabiladi.com

a hub at the center of the network



an established and hybrid website



> 2.8 Millions of archived pages

As corpora of web archives are wide and sparse :

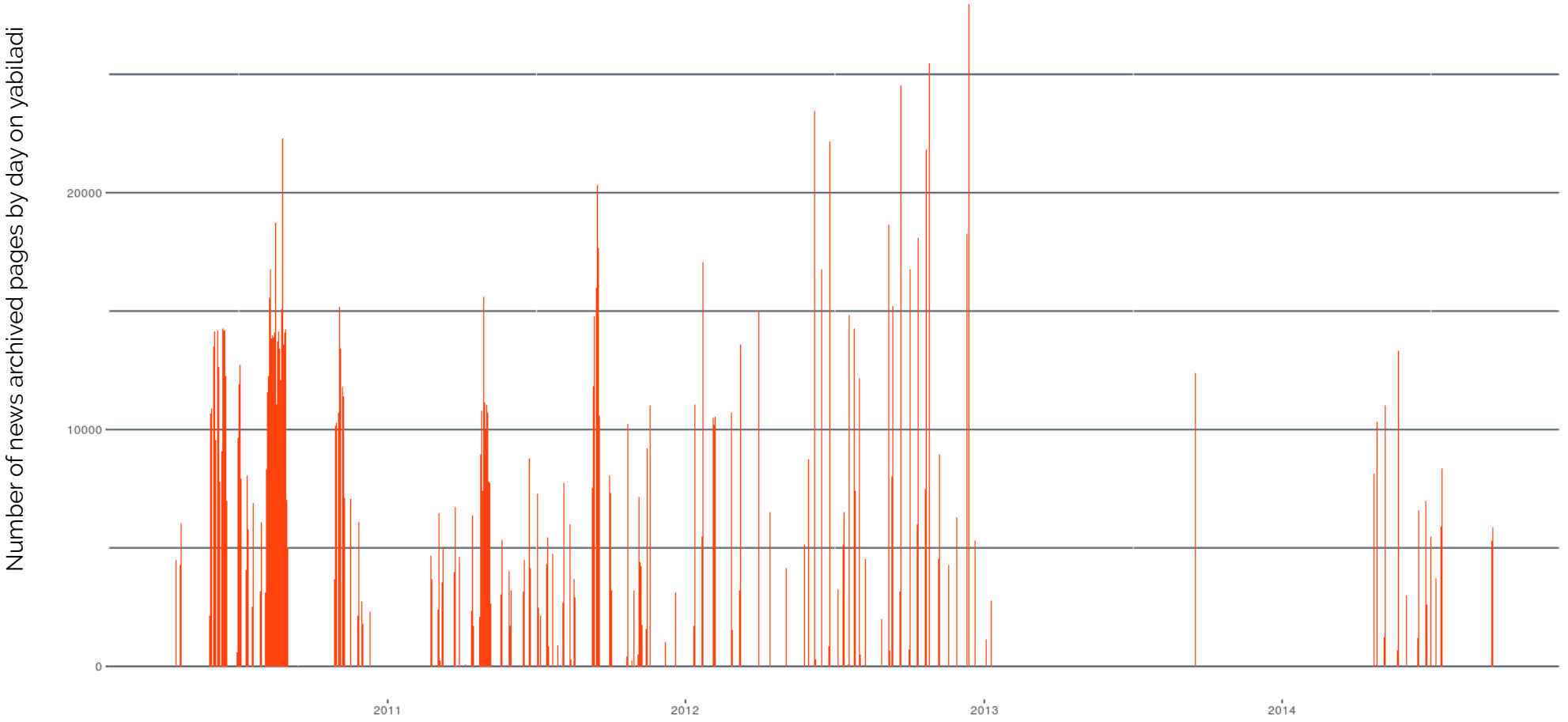
Is there a way to effectively guide a researcher through exploration of web archives ?

Like around a particular event

How can we follow and analyse the traces of an event and its genesis by restoring it in the dual temporality of the web and the real?

The naive approach

> considering all the archived pages as traces of activities on the website



> Are those peaks and valleys relevant ?

The naive approach

> considering all the archived pages as traces of activities on the website

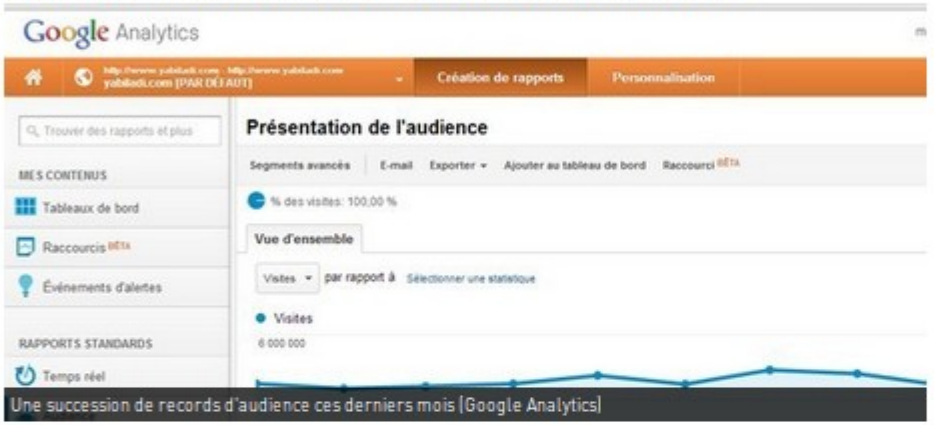
MÉDIA

Publié le 02/02/2013 à 08h00

L'année 2013 débute avec un record d'audience pour Yabiladi.com

Avec près de 11 ans d'existence, le portail Yabiladi.com s'est inscrit comme un média incontournable pour la diaspora marocaine. Même au Maroc, il est devenu un site d'information francophone incontournable depuis notre implantation à Casablanca.

23 partages



Une succession de records d'audience ces derniers mois (Google Analytics)

L'année 2013 débute de la plus belle façon qui soit. Janvier se termine avec un record d'audience puisque le nombre de visites mensuelles a atteint plus de 5 millions de visites et 3 534 000 visiteurs uniques sur le mois (Source Google Analytics).

Alors que près de 20% de l'audience provient du Maroc, plus de 80% des visites est constitué de Marocains résidant à l'étranger mais aussi des amoureux du Maroc à travers le monde.

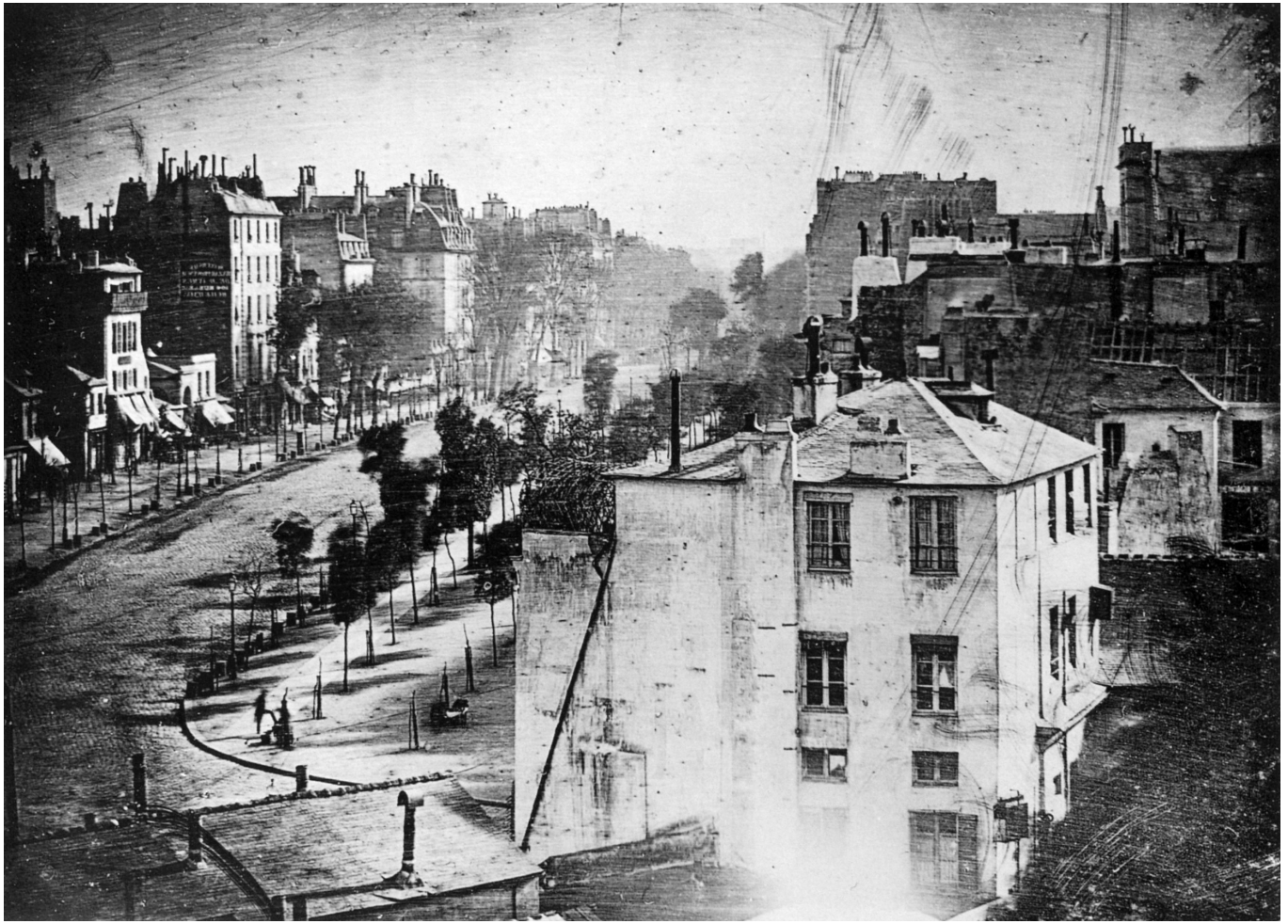
Une tendance importante s'est dessinée sur l'année 2012 et le premier mois de l'année 2013 : le formidable essor du mobile chez nos internautes. Ainsi, en janvier plus de 26% de l'audience provient d'un appareil mobile (téléphone,

