SciencesPo. | médialab

Websites as sources

20 Mars 2011

The Hypertext Corpus Initiative

methods and tools for Social Sciences to build corpus from the web

Paul Girard, Mathieu Jacomy, Audrey Baneyx, Tommaso Venturini

What is actually a website ?

Or why we need to define web entities ?

From sources to corpus ?

a method to select and collect sources to build a corpus

What for ?

Analysis opportunities and limitations
Interactions with web archive initiative
From tools to equipment

WWW : a network of ressources

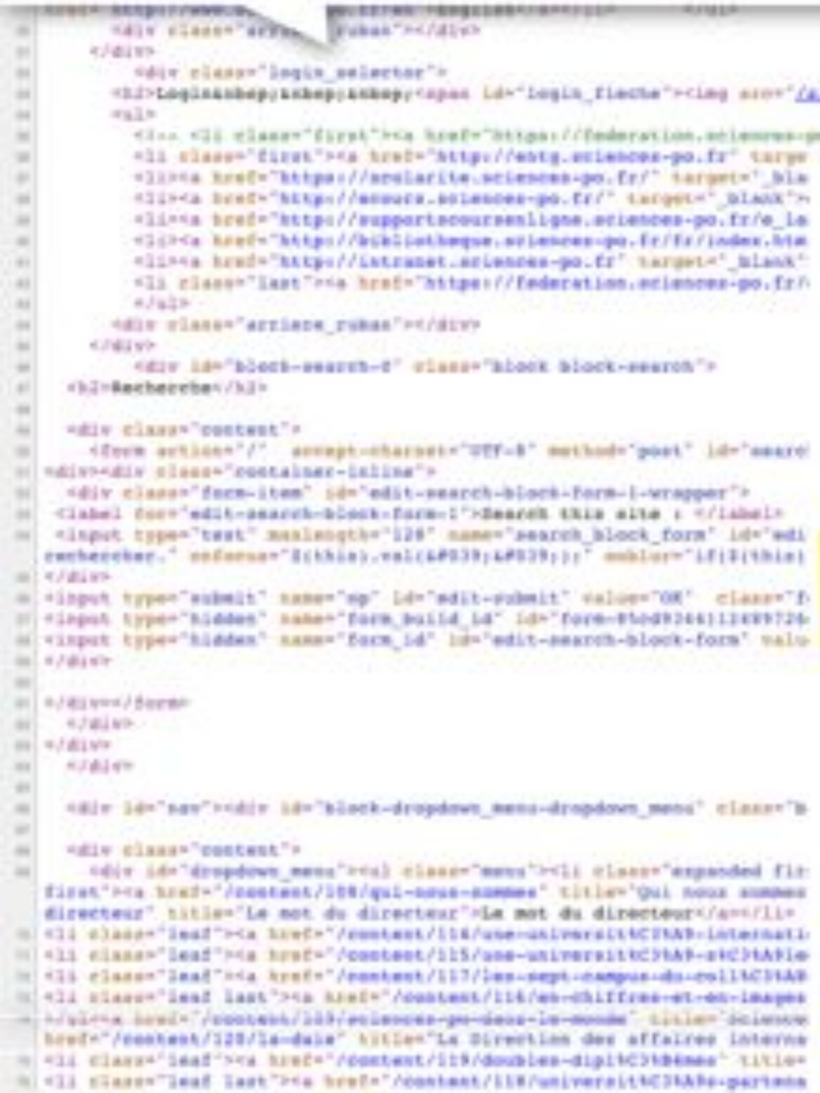
Luc Viatour © GFDL
www.lucnix.be

Uniform Ressource Locator
Hypertext REFerence

Nick Finck @ flickr

A network of references

A HREF



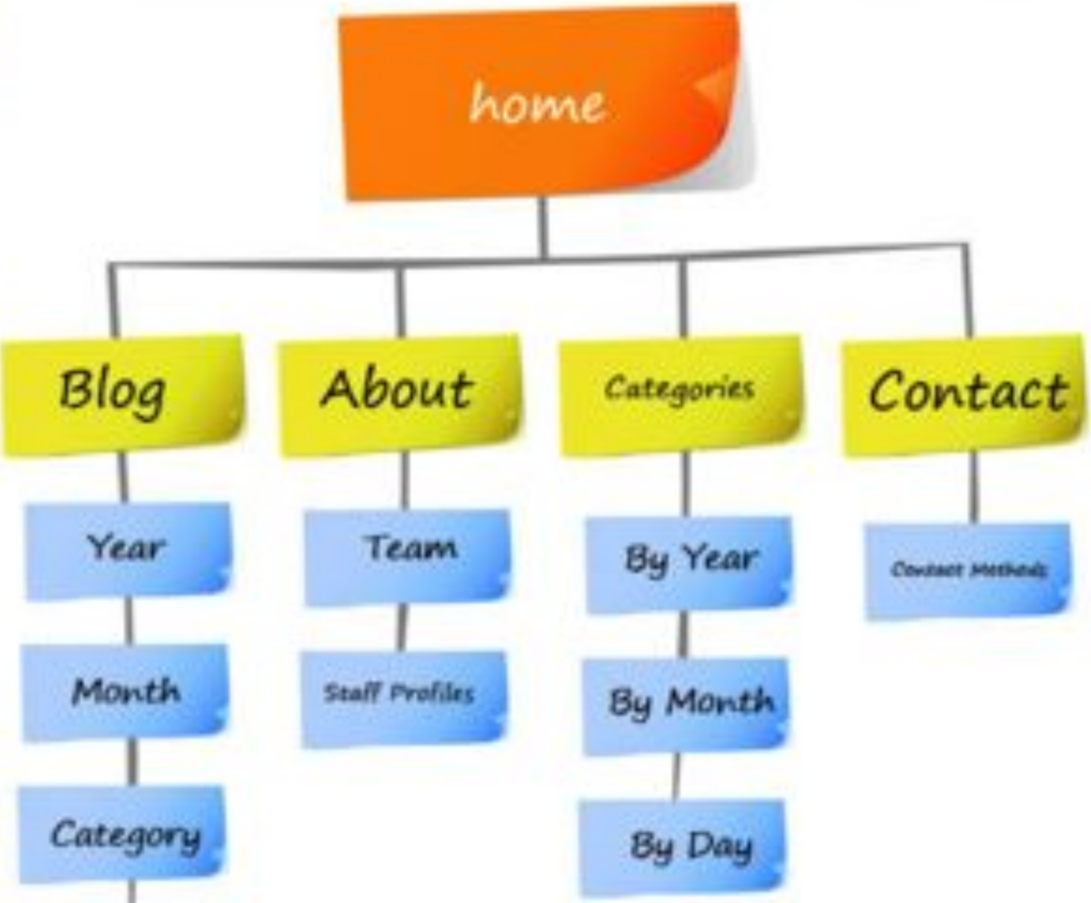
```
href="http://www.s        po.fr/en">English</a></li>        </ul>
        <div class="arri      ruban"></div>
    </div>
        <div class="login_selector">
      <h2>Login&nbsp;&nbsp;&nbsp;<span id="login_fleche"><img src="/sites/all/themes/sciences_po/images/icon_flechedroite.png" alt="fleche_droite"/></span></h2>
      <ul>
        <!-- <li class="first"><a href="https://federation.sciences-po.fr/cas/login?service=http%3A%2F%2Fwww.sciencespo.fr%2Fcas" target="_blank">Portail</a></li> -->
        <li class="first"><a href="http://entg.sciences-po.fr" targe
        <li><a href="https://scolarite.sciences-po.fr/" target="_bla
        <li><a href="http://moodle.sciences-po.fr/" target="_blank">
        <li><a href="https://supportscoursenligne.sciences-po.fr/e_la
        <li><a href="http://bibliotheque.sciences-po.fr/fr/index.htm
        <li><a href="http://intranet.sciences-po.fr" target="_blank"
        <li class="last"><a href="https://federation.sciences-po.fr/
