Introducing Web Fragments

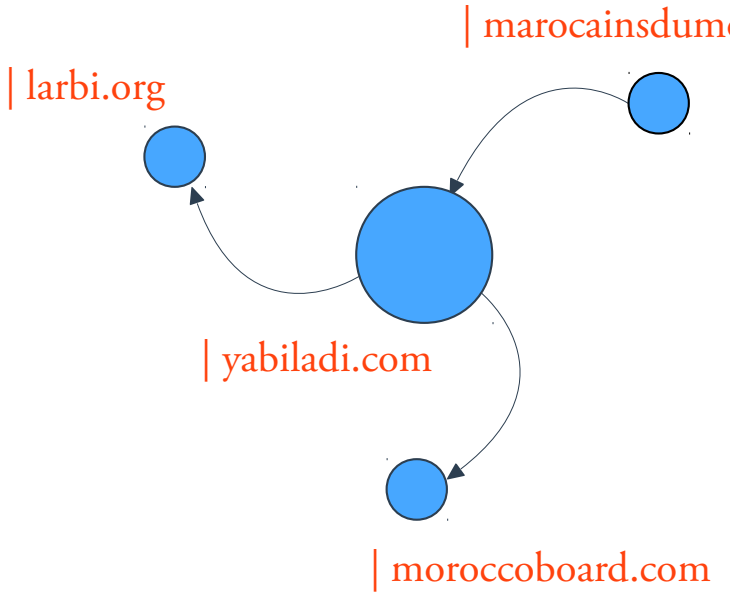
An exploration of web archives beyond the webpages

Quentin Lobbé (LTCI, Télécom ParisTech & Inria Paris)

DBWeb seminar – May 31, 2017

The e-diasporas Atlas

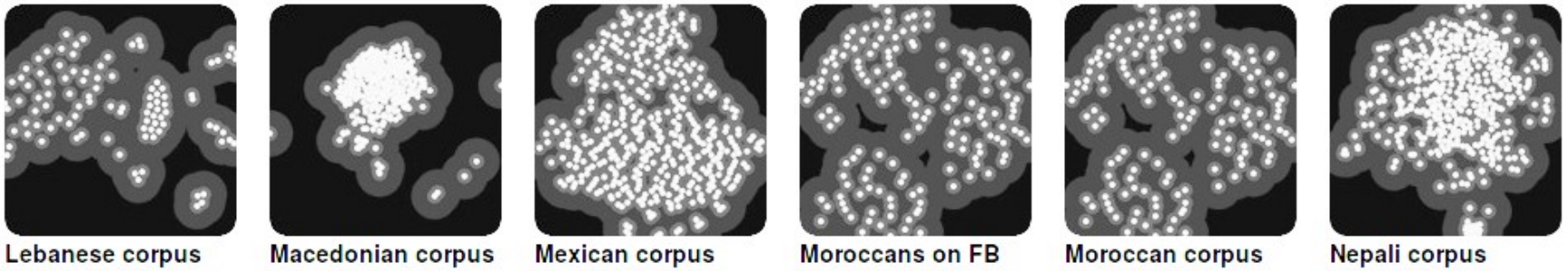
> A collection of online migrant collectives



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

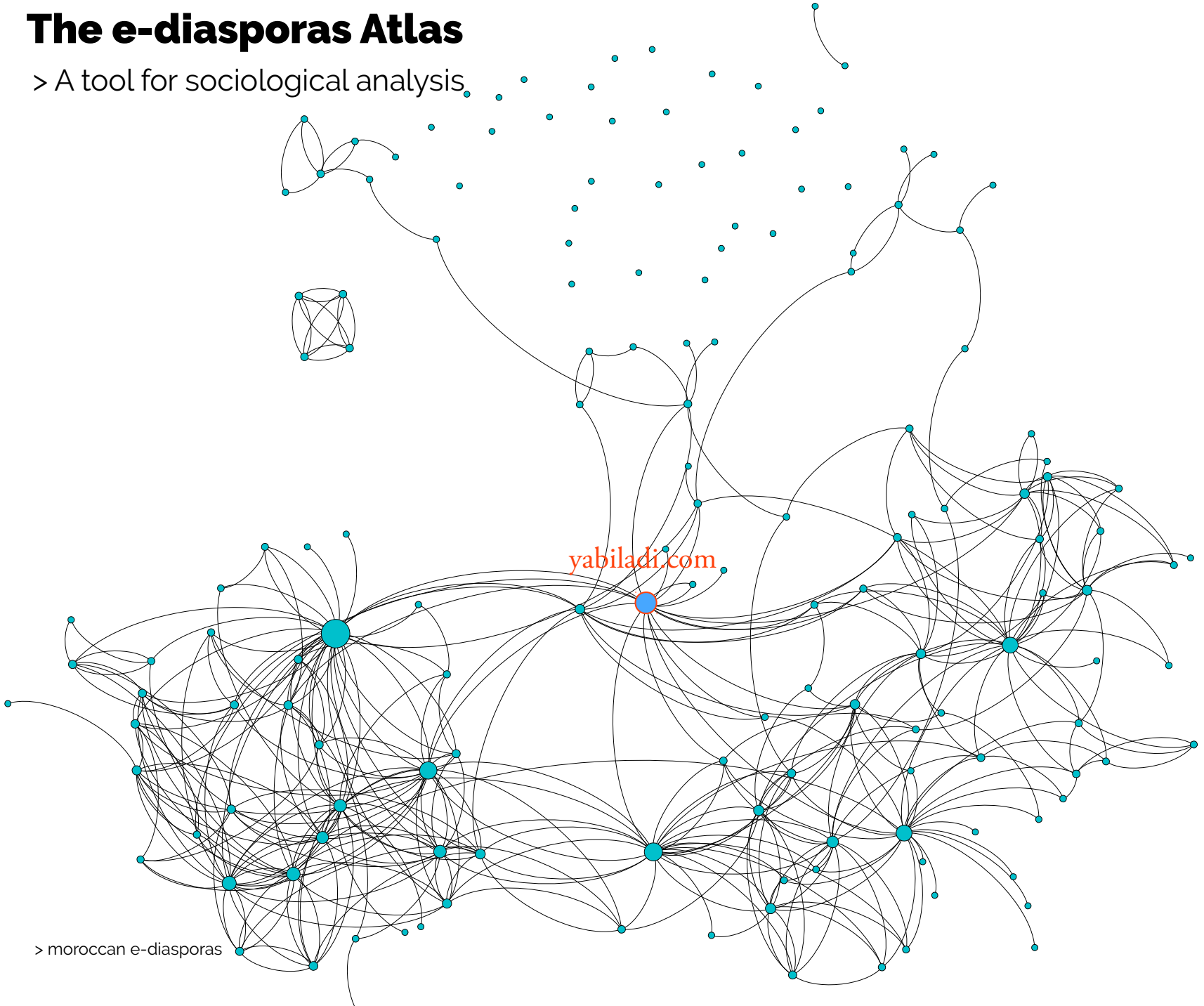
An e-Diaspora is a directed network of migrant websites linked by url