No it's not relevant ...
It's easy to find pages edited in 2013
So what have we seen ?

Keep calm and ...
go back to the basics of the structure of a web archive corpus

Archiving is all about selecting and destroying

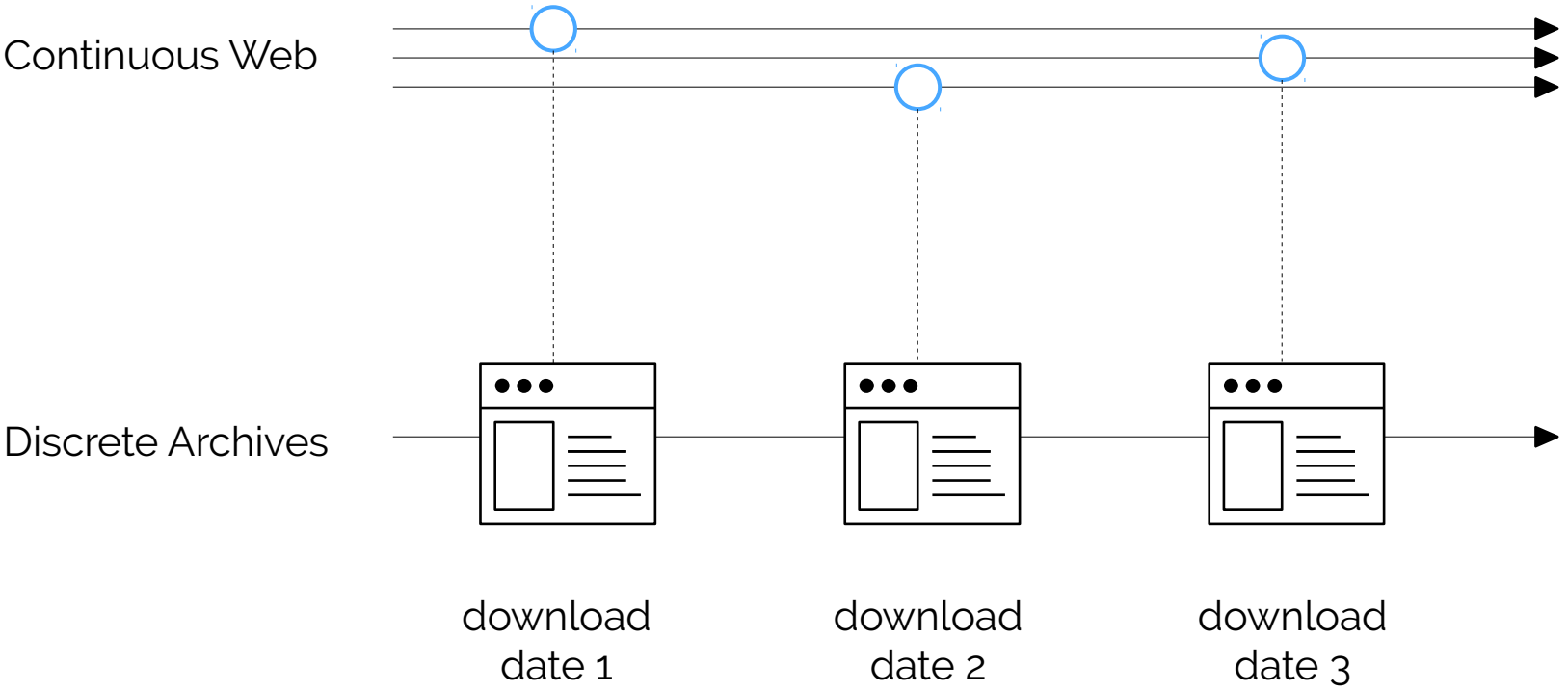
> as webpages change over time



> "Boulevard du Temple", Louis Daguerre, 1838

Web archives are not direct traces of the web

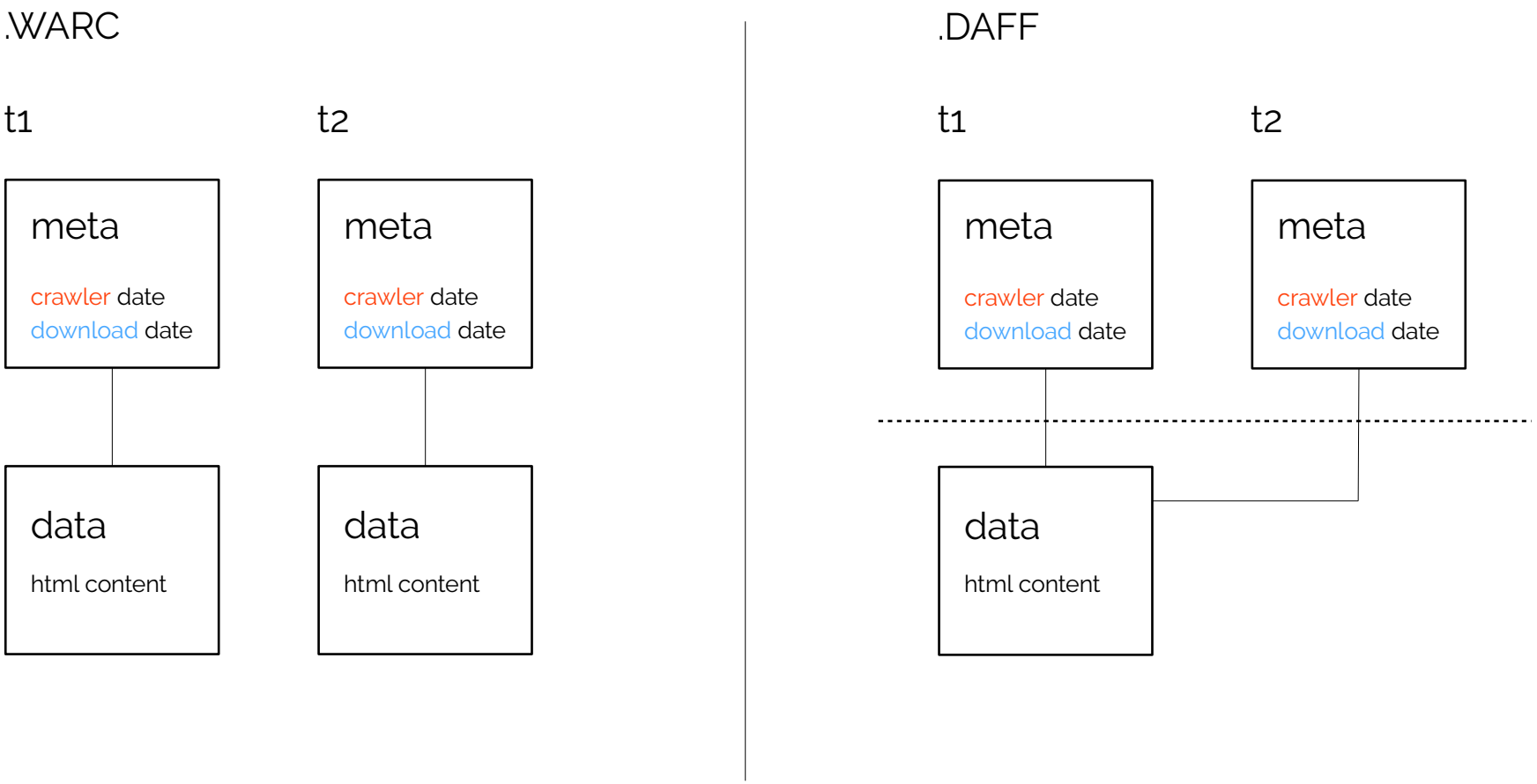
> web archives should be considered as direct traces of the crawler