        </ul>
      <div class="arriere_ruban"></div>
    </div>
        <div id="block-search-0" class="block block-search">
  <h2>Recherche</h2>

  <div class="content">
    <form action="/"  accept-charset="UTF-8" method="post" id="searc
<div><div class="container-inline">
  <div class="form-item" id="edit-search-block-form-1-wrapper">
 <label for="edit-search-block-form-1">Search this site : </label>
 <input type="text" maxlength="128" name="search_block_form" id="edi
rechercher." onfocus="$(this).val(&#039;&#039;);" onblur="if($(this)
</div>
<input type="submit" name="op" id="edit-submit" value="OK"  class="f
<input type="hidden" name="form_build_id" id="form-8Ved9366111689726
<input type="hidden" name="form_id" id="edit-search-block-form" valu
</div>

</div></form>
  </div>
</div>
  </div>

  <div id="nav"><div id="block-dropdown_menu-dropdown_menu" class="b

  <div class="content">
    <div id="dropdown_menu"><ul class="menu"><li class="expanded fir
first"><a href="/content/108/qui-nous-sommes" title="Qui nous sommes
directeur" title="Le mot du directeur">Le mot du directeur</a></li>
<li class="leaf"><a href="/content/116/une-universit%C3%A9-internati
<li class="leaf"><a href="/content/115/une-universit%C3%A9-s%C3%A9le
<li class="leaf"><a href="/content/117/les-sept-campus-du-coll%C3%A8
<li class="leaf last"><a href="/content/114/en-chiffres-et-en-images
</ul><a href="/content/103/sciences-po-dans-le-monde" title="Science
href="/content/120/la-dai" title="La Direction des affaires interna
<li class="leaf"><a href="/content/119/doubles-dipl%C3%B4mes" title=
<li class="leaf last"><a href="/content/118/universit%C3%A9s-partena
```