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



The e-diasporas Atlas

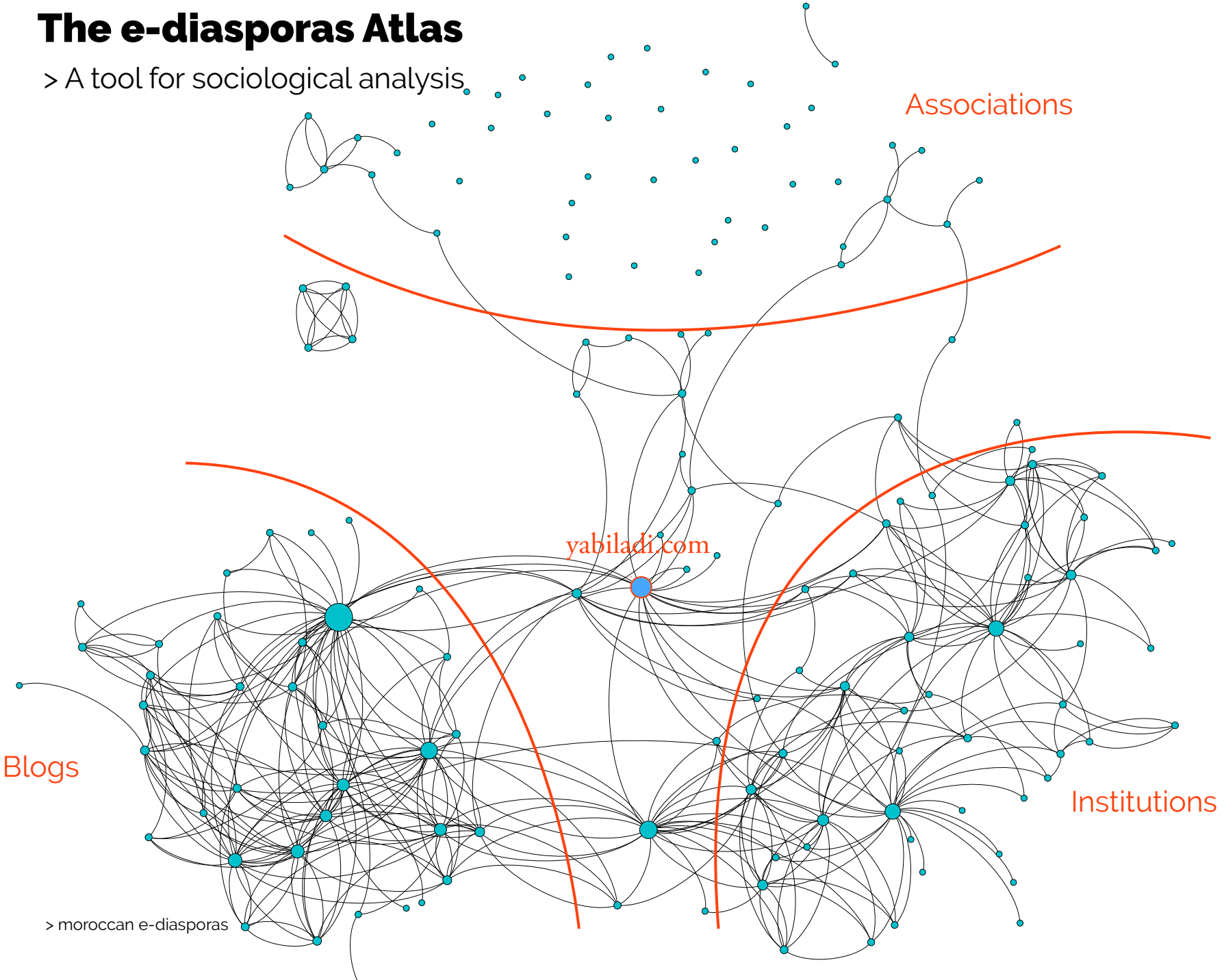
> A tool for sociological analysis



> moroccan e-diasporas

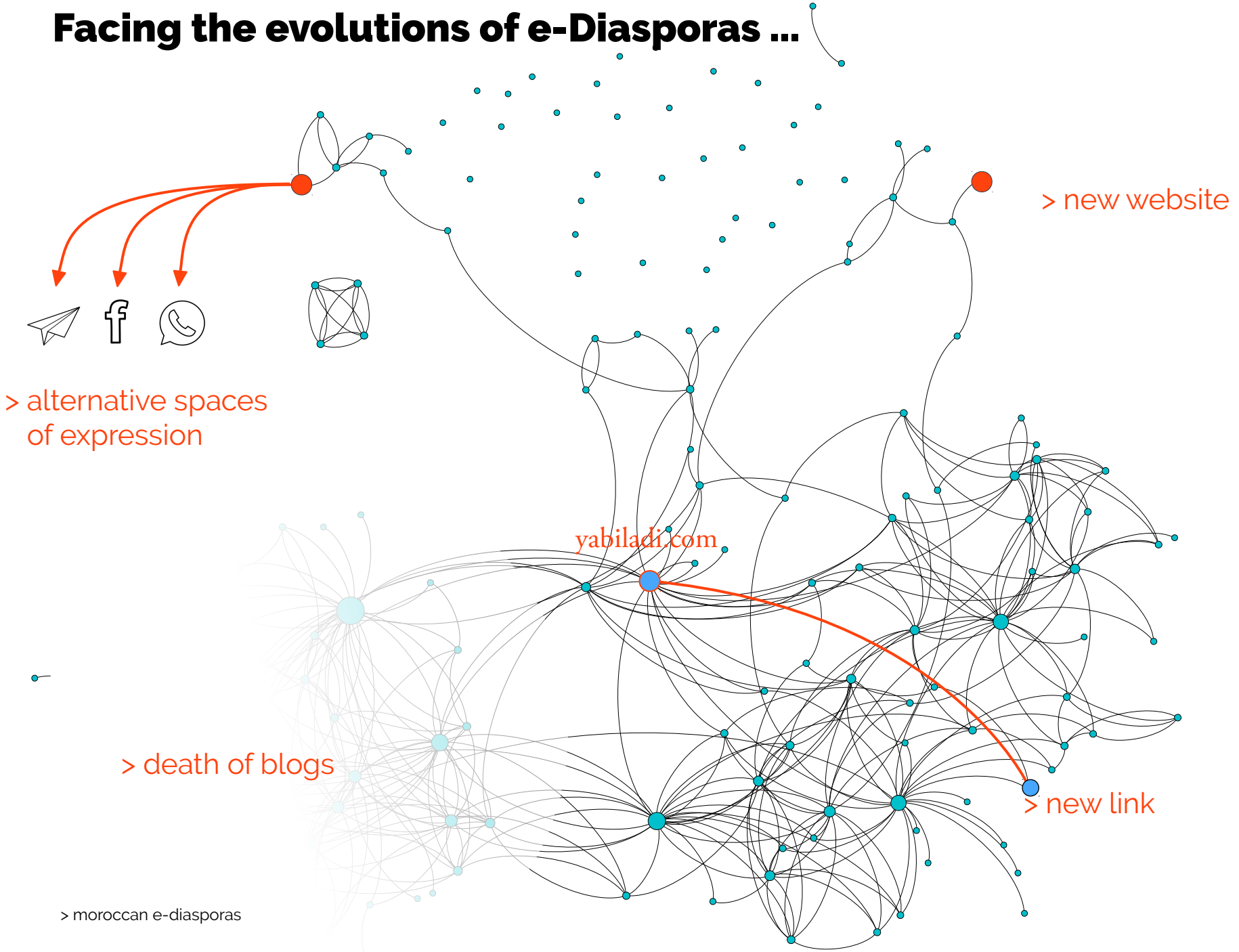
The e-diasporas Atlas

> A tool for sociological analysis



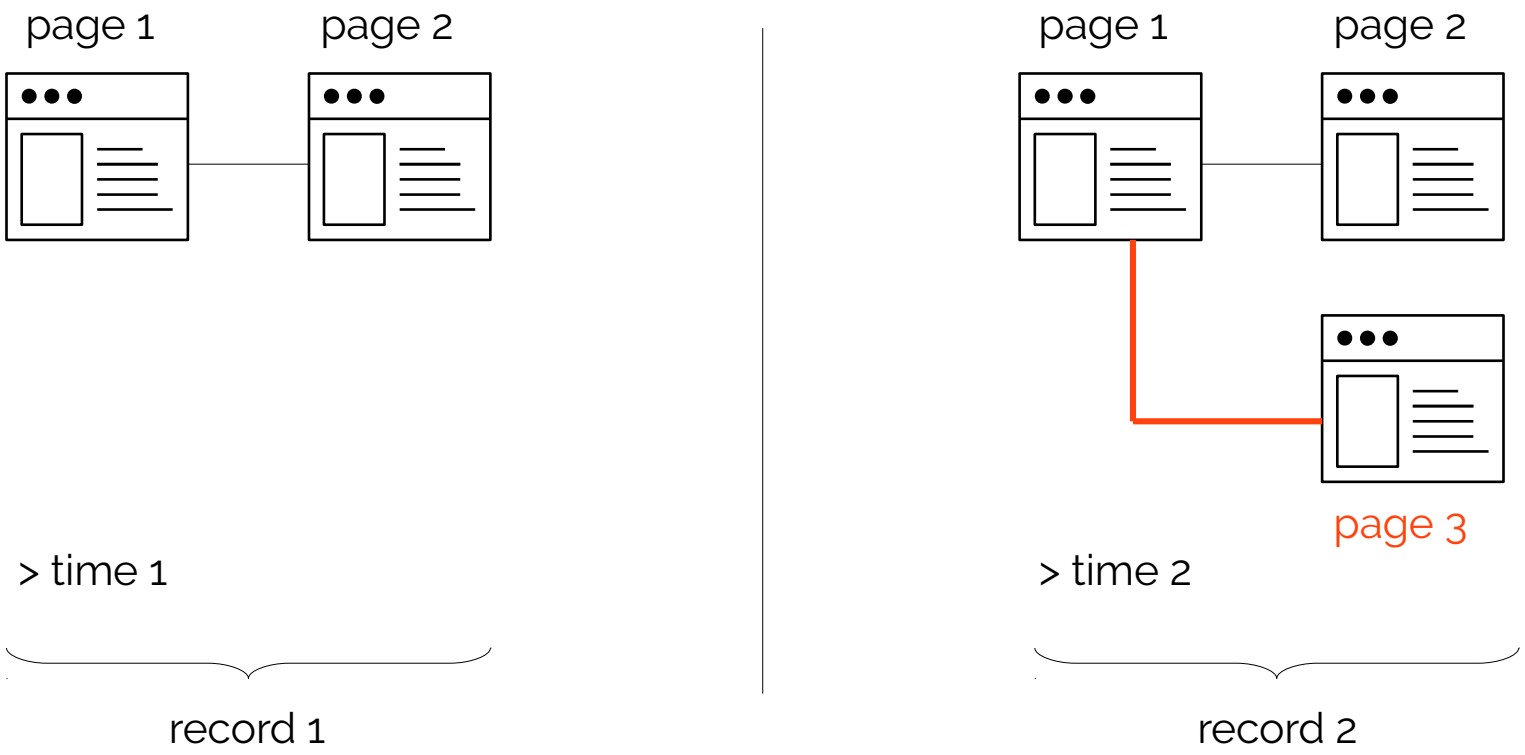
> moroccan e-diasporas

Facing the evolutions of e-Diasporas ...



... we build a corpus of web archives

> To keep a trace of the evolutions of websites



> Our corpus is a 70 To web archive, categorized by e-diasporas corpus, crawled weekly or Monthly, between 2010 and 2015 hosted at the INA

Our original research questions

> Considering the e-Diasporas archived corpus

Can the structure and content of the archived e-Diasporas be permeable to the effects of shocks and external events such as political and social mobilizations?