> We saw what we call a crawl legacy effect

The original scale of web archives is the webpage

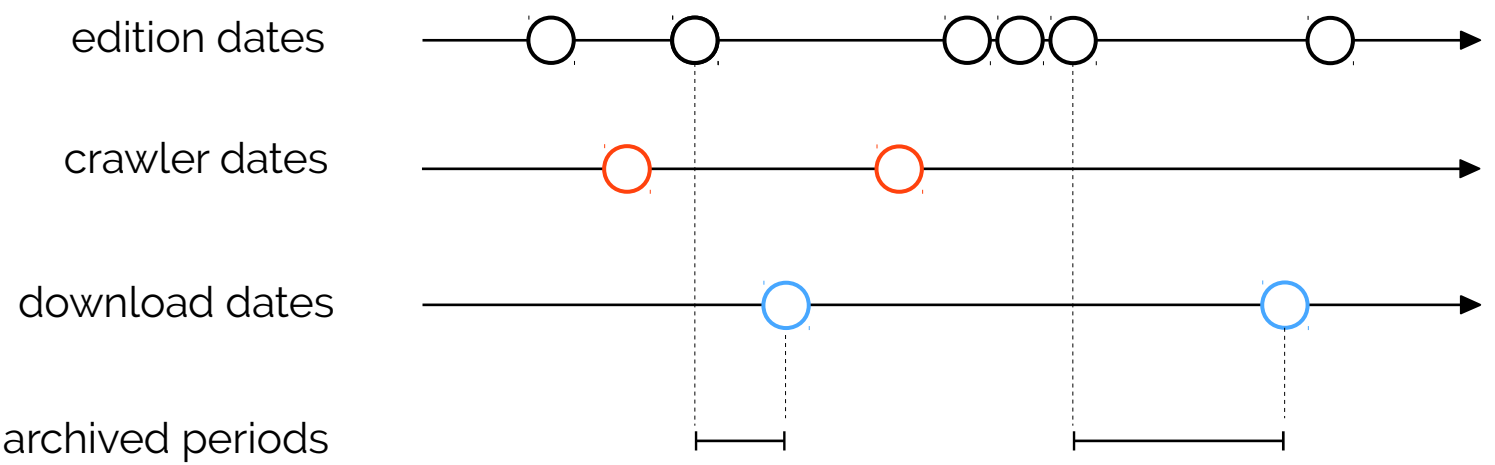
> what can we learn from the structure of web archives files?



> by definition, web archives are built on top of webpages

Archiving on top of webpages goes with many challenges

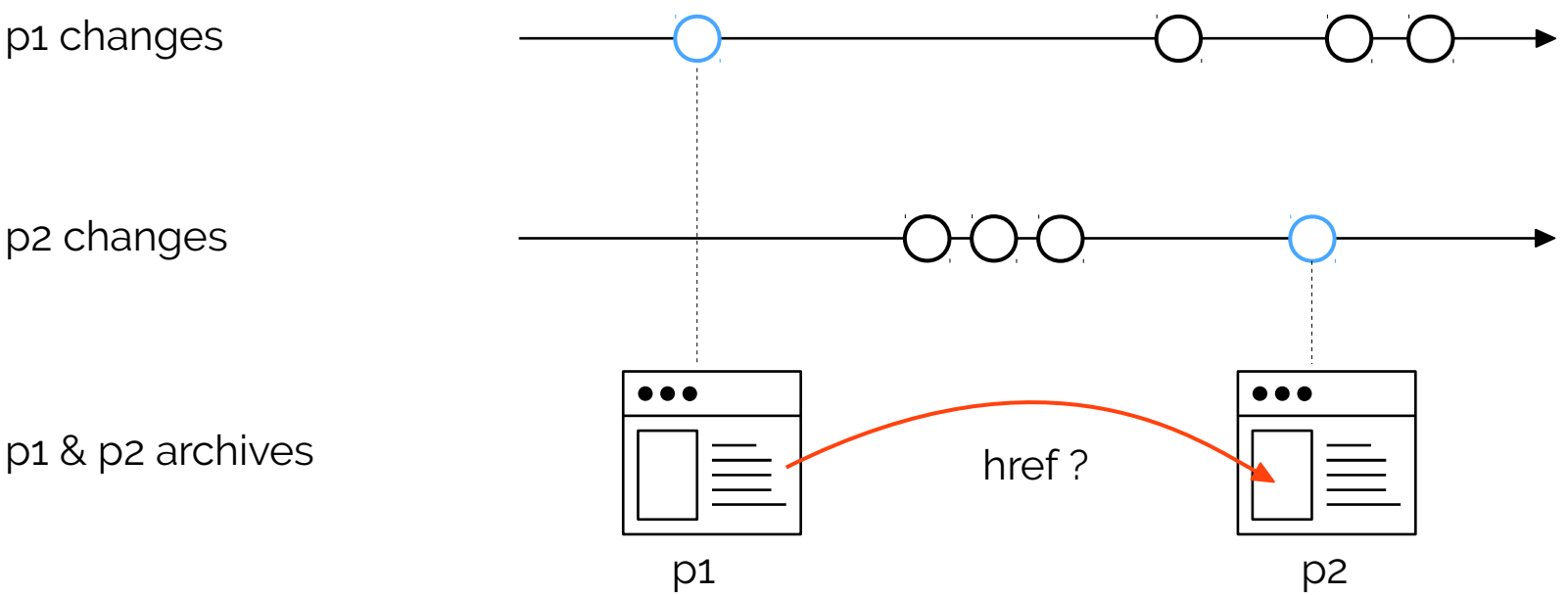
> Crawler blindness and archive quality



> Web archiving goes with construction locks

Archiving on top of webpages goes with many challenges

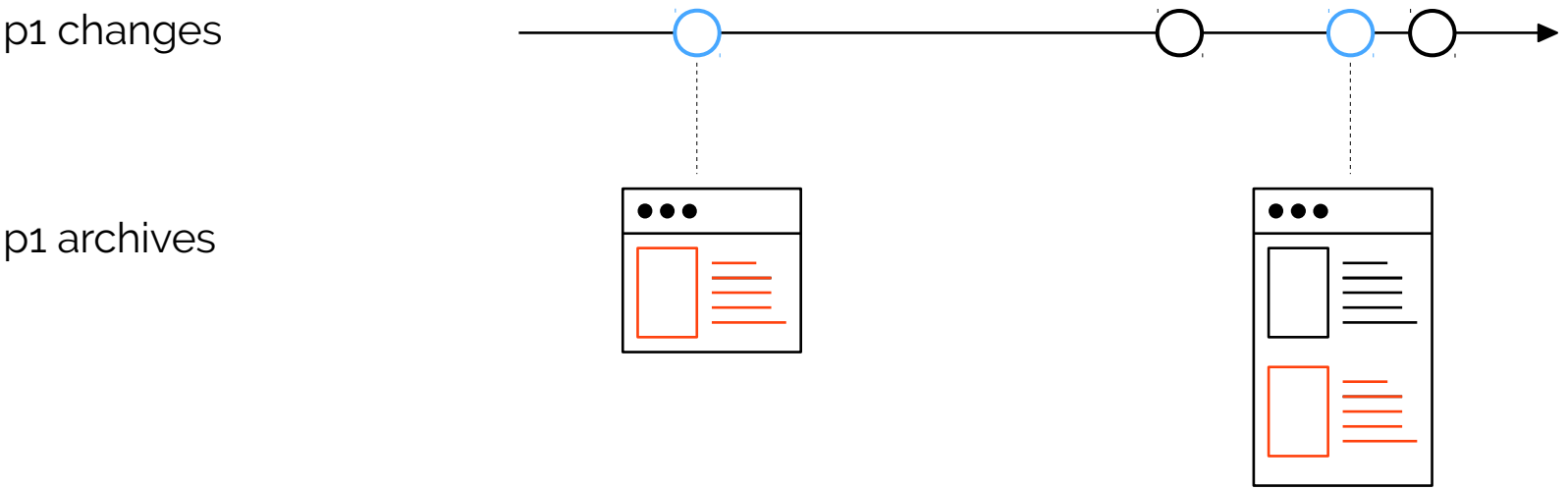
> Archive consistency across pages



> Web archiving goes with navigation locks

Archiving on top of webpages goes with many challenges

> Pages with archive-like content



> Archiving goes with discrete and continuous interpretation locks

To face or reduce crawl legacy effects and effectivelly guide researchers through exploration of web archives :

We introduce a new entity called web fragments

> related works for the extraction of individual components out of webpages

Named entities
Dates
N-grams
Titles
Keywords
Text features
...

But mostly designed for for automatic large scale processes

We need an entity that can be equaly understood by computer scientists sociologists or historians.