Hierarchies of addresses

rfc3986

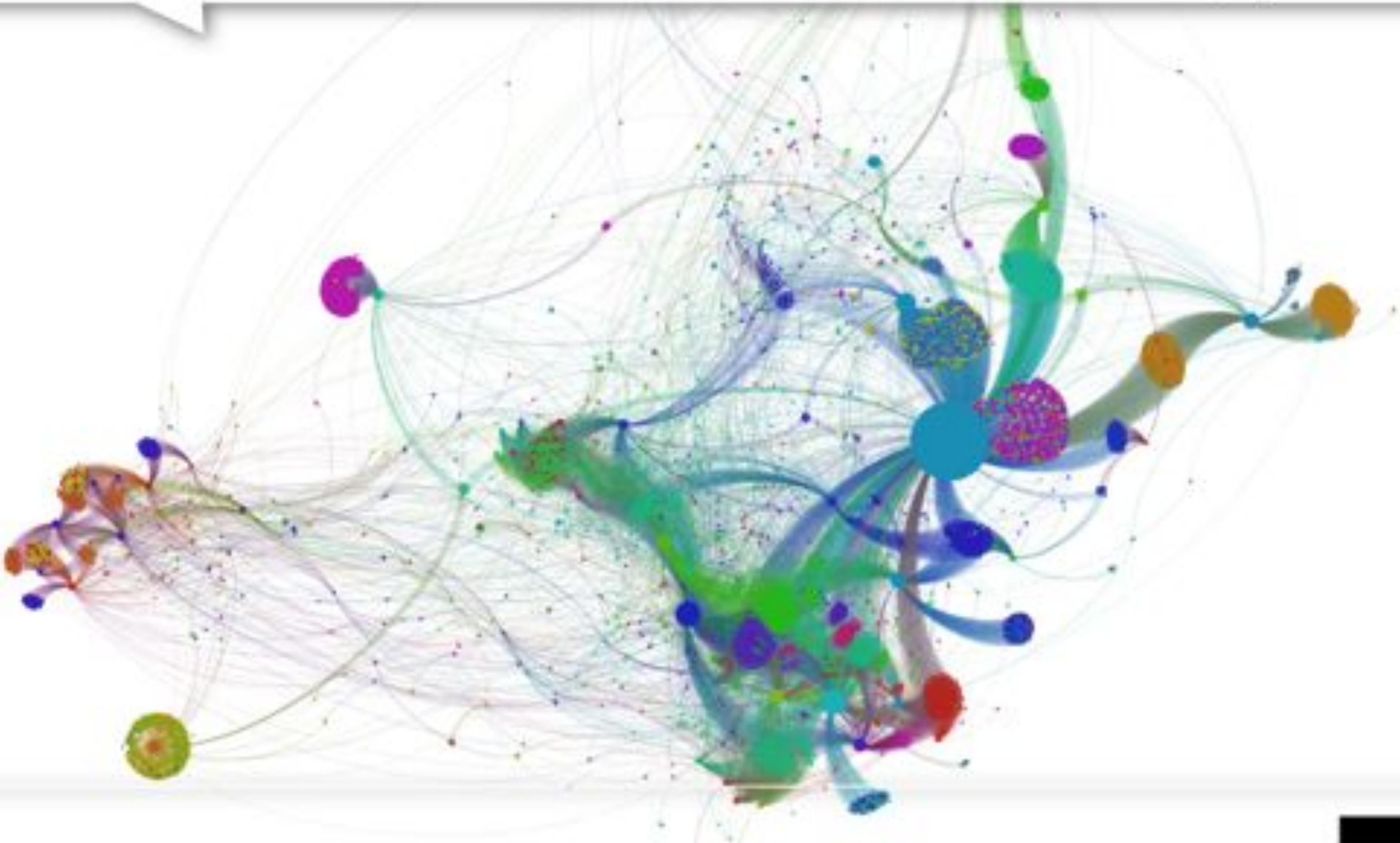
« The URI syntax is organized hierarchically […]

It is ***often*** the case that a group or "tree" of documents has been constructed to serve a common purpose, wherein the vast majority of URI references in these documents point to resources within the tree rather than outside it.

Similarly, documents located at a particular site are ***much more likely*** to refer to other resources at that site than to resources at remote sites. »