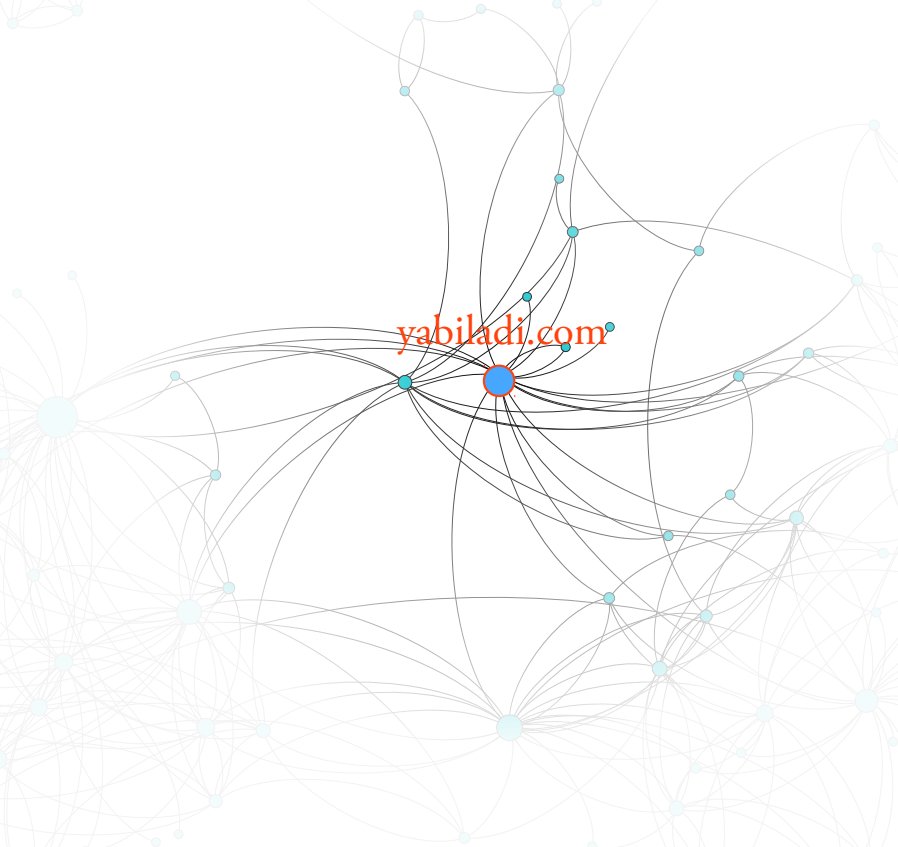
> Considering any archived corpus

How can we follow traces through web archives in order to deal with a given event and its genesis by restoring it in the dual temporality of the web and the real?

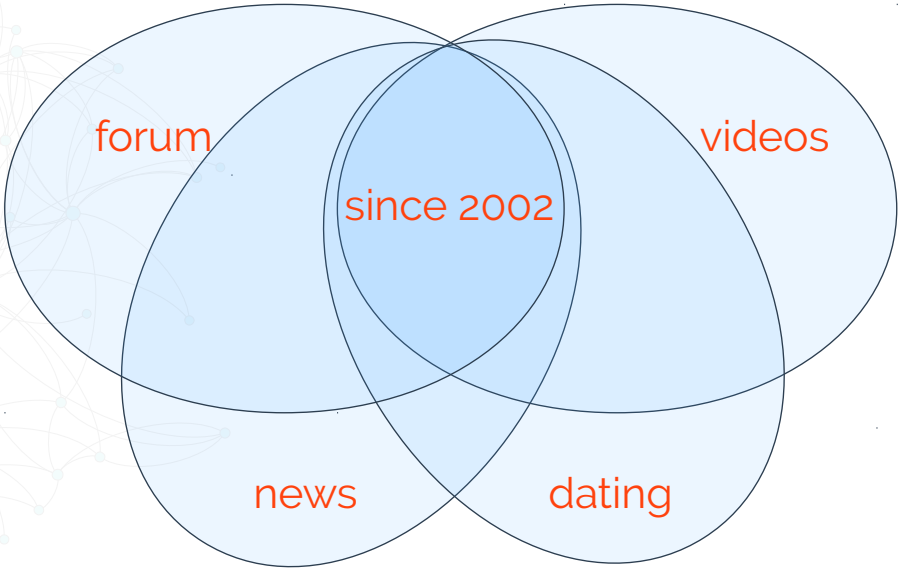
The naive approach

> focusing on the particular case of yabiladi.com

a hub at the center of the network



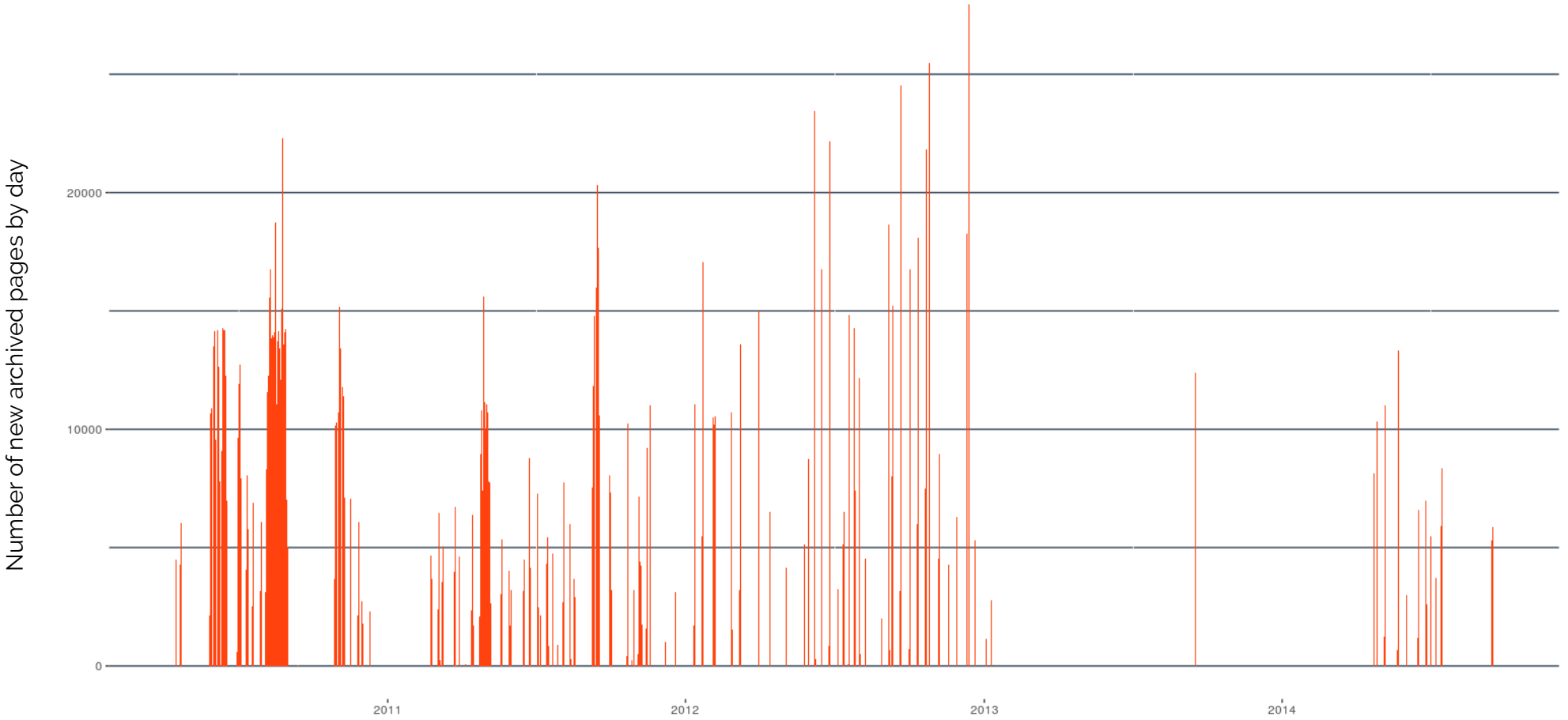
an ancient and hybrid website



> 2.8 Millions of archived pages

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

POLITIQUE

Publié Le 29/08/2013 à 17h00

Naturalisation : La France simplifie les procédures pour revenir aux 100 000 acceptations par an

La simplification des procédures de naturalisation voulu par le ministère français de l'Intérieur et qui faisait l'objet d'une circulaire en octobre dernier, se concrétise. Manuel Valls vient d'émettre deux décrets à ce sujet. La nouvelle est source de critique au sein de l'opposition, mais réjouit les étrangers.

3 partages

Manuel Valls souhaite doubler le nombre de naturalisations / DR

Le ministre français de l'Intérieur, Manuel Valls, veut revenir à un rythme annuel de 100 000 naturalisations par voie réglementaire, a-t-il fait savoir, mercredi, lors du Conseil des ministres. Il entend rompre avec la politique de son prédécesseur qui, selon lui, a fait preuve d'«un manque de transparence et de justesse dans l'appréciation» de l'accès à la nationalité.