The web fragment

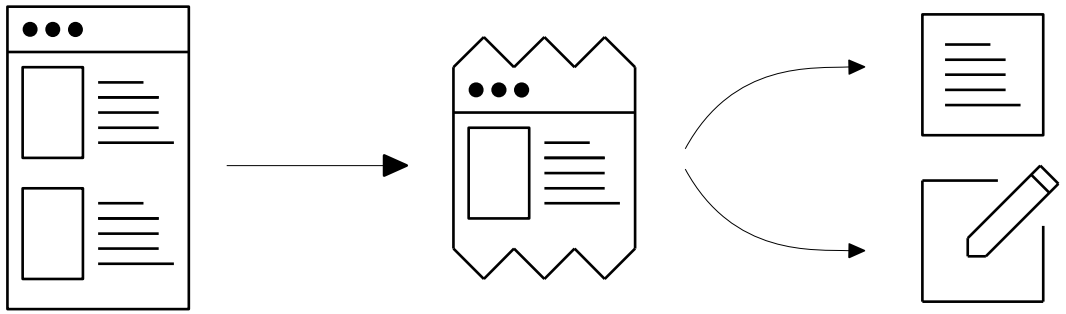
> Definition

Considering the webpage :

- as the unit of access and consulation to the web
- built using it's own writing modalities

Noticing that from the point of view of human perception, a webpage is the result of a logical arrangement of distinct semantic components

> The web fragment is a semantic and syntactic subset of a given webpage ...

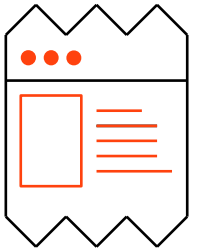


... that deals with a web content and catches the way it has been written and published online

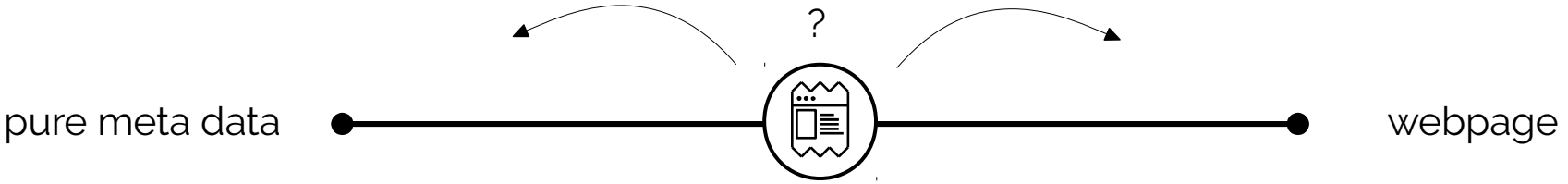
The web fragment

> Definition

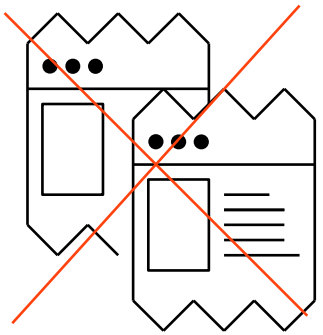
> It's a coherent set of textual, visual or audio content that can be understood on it's own



> There is a scale relationship between a webpage and its fragments



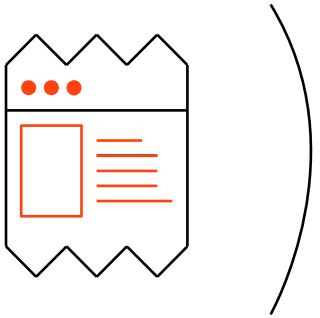
> Within the same webpage, two fragments cannot overlap



The web fragment

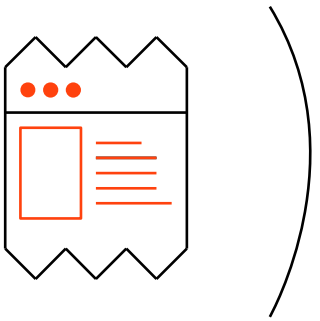
> Definition

> It goes with an associated set of extracted meta contents



Title
Author
Date (edition date)

> It encompass the writing and sharing elements used for publishing and sharing its content



CMS widgets
Integrated text editor
Href links
Rss feed

The web fragment

> Examples

an article

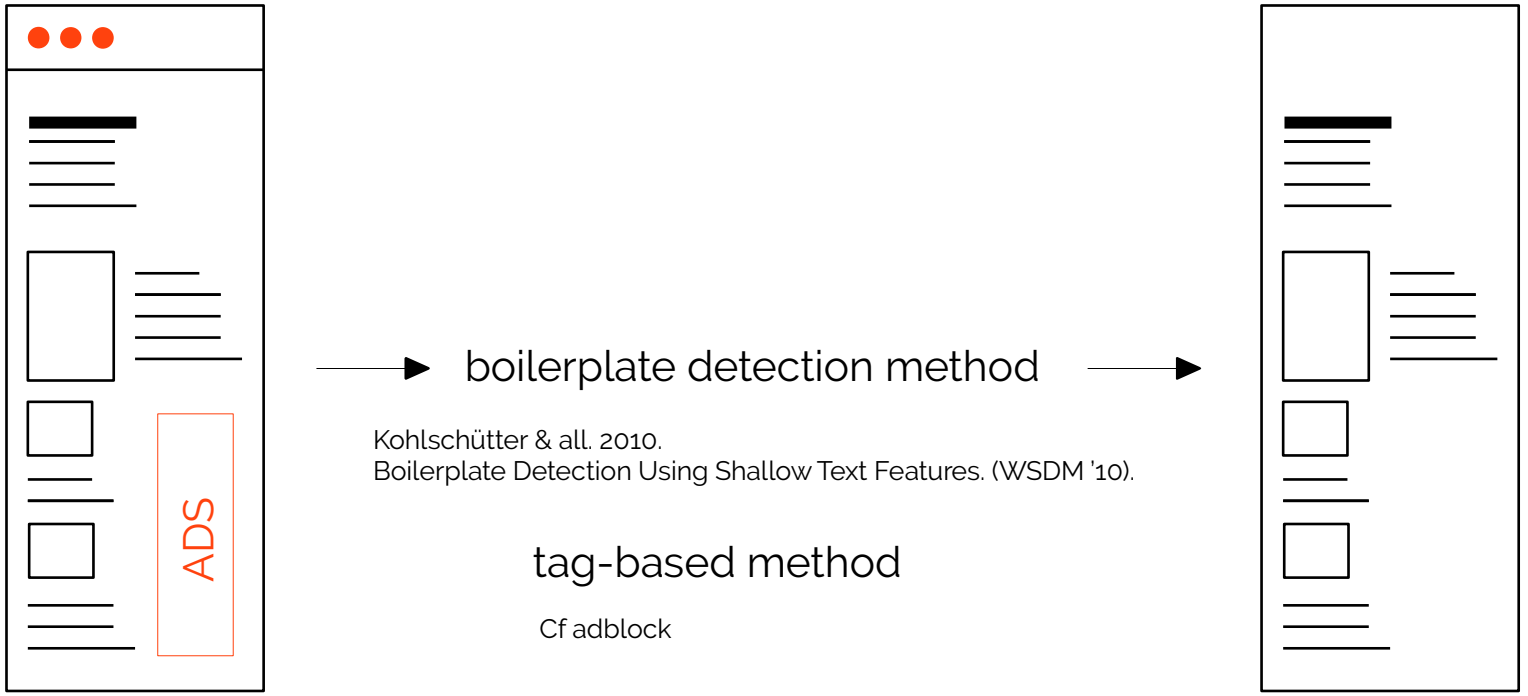
a news item

a comment

> Let's try the firefox module !

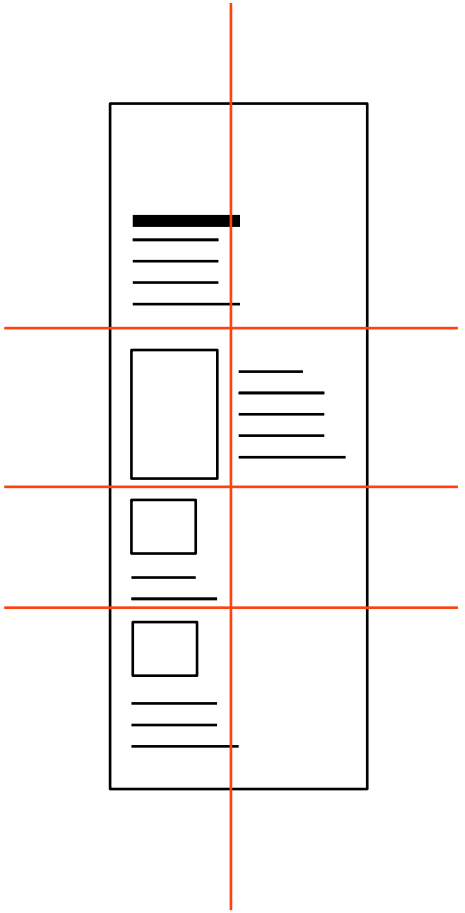
Finding web fragments

> Clean the page



Finding web fragments

> Segmentation and extraction



vision-based method

Cai & all 2003. Vips: a vision-based
page segmentation algorithm. (2003).