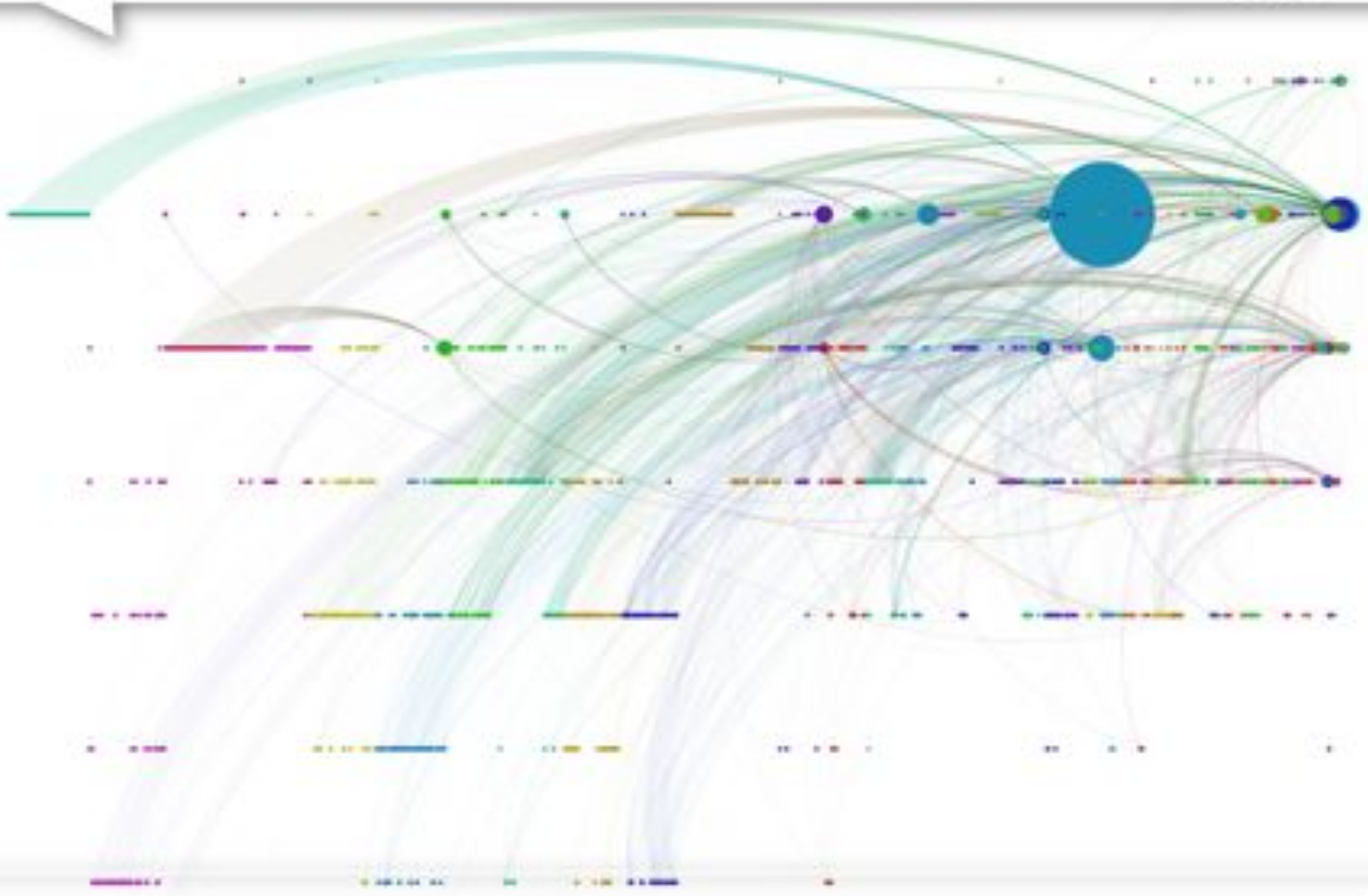
clusters of HTML pages

sciences-po.fr domain



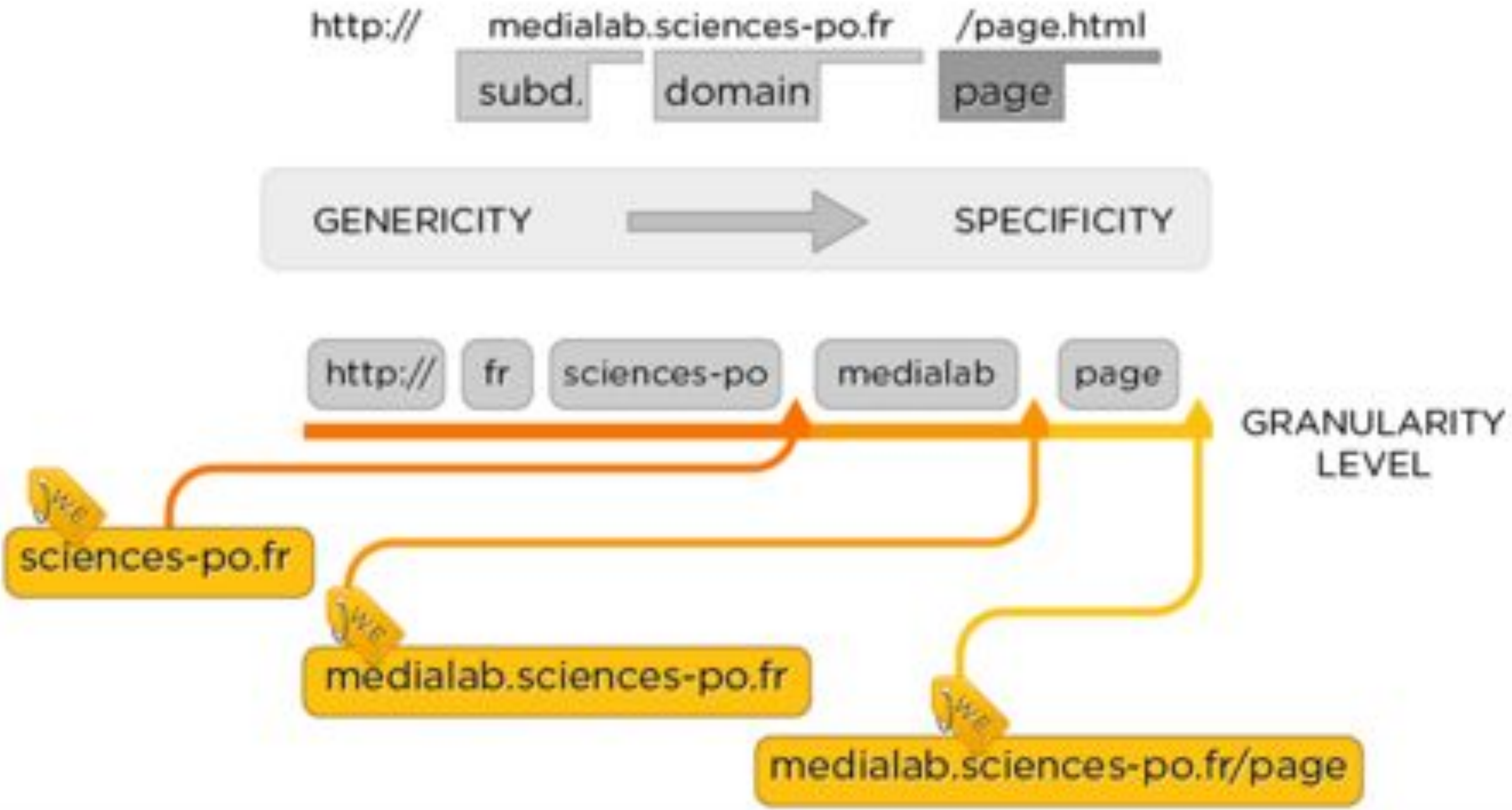
hierarchical namespace

sciences-po.fr domain



pointers in the URL hierarchy

web entities



pointers in the URL hierarchy

web entities

RFC3986 : about URI

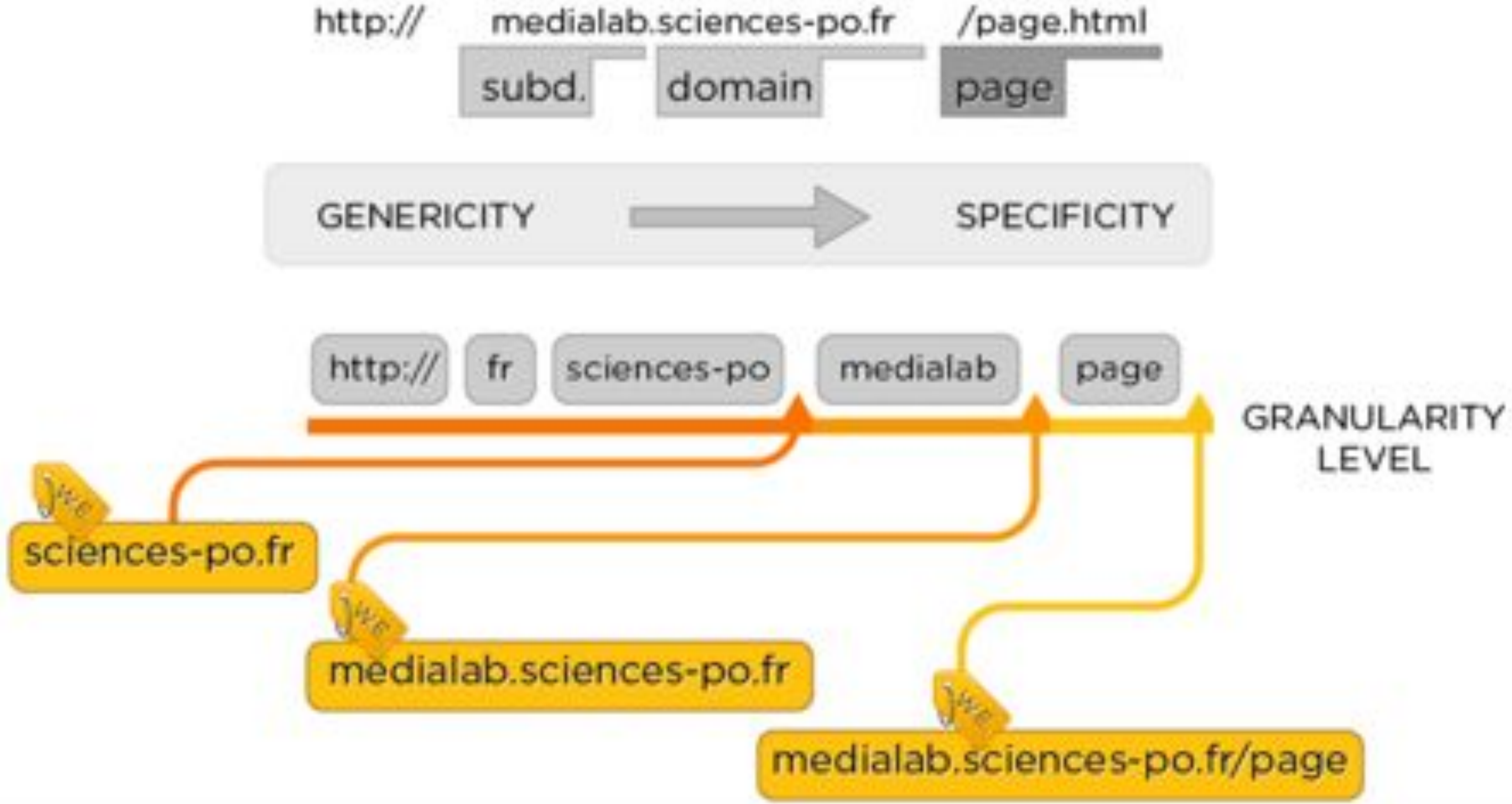
“ The URI syntax is organized hierarchically, with components listed in order of ***decreasing significance from left to right***.”

RFC882 : about domain names

“by convention, the labels that compose a domain name are read left to right, ***from the most specific (lowest) to the least specific (highest).***”