En effet, le nombre de naturalisations avait fortement chuté avec la politique de

Radio Yabiladi

Amaziria

72 auditeurs

Election Présidentielle

franceinfo: + Yabiladi

FIL INFO

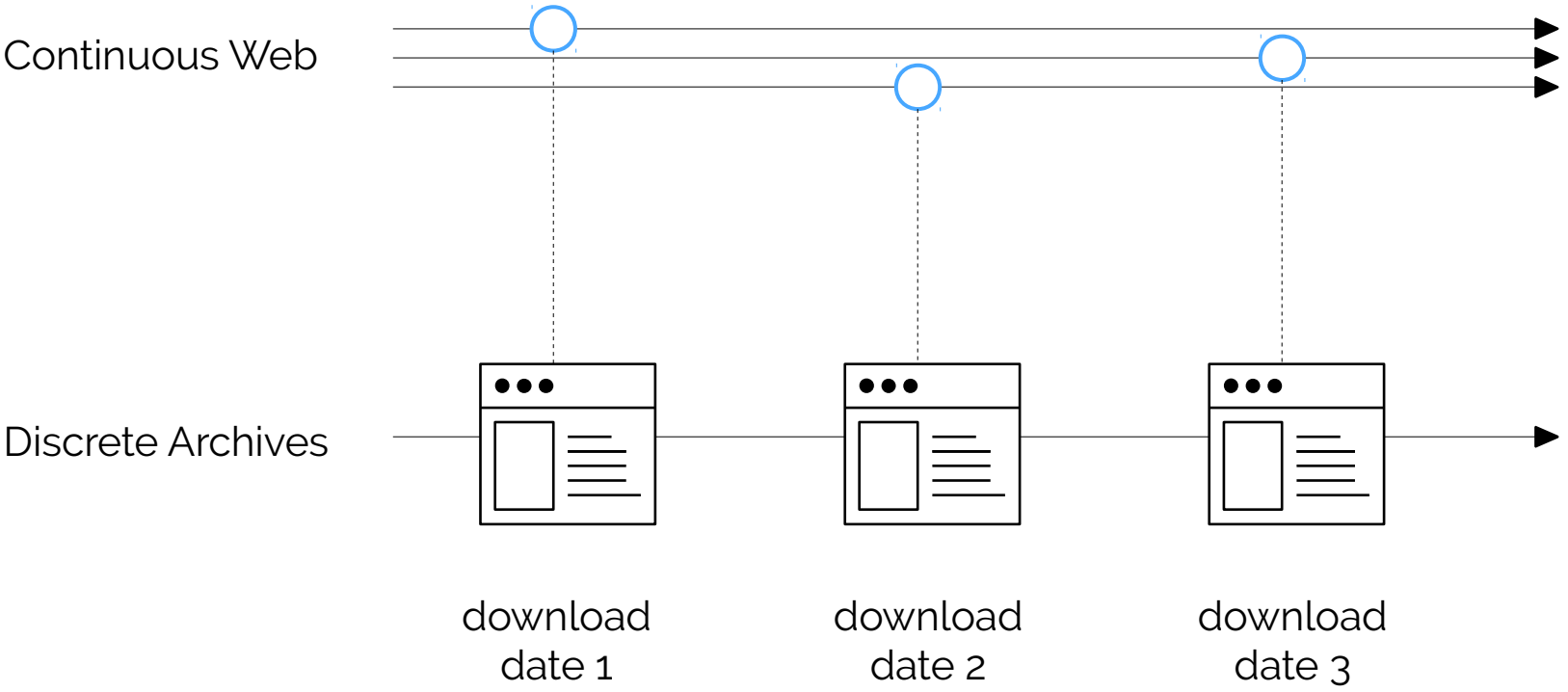
14H30 Affrontement près de Kelaât Es-Sraghna suite aux limogeages de cadres d'Al Adl Wal Ihsane

13H36 Marrakech : Incendie maitrisé à l'Hôtel La Mamounia

12H50 L'abattage rituel sans étourdissement interdit en Wallonie dès septembre 2019

Web archives are not direct traces of the web

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



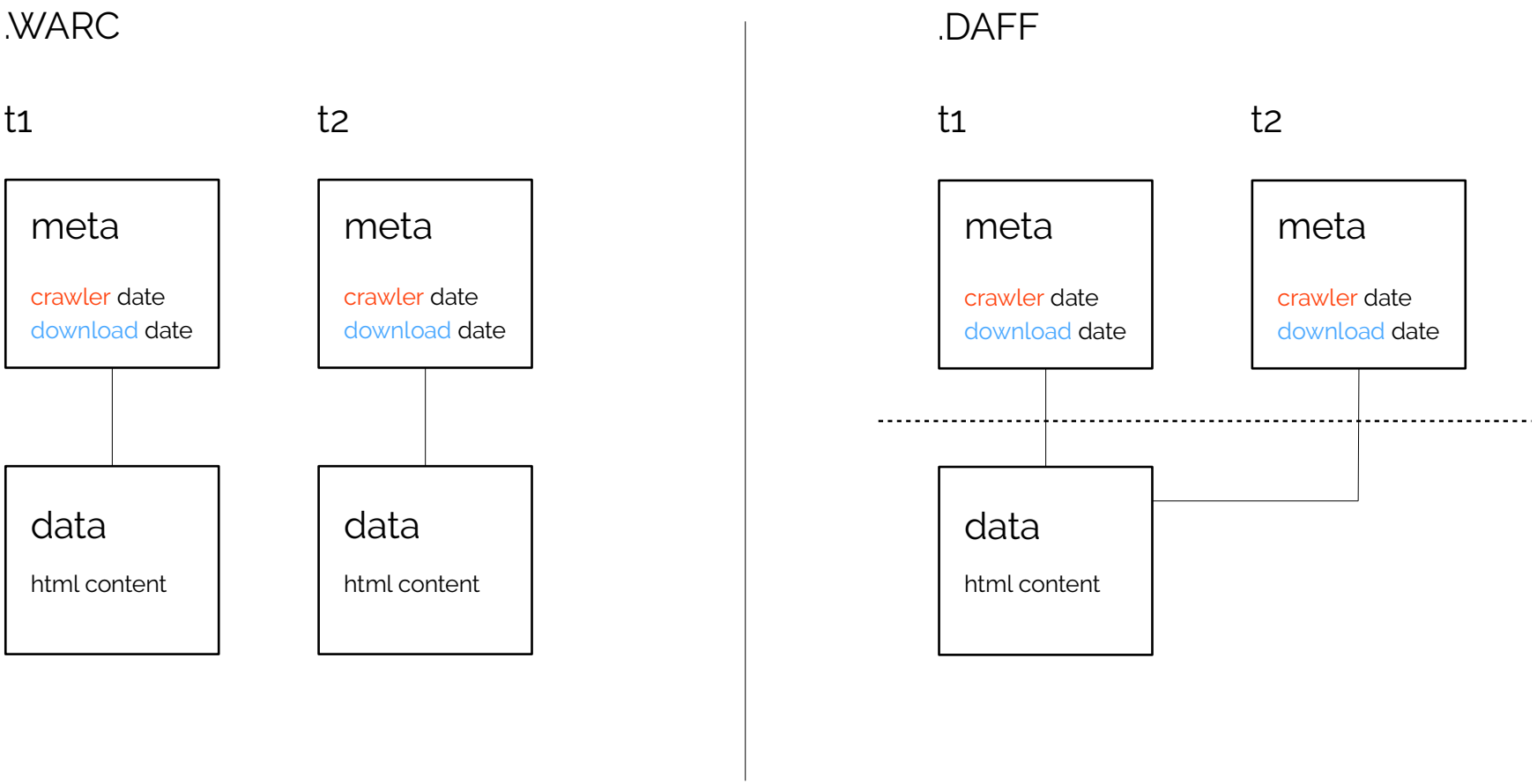
> We saw what we call a crawl legacy effect

To avoid the crawl legacy effect

We propose to conduct an exploratory analysis of web archives which would go beyond the level of the webpages