tag-based method

Jatowt & all 2007. Detecting age of page content.
In Proceedings of the 9th annual ACM international workshop on Web information and data management. ACM, 137–144.

ad-hoc rules

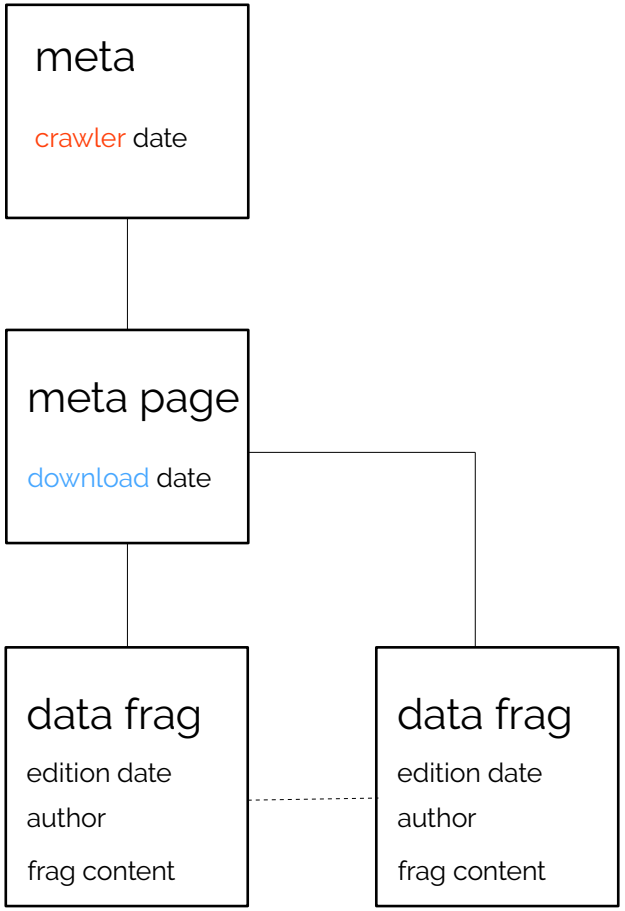
Title U Author U Text

distance score node by node

> Implementation inspired by readability & fathom from Mozilla

The web fragment

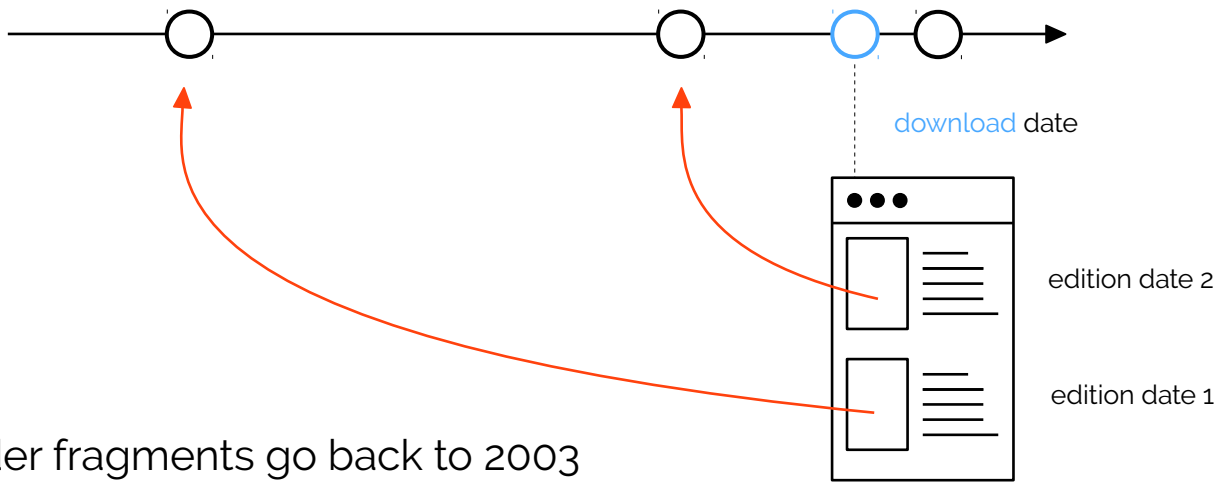
> New structure for web archives




```
<field name="id"                          type="string"

<!-- archive fields -->
<field name="archive_active"              type="boolean"
<field name="archive_corpus"              type="string"
<field name="archive_mime_type"           type="string"
<field name="archive_country"             type="string"
<field name="archive_lang"                type="string"

<!-- crawler fields -->
<field name="crawl_id"                    type="string"
<field name="crawl_date"                  type="date"
<field name="crawl_date_first"            type="date"
<field name="crawl_date_last"             type="date"

<!-- page fields -->
<field name="page_downl_id"               type="string"
<field name="page_downl_date"             type="date"
<field name="page_downl_date_first"       type="date"
<field name="page_downl_date_last"        type="date"
<field name="page_domain"                 type="string"
<field name="page_url"                    type="string"
<field name="page_url_id"                 type="string"
<field name="page_space"                  type="string"
<field name="page_title"                  type="string"
<field name="page_description"            type="text"
<field name="page_published_date"         type="date"
<field name="page_publisher"              type="string"

<!-- fragment fields -->
<field name="frag_type"                   type="string"
<field name="frag_author"                 type="string"
<field name="frag_date"                   type="date"
<field name="frag_date_first"             type="date"
<field name="frag_date_level"             type="dateLevel"
<field name="frag_href"                   type="string"
<field name="frag_href_id"                type="string"
<field name="frag_ratio"                  type="int"
<field name="frag_node"                   type="text"
<field name="frag_offset"                 type="int"
<field name="frag_text"                   type="text"
<field name="frag_text_id"                type="string"
<field name="frag_text_shingle"           type="shingle"
```

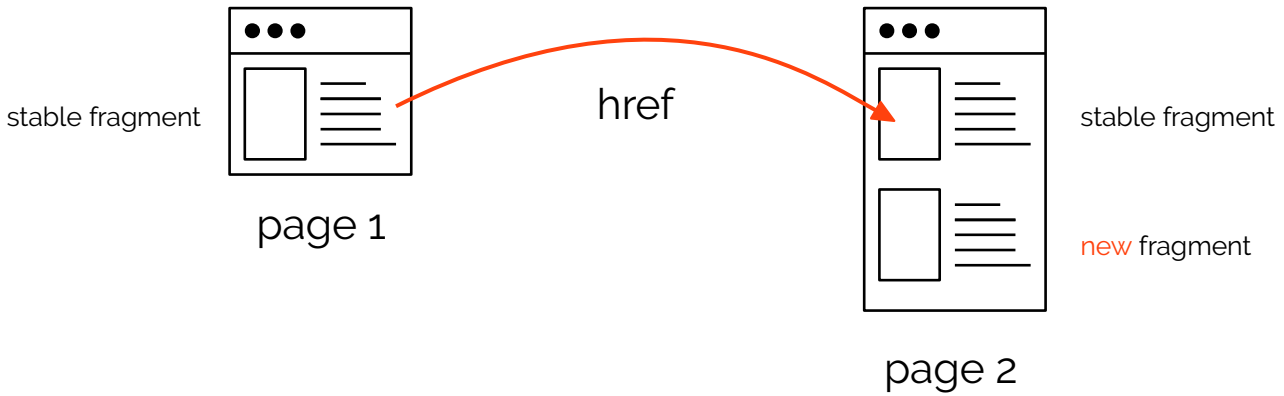
Rethinking archive challenges using web fragments

> Crawler blindness can be reduced and archive quality increased



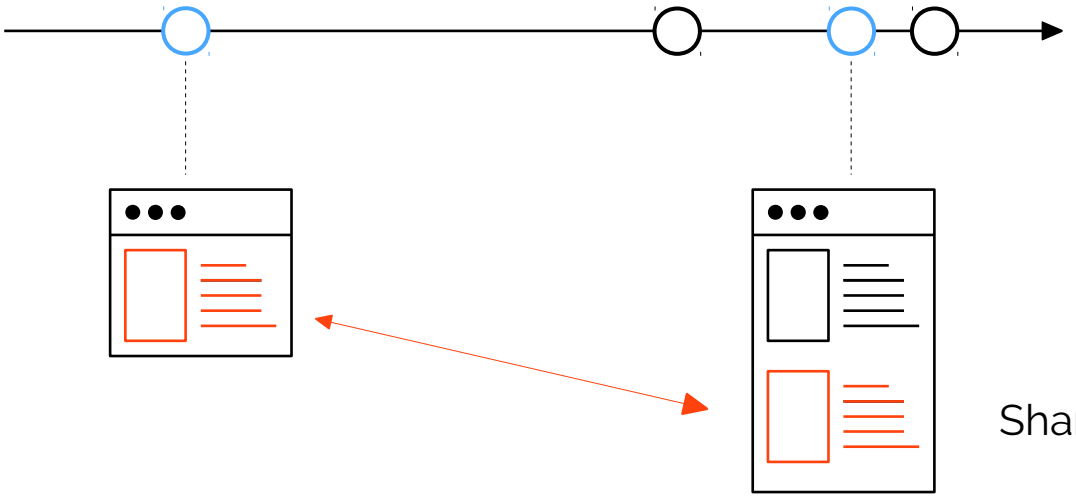
Yabiladi's older fragments go back to 2003