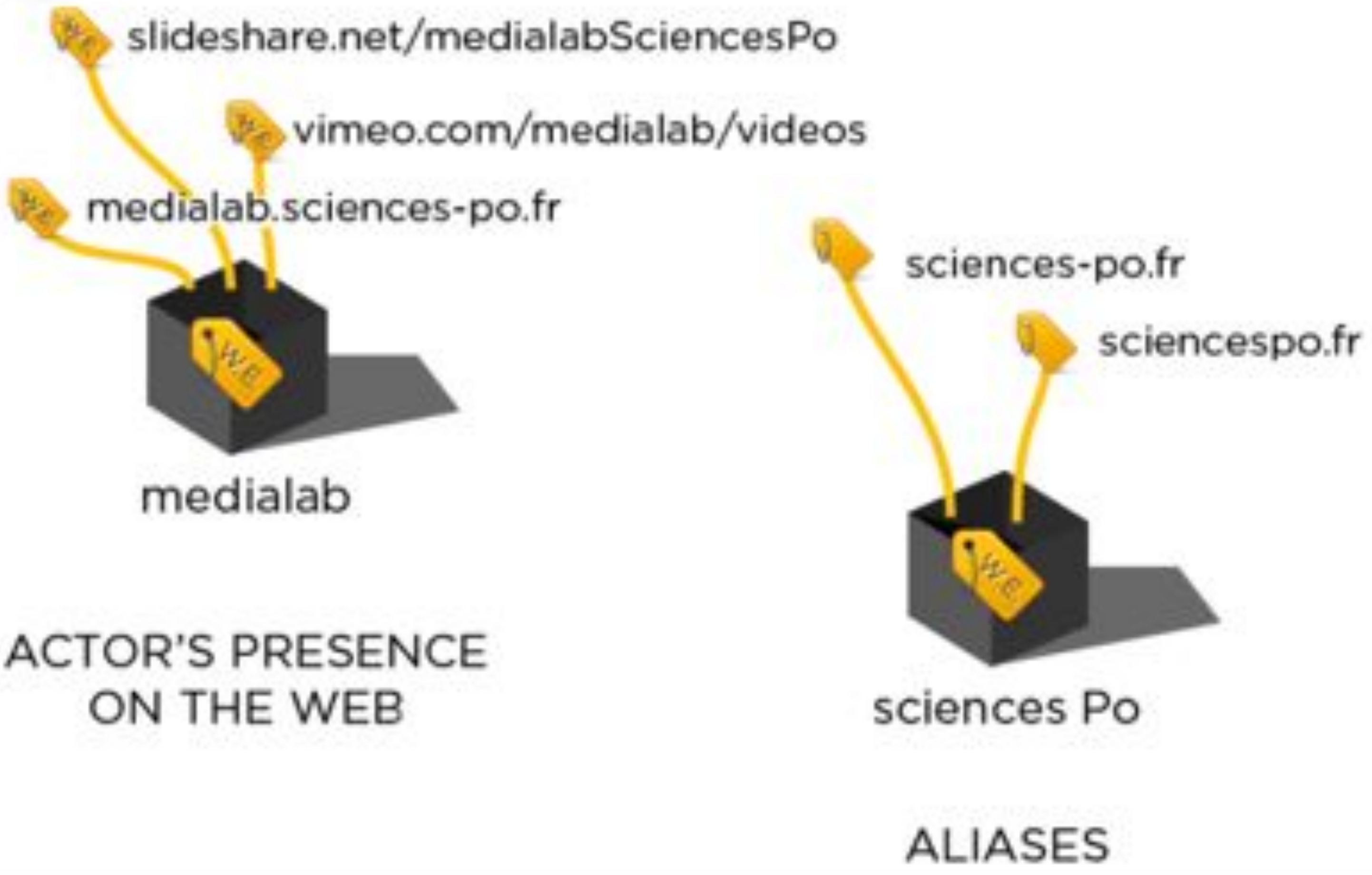
pointers in the URL hierarchy

web entities



handle heterogeneity : aliases

web entities



slideshare.net/medialabSciencesPo

vimeo.com/medialab/videos

medialab.sciences-po.fr

medialab

ACTOR'S PRESENCE ON THE WEB

sciences-po.fr

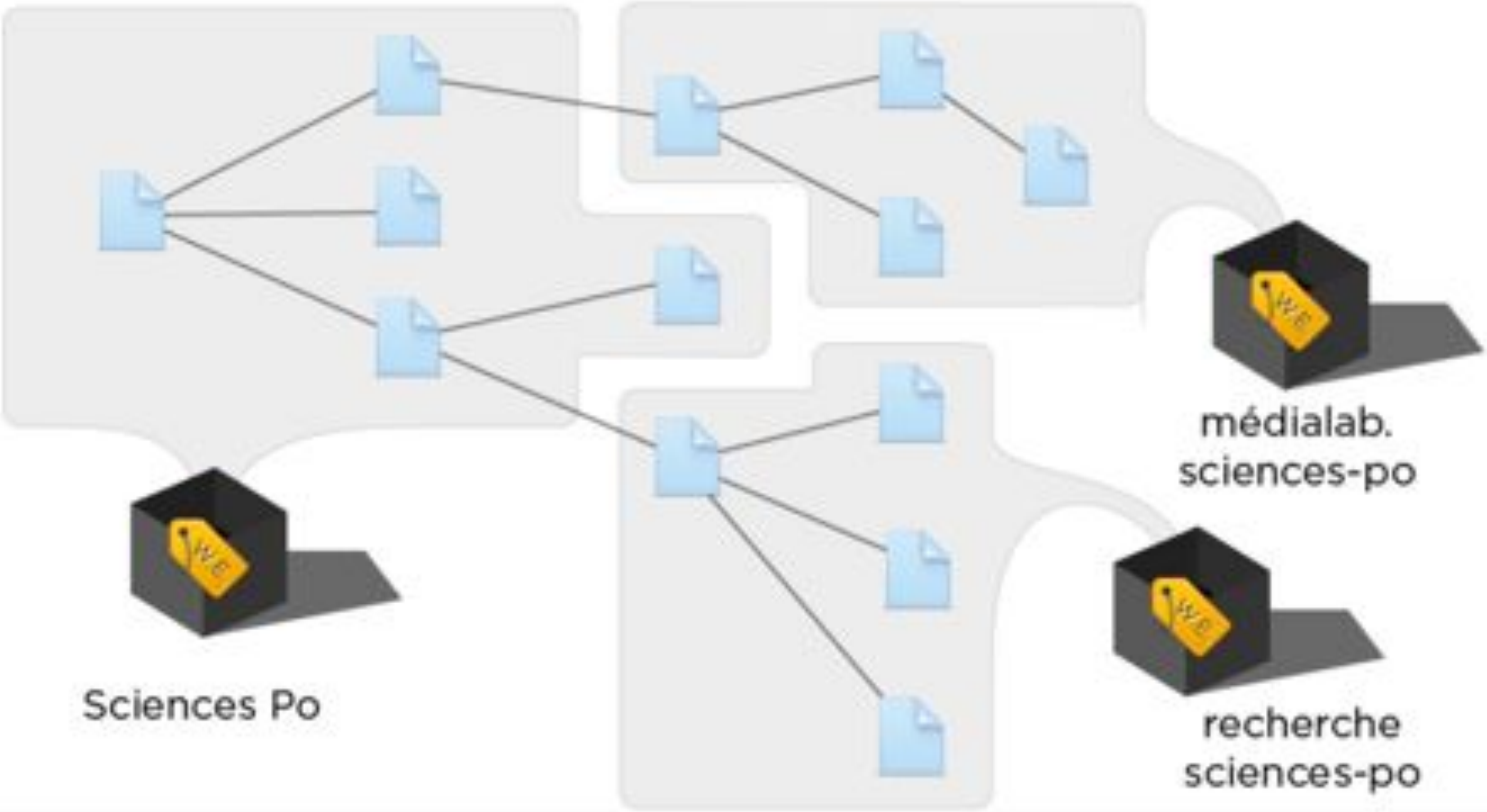
sciencespo.fr

sciences Po

ALIASES