The original scale of web archives is the webpage

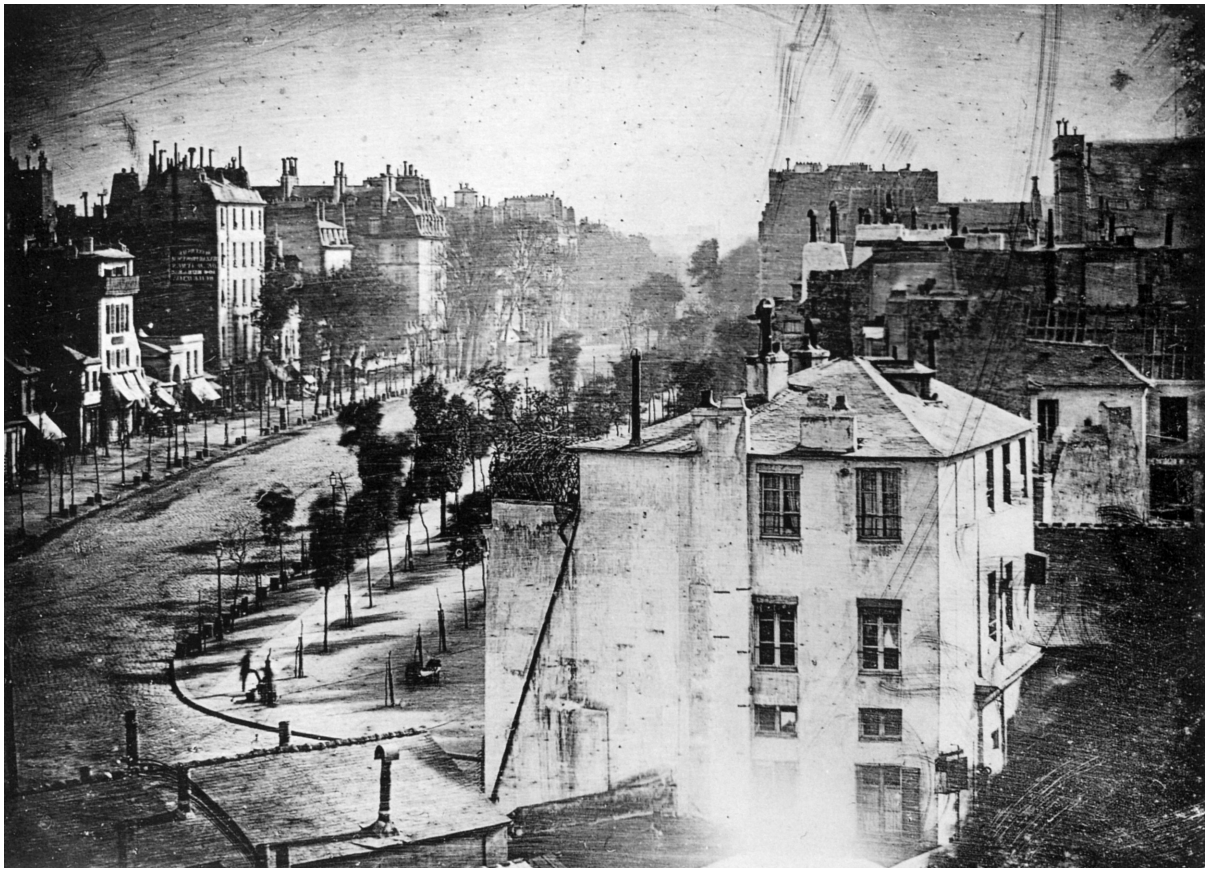
> what can we learn from the structure of web archives files?



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

Archiving is all about selecting and destroying

> as webpages change over time



> "Boulevard du Temple", Louis Daguerre, 1838

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

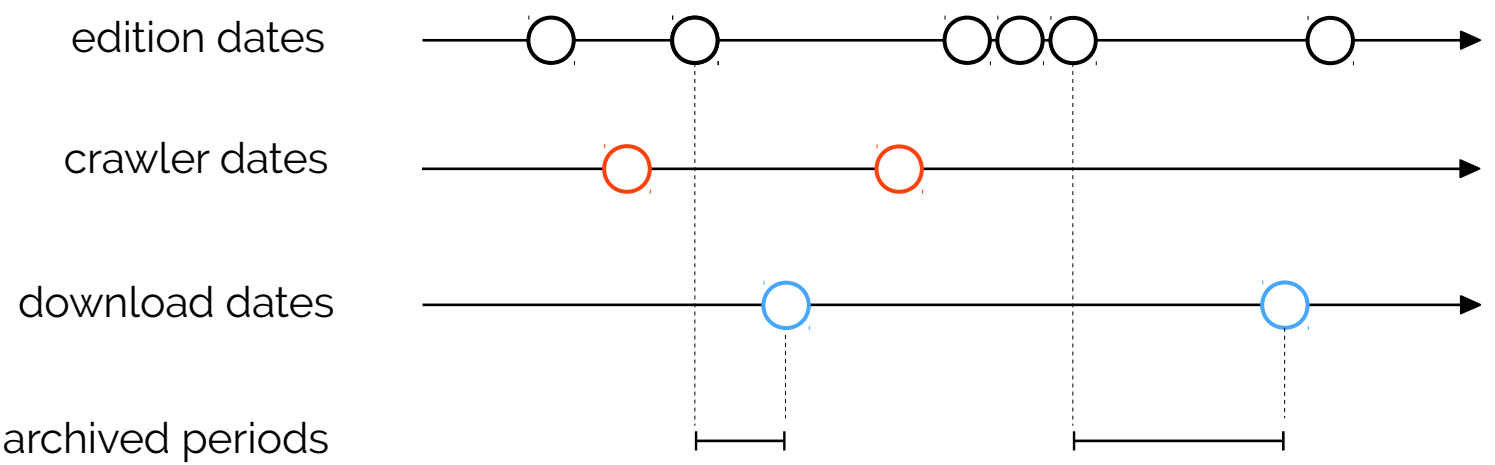
> attribute changes
css, font ...