> We introduce a more permissive archive consistency based on fragments and user requests



Rethinking archive challenges using web fragments

> Pages with archive-like content is no more a problem with web fragments as a search unit base



Sharing the same id (sha256)

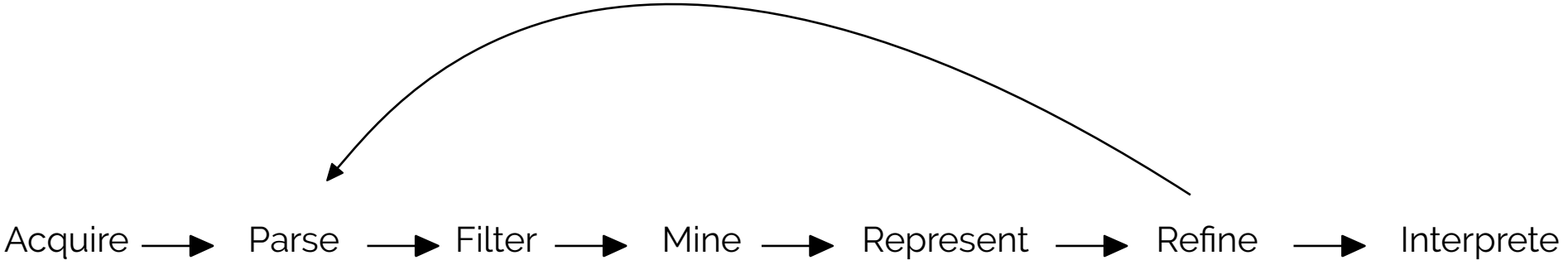
> Web fragments help us expanding web archives beyond web pages

Now let's see how we can concretely conduct an exploratory archive analysis ...

Exploratory analysis of Web archives

> Following John Wilder Tukey's work

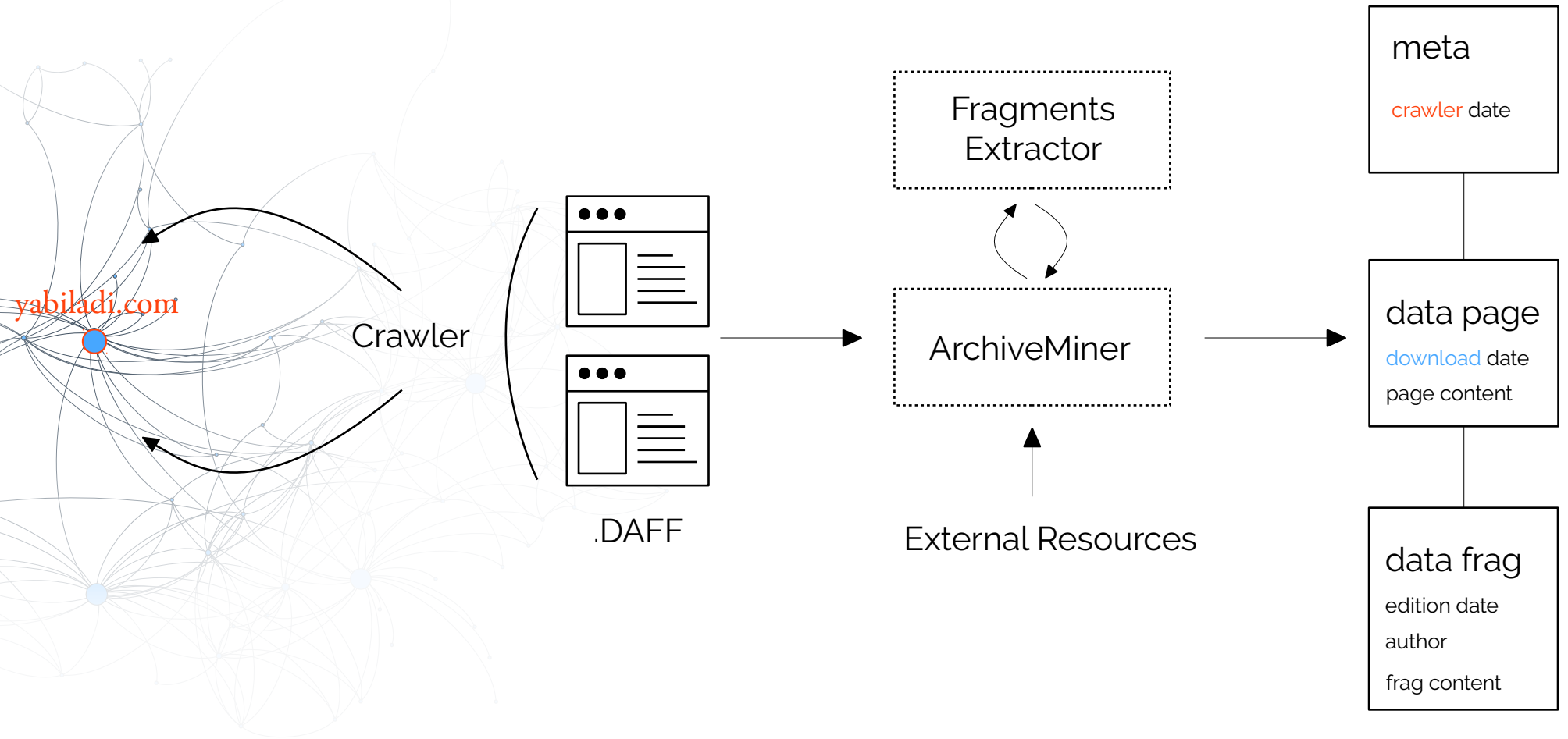
An iterative process that is deliberately part of a logic of observation, discovery and astonishment



Archives extraction engine

Acquire → Parse → Filter → Mine → Represent → Refine → Interprete

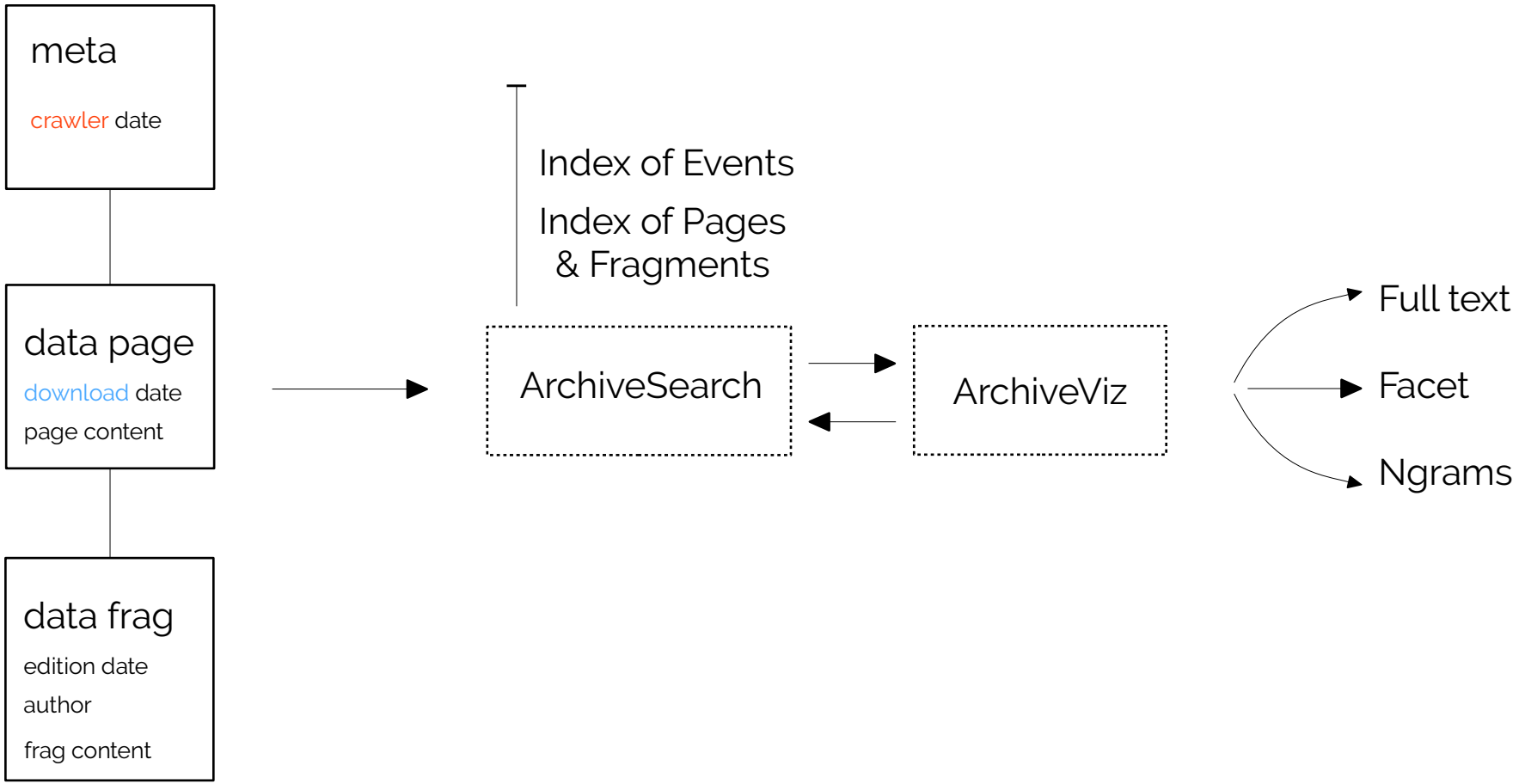
> The Web Archives Explorer (part 1)