maintain complexity

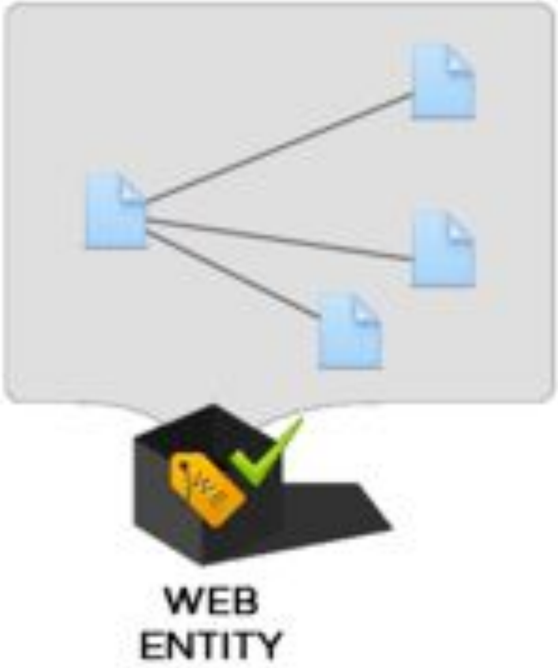
web entities



From sources to corpus

sourcing

Define



Select

status

✓ included

✗ excluded

? undefined

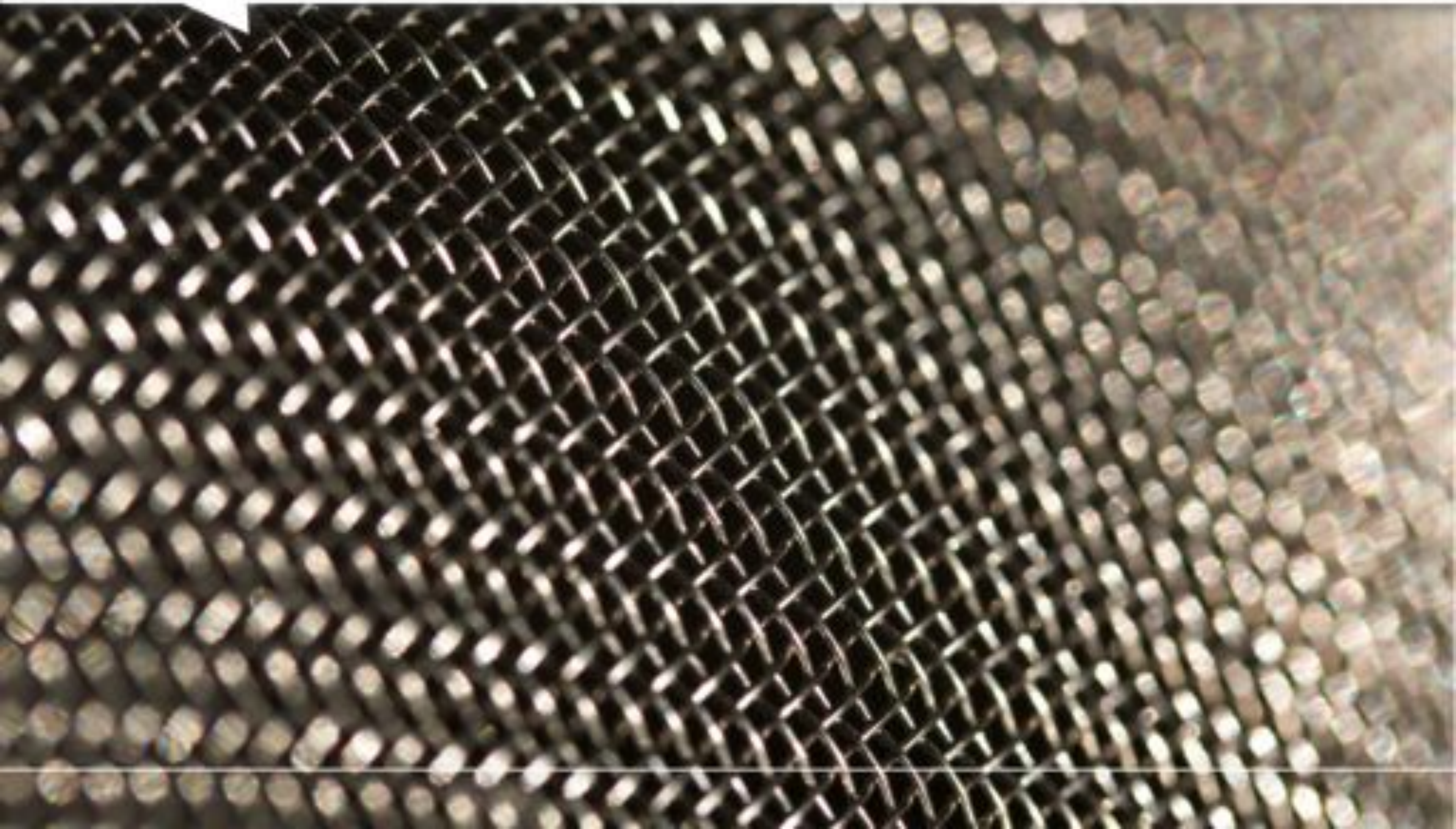
∅ unknown

Describe

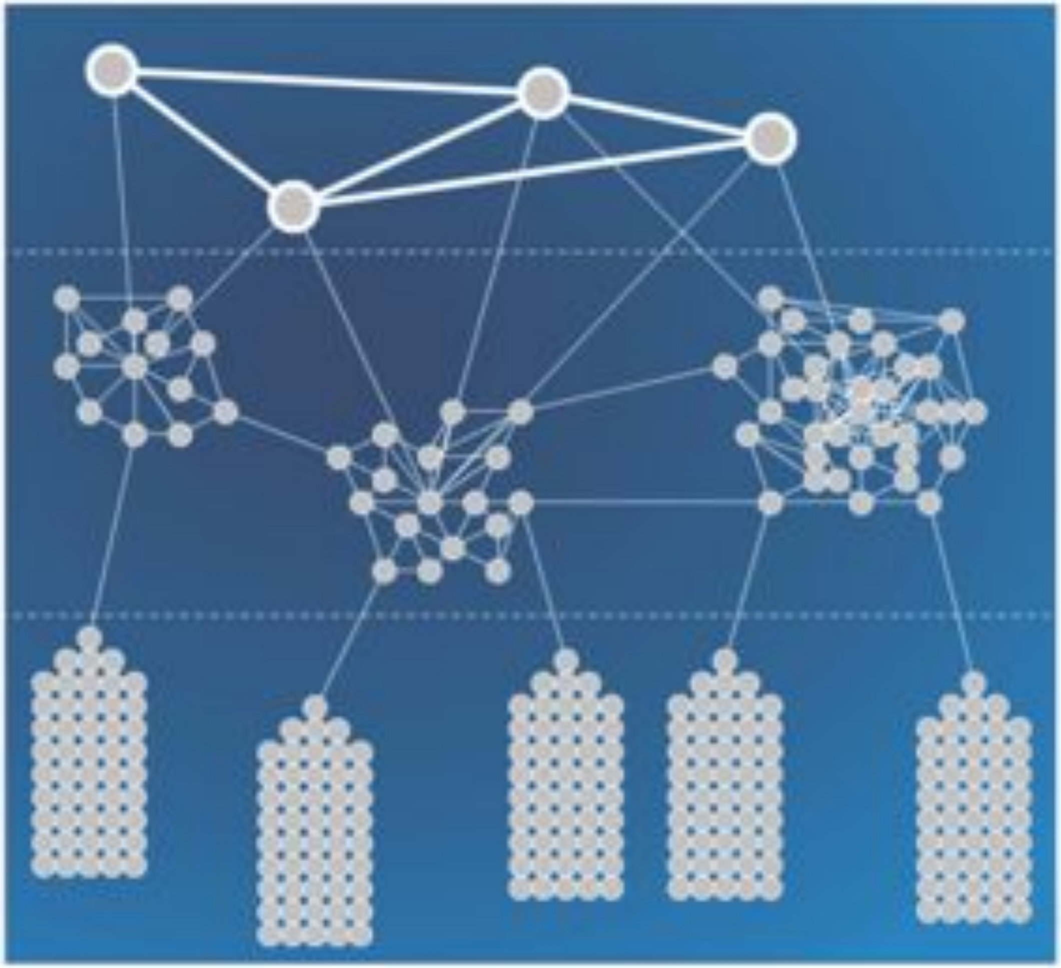
tagging

Sourcing : how to sieve the web ?

© Mayhem Chaos



WWWeb is an organized space



Couche **la plus visible** du web : Google, Amazon, Voila, SNCF, etc...

Couche intermédiaire : **agrégats**, communautés en ligne

Couche profonde : **bases de données**

Select : from sources to corpus

Sourcing : a qualitative task which gives entry points

- field enquiry, interviewing the actors
- use search engines and browse

But how to construct and