> type changes
 to <p>

> semantic changes

Archiving on top of webpages goes with many challenges

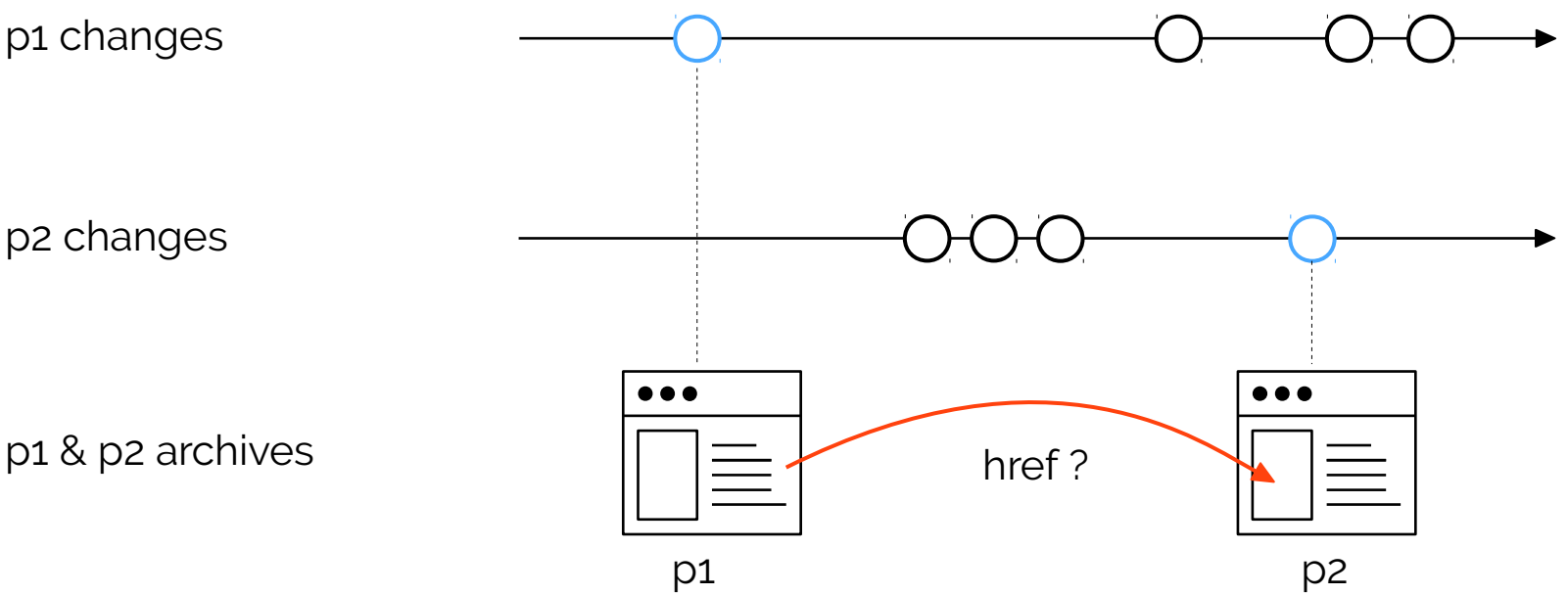
> Crawler blindness and archive quality



> Web archiving goes with construction locks

Archiving on top of webpages goes with many challenges

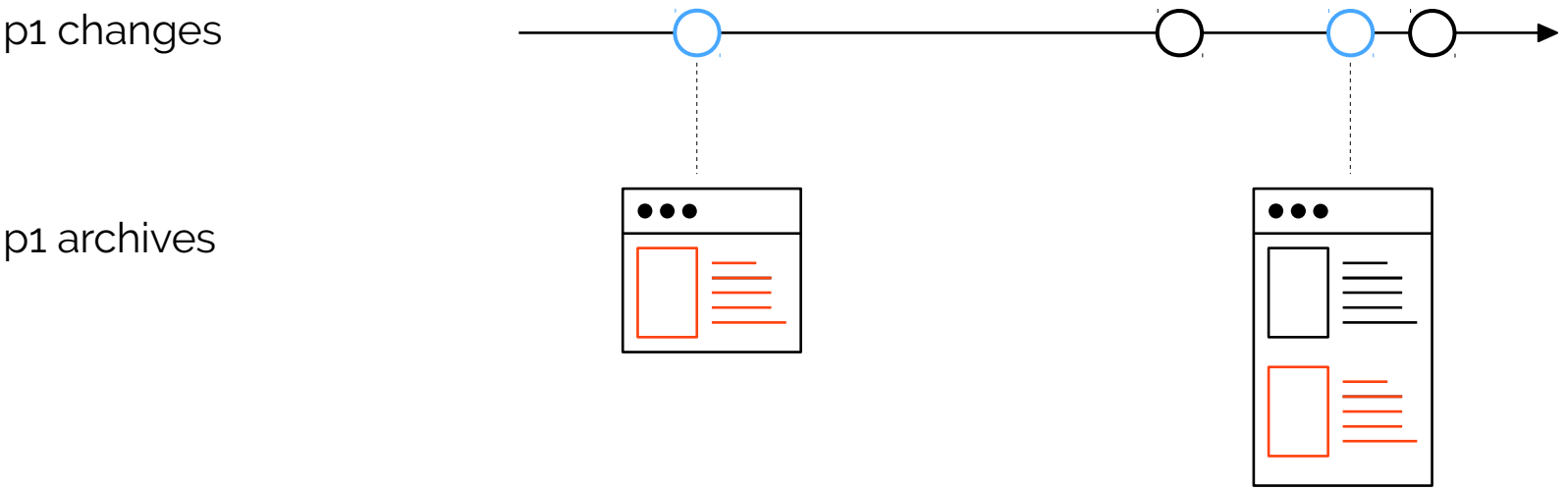
> Archive consistency across pages



> Web archiving goes with navigation locks

Archiving on top of webpages goes with many challenges

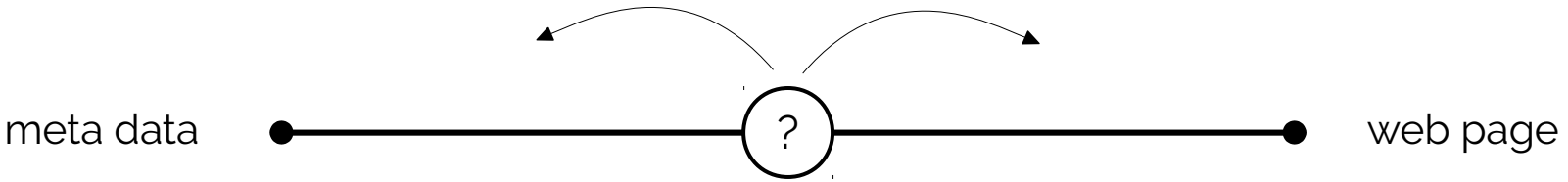
> Pages with archive-like content



> Archiving goes with discrete and continuous interpretation locks

To face or reduce these challenges

We propose to build a new entity from based on web archives called
web fragments



The web fragment

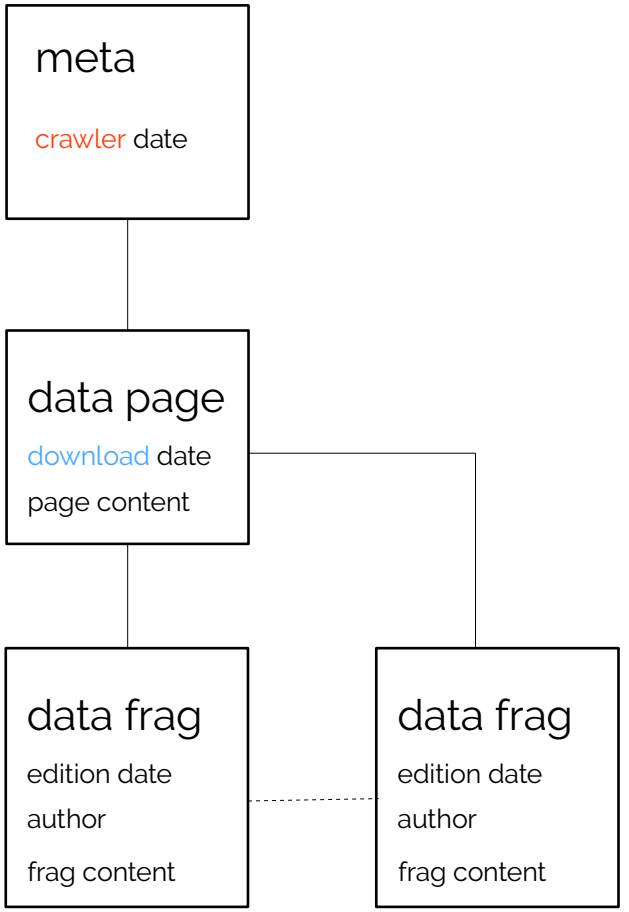
> A structured part of a webpage with high informationing contents

an article

a news item

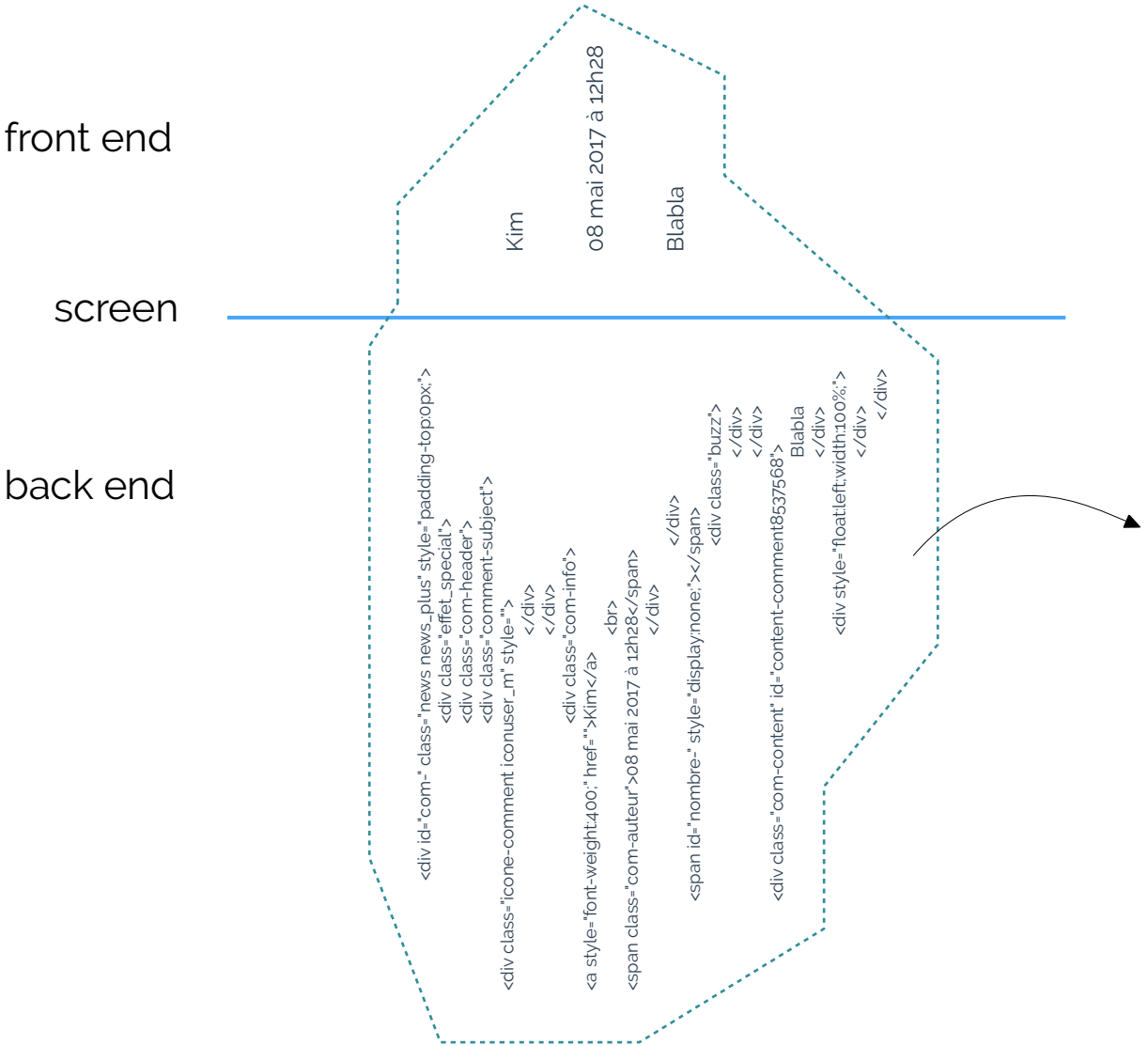
a comment

> New structure for web archives



Finding web fragments

> We must see a webpage as a front & back end object



Related works :
- a flat-file
- or an unordered tree
- or an ordered tree

Finding web fragments

> A webpage is a 2D hierarchical list of HTML nodes

> Nodes are categorized among : title, author, date and text

Finding web fragments

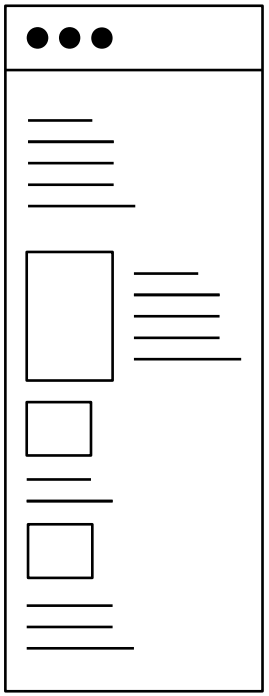
> Nodes are selected based on markup & class & id using regex

<h1 id = "title" class = "title_comment"> Hello archives </h1>

> Nodes are incrementally grouped into web fragments using ad-hoc rules

[U text] or [text U _text] or [title U text] or [date U _text] or [author U date] ...