Archives exploration engine

Acquire → Parse → Filter → Mine → Represent → Refine → Interprete

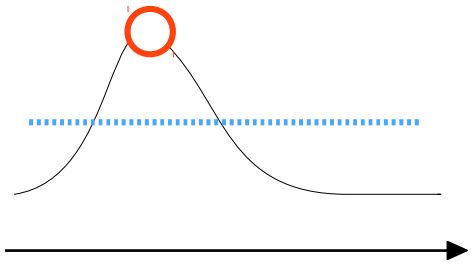
> The Web Archives Explorer (part 2)



Using web fragments

Acquire → Parse → Filter → Mine → Represent → Refine → Interprete

> Using an event detection system



2. identification with titles of news articles

1. threshold-based detection

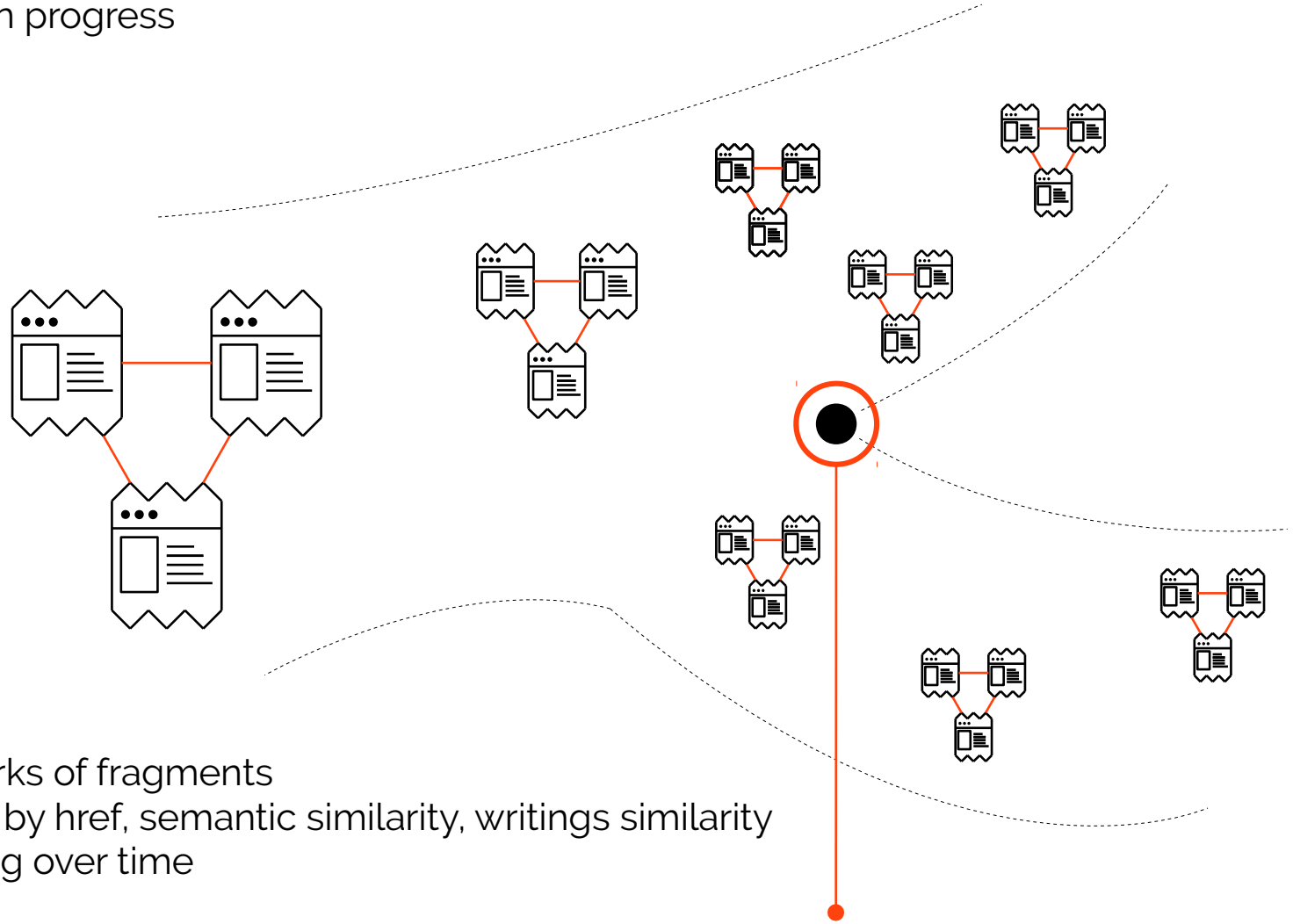
3. fields and experts interpretations

> Let's see the Web Archives Explorer in action

video presentation for CIKM2017

Using web fragments

> Work in progress



> Networks of fragments
Linked by href, semantic similarity, writings similarity
Evolving over time

End of part 1 !
Thank you

> e-Diasporas web archives are "just" one example among many of archived digital traces of migration … (let's go to part 2)

Further studies

> Navigation logs at BPI

Free & anonymous wifi access,
no reservation, during 40 mn

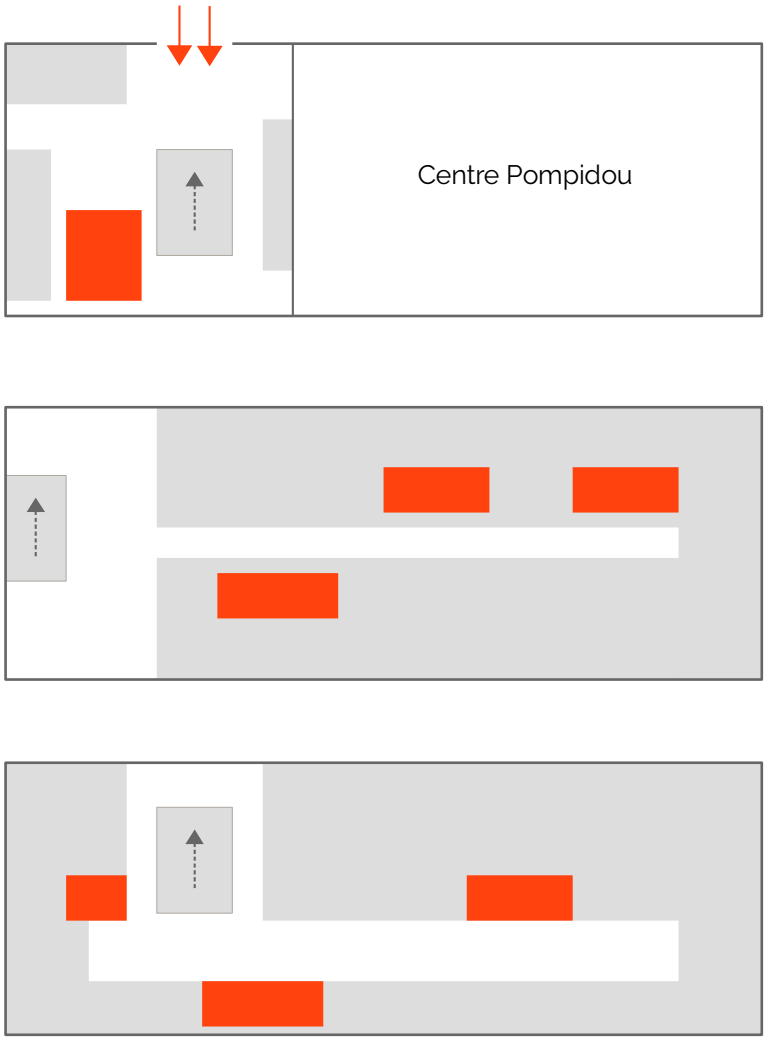
170 computers distributed on the 3 floors
among shared tables

Open from 12h - 22h every day except the
Tuesday

Experiments of up to 250 free access computers
without time limit
> May 2016

A precarious public (homeless, migrant,
precarious intellectual …)