> Algorithm



1, Select nodes in DOM

text
text
title
author
text
author
date
text
author
date
text

2, Group in fragments

[text text]

[title author text]

[author date text]

[author date text]

3, Group by list of fragments

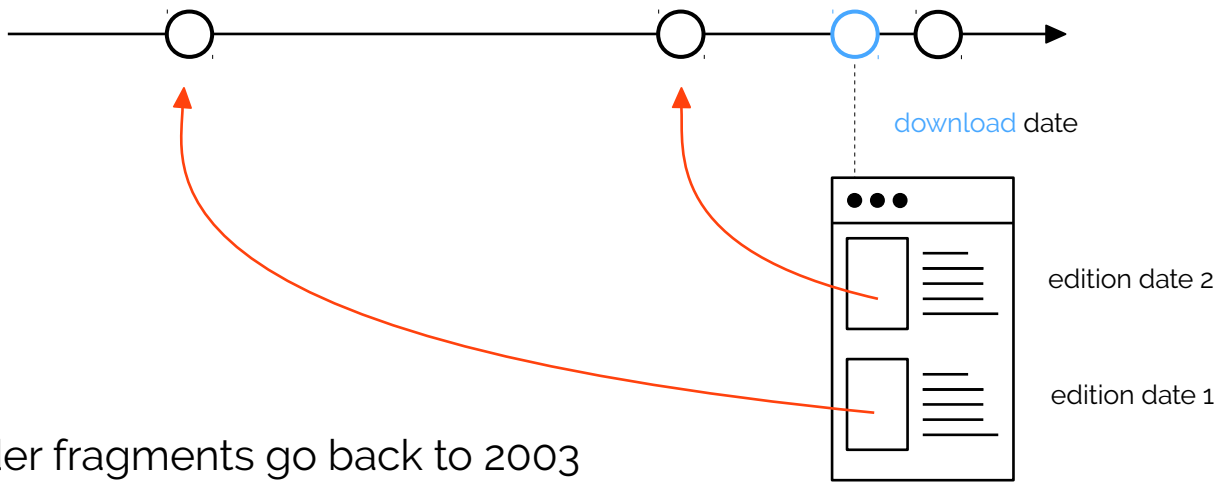
[text text]

[title author text]

[author date text], [author date text]

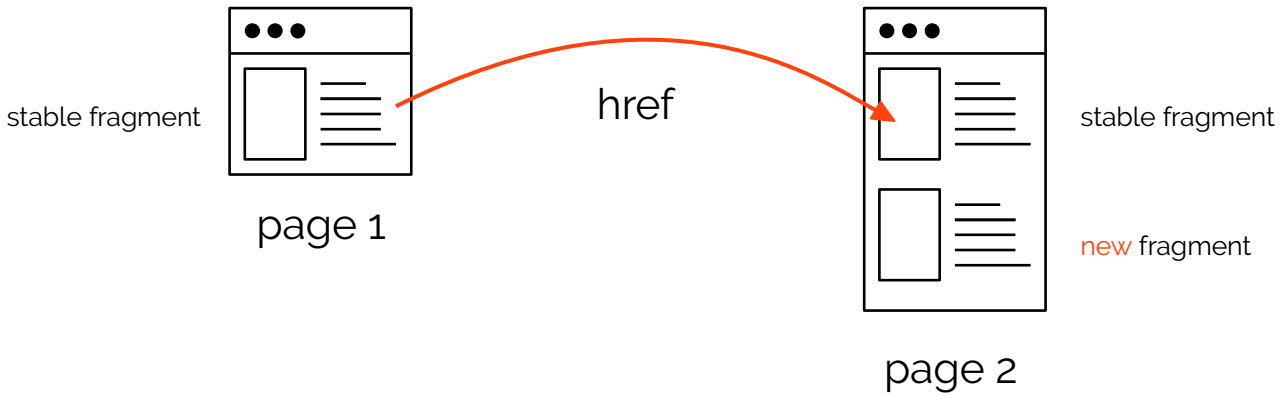
Rethinking archive challenges using web fragments

> Crawler blindness can be reduced and archive quality increased



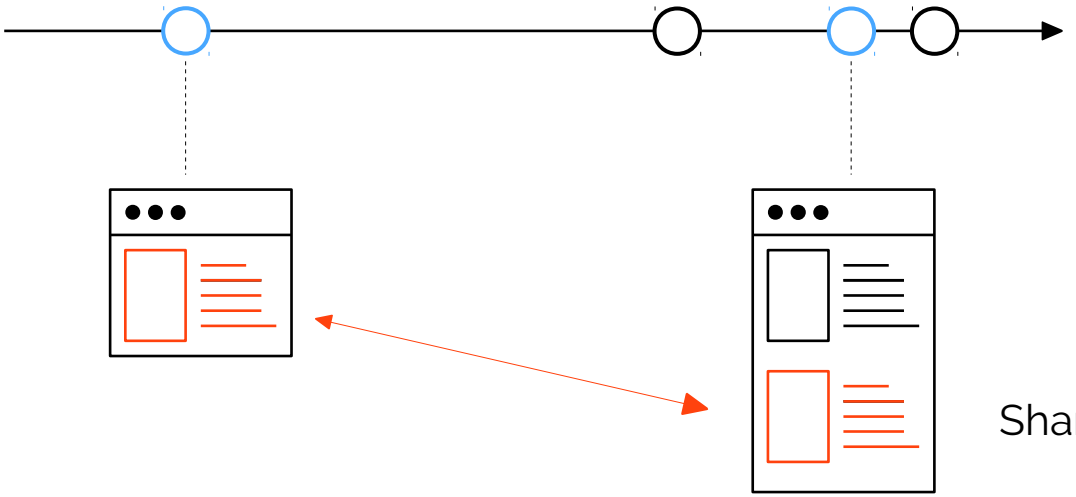
Yabiladi's older fragments go back to 2003

> 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)

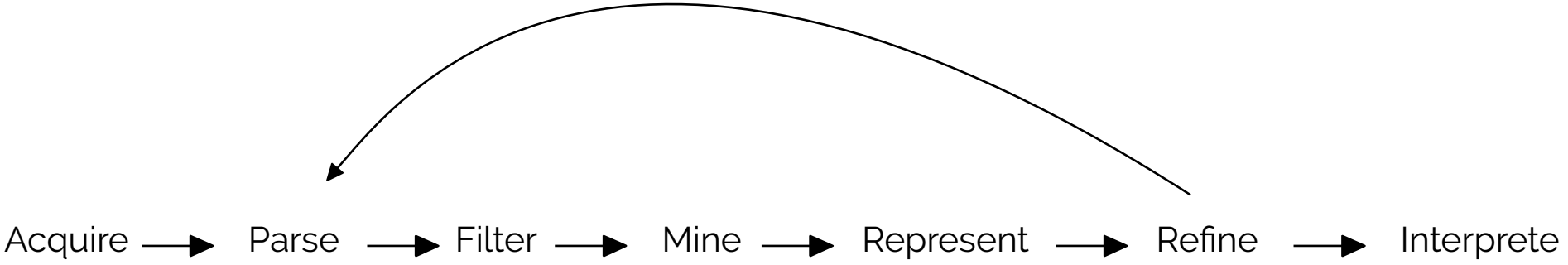
> Web fragments help us expanding web archives beyond web pages

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

Exploratory analysis of Web archives

> Following John Wilder Tukey's work

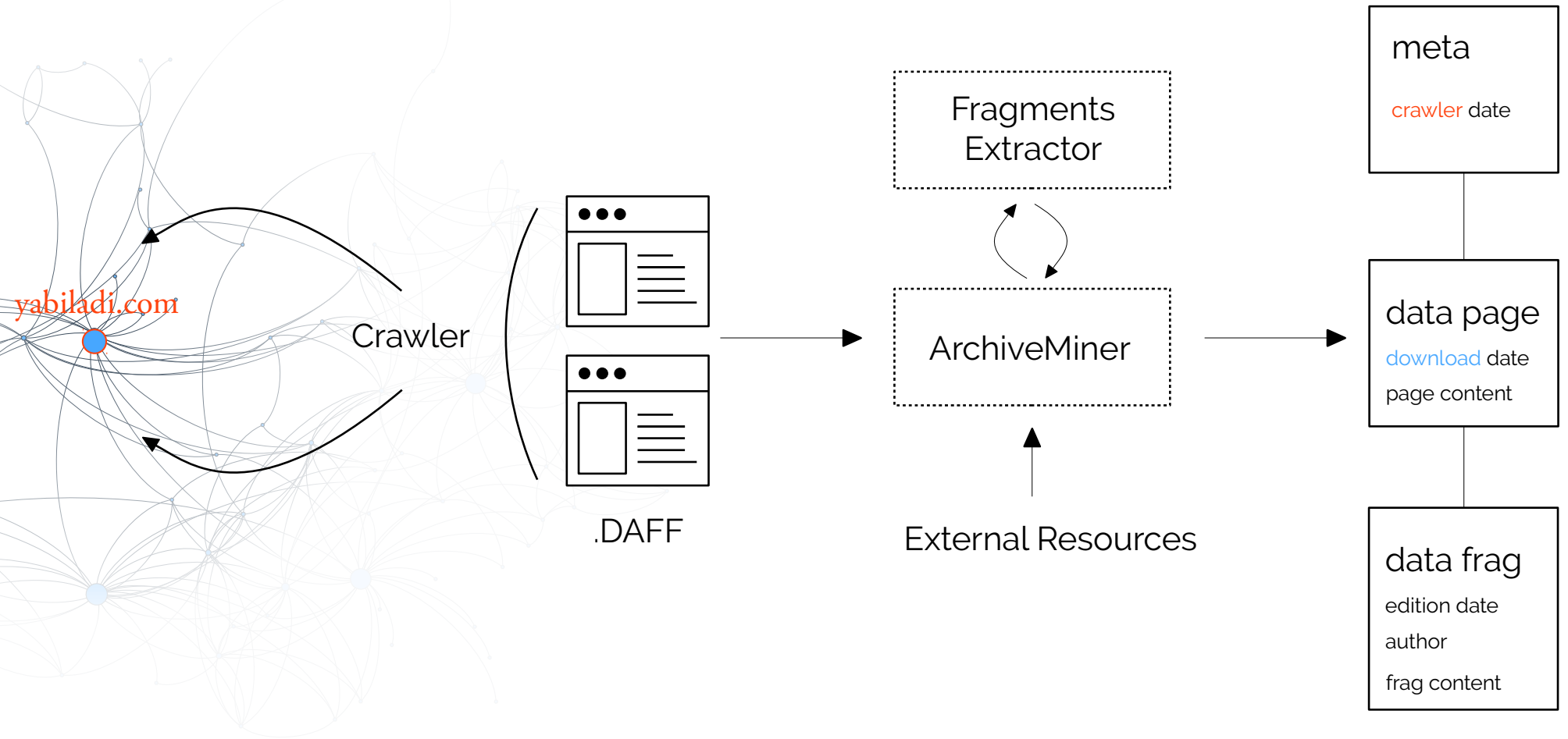
An iterative process that is deliberately part of a logic of observation, discovery and astonishment



Archives extraction engine

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

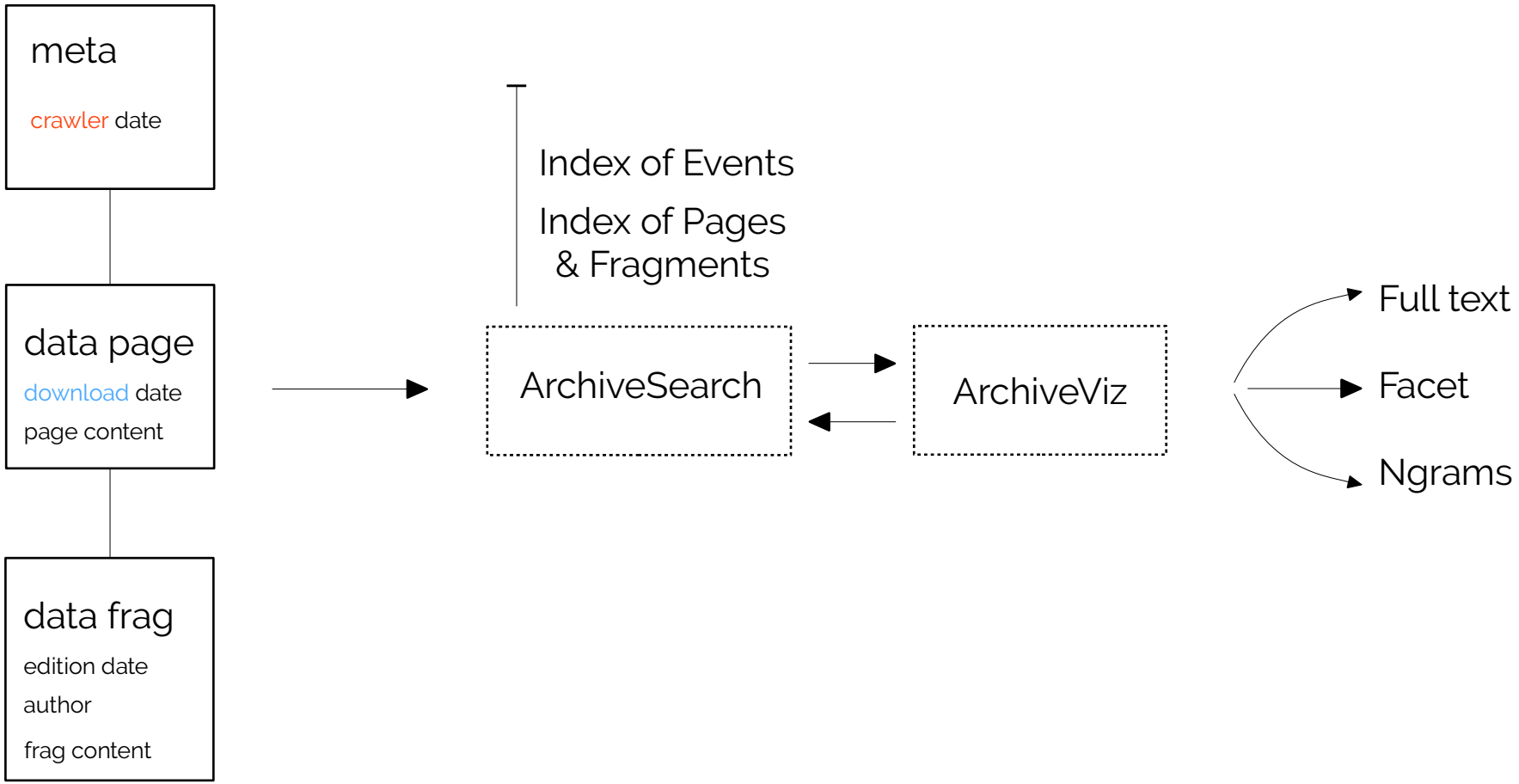
> The Web Archives Explorer (part 1)



Archives exploration engine

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

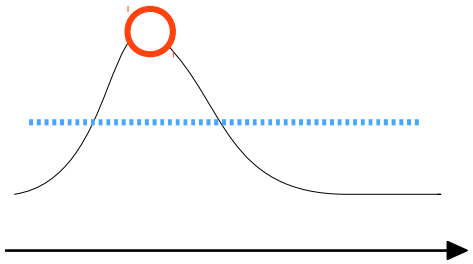
> The Web Archives Explorer (part 2)



The validation of 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

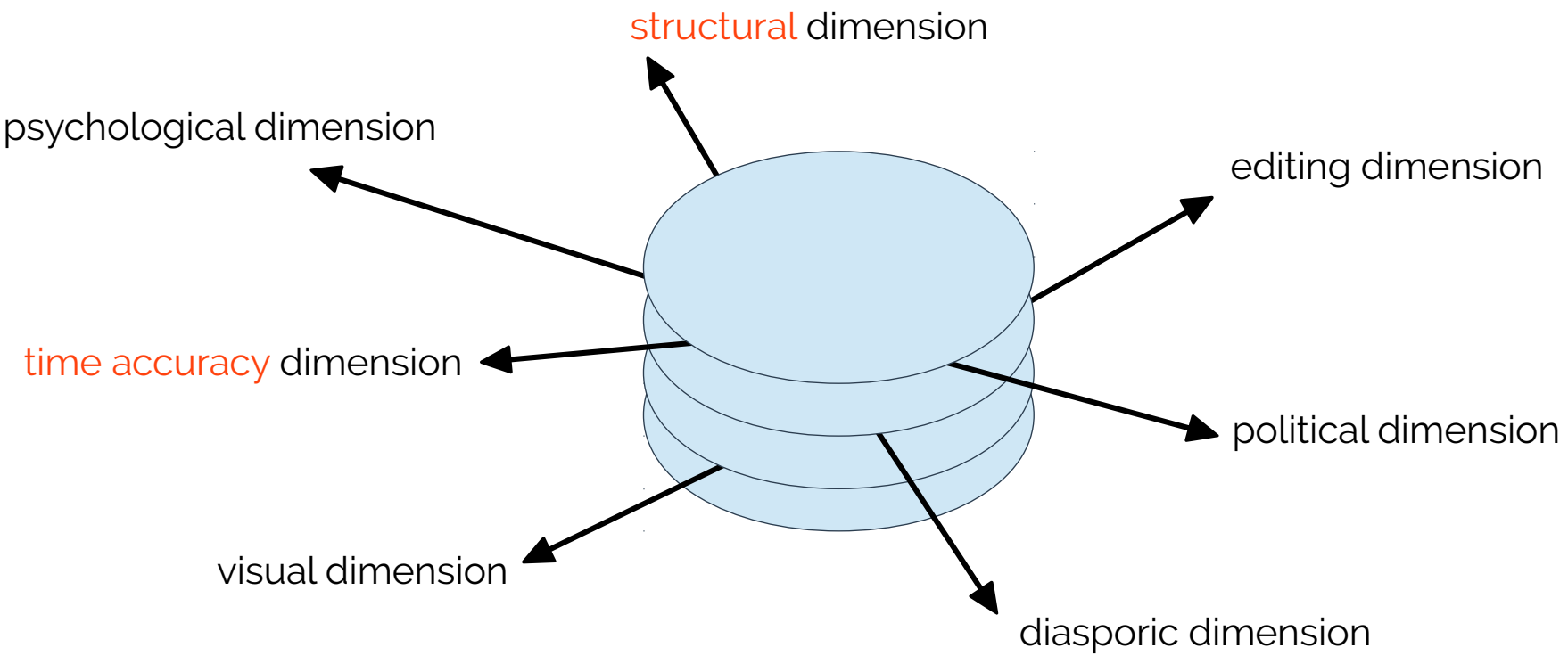
3. fields and experts interpretations

> Let's see the Web Archives Explorer in action

video presentation for CIKM2017

Going deeper through the definition of web fragment

> A more abstract & pluridisciplinar characterization of web fragments



> More validation process based on thematical workshops (such as event detection) and field interpretations

Hard work in progress here !

Reading :
- Information extraction
- digital edition
- history of the web
- the concept of Rhizome developed by Deleuze and Guattari

Thank you!
Questions?