**What is the role of the free Internet
in the daily life of the public of the BPI ?**



Source Data

> web navigation logs

```
10.6.8.85   - PUB-2-INT-242 [23/02/2017 13:15:49] "GET https://www.google-analytics.com:443 HTTP/1.1"    200 - - 1000 -
10.6.6.218  - PUB-2-INT-239 [23/02/2017 13:15:47] "GET https://www.youtube-nocookie.com:443 HTTP/1.1"    200 - - 1207 -
10.6.6.218  - PUB-2-INT-239 [23/02/2017 13:15:45] "GET https://s.youtube.com:443 HTTP/1.1"                200 - - 1207 -
10.6.6.218  - PUB-2-INT-239 [23/02/2017 13:15:43] "GET https://s.ytimg.com:443 HTTP/1.1"                  200 - - 1207 -
10.6.8.85   - PUB-2-INT-242 [23/02/2017 13:15:49] "GET https://www.google.com:443 HTTP/1.1"               200 - - 1000 -
10.6.6.218  - PUB-2-INT-239 [23/02/2017 13:15:47] "GET https://www.googleapis.com:443 HTTP/1.1"          200 - - 1000 -
10.6.6.218  - PUB-2-INT-239 [23/02/2017 13:15:47] "GET https://www.youtube.com:443 HTTP/1.1"             200 - - 1207 -
10.6.8.121  - PUB-1-INT-125 [23/02/2017 13:03:35] "GET http://www.bpe.europresse.com:443 HTTP/1.1"       200 - - 1000 -

                                                  ...

10.6.8.121  - PUB-1-INT-125 [23/02/2017 13:43:27] "GET https://intensedebate.com:443 HTTP/1.1"  200     - - 1228 -
```

Around 2.000.000 logs / day

An exploration pipeline

> for extracting informations

Filter the logs
Drop ads, image, css ...
Aggregate by session
Arround 200,000 cleaned logs / day

Search inside the logs :

| date
| url
| domain name
| extension (.com .fr .ru ...)
| category (olfeo)
| catégory (bpi)
| session id
| stage #
| table #
| computer #

Visualize the logs by fields on the fly

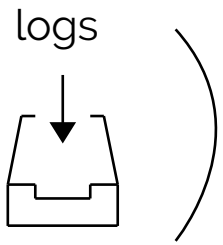
Can we clean everithing ?

The relation between what we see on the screen and what is recorded

With the participation of Chaima Berrachdi

First results

> From the fields and from the data



navigation profils ?

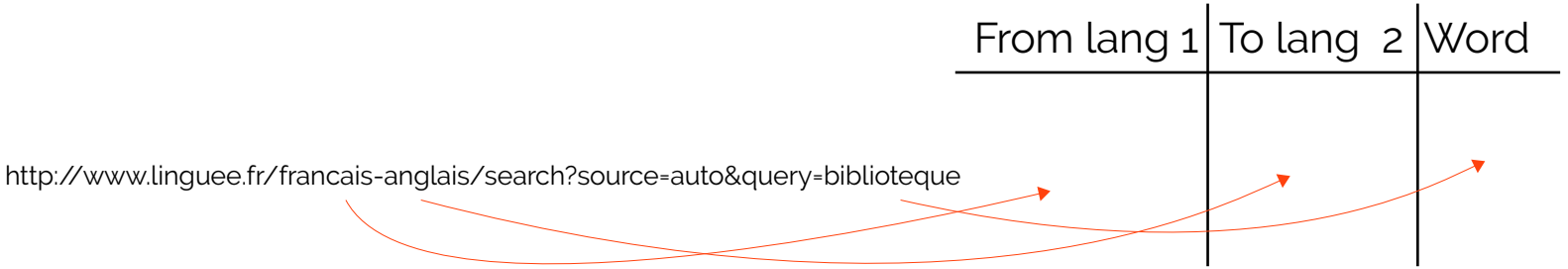
Clustering users by category of url :

- Watching videos (youtube, dailymotion)
- Social networks
- Google services

> What about the long tail ?

Traduction – Online Games – Regional news – Dating sites – Public services

> How to find structured and miningfull information out of an url ?



With the participation of Marie Pierru & George Legrady

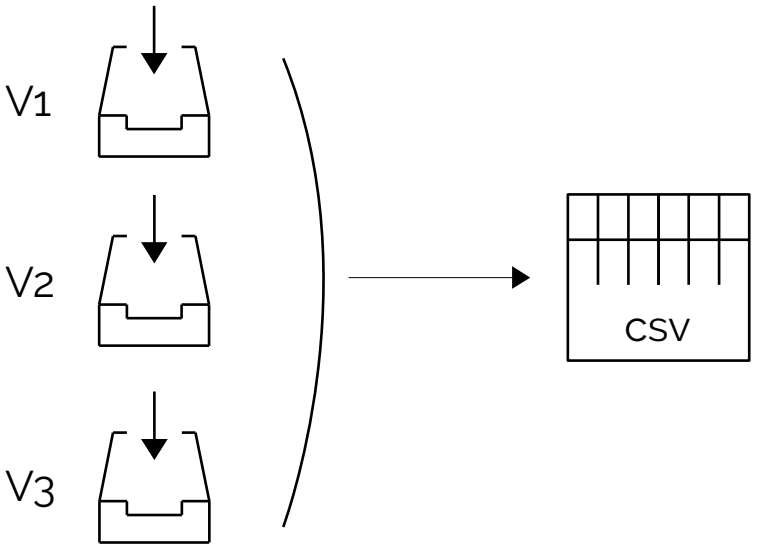
Further studies

> Calm letters

11892 form letters to the CALM (*Comme à la Maison*) program of Singa for hosting refugees

cp	age	Motivation	Type de logement	Taile	Parlez nous de vous	Enfant ?	Enseignement métier ?	Partage Réseau	Capacité d'accueil	Autres info
INT	INT	TEXT	TEXT	INT	TEXT	BOOL	BOOL	BOOL	BOOL	TEXT

From June 2015 to June 2017



Hosting organized by digital platforms

The vocabulary of hosting

The role of the media

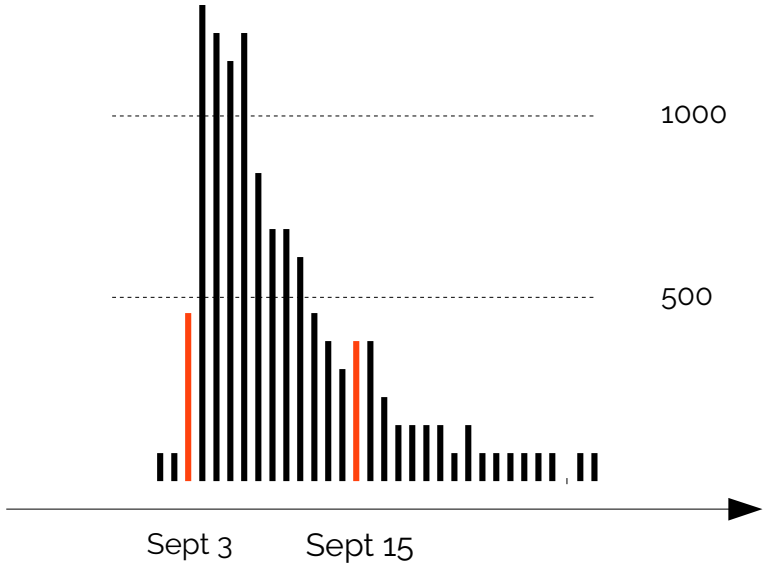
With the participation of Students of Telecom ParisTech

Furst Results

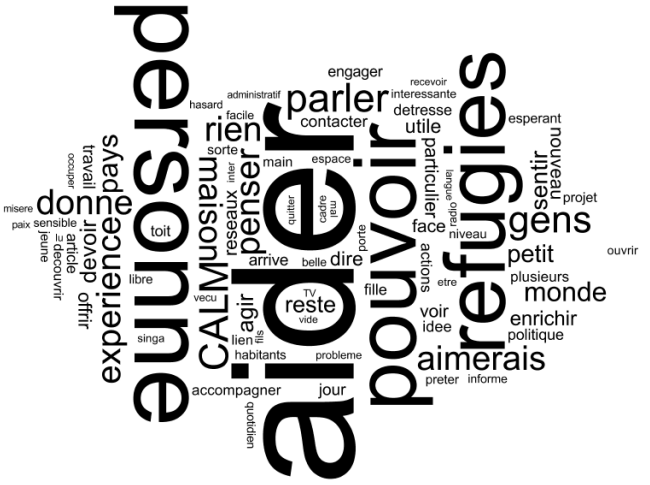
> Calm letters

Demographic profil of the families : 30's & 50's year old – educated – working on education care and cultural world

The relation with mediatic events :



The vocabulary :



From Proposer to Partager

From Refugier to Personne

With the participation of Students of Telecom ParisTech

Thank you !
Questions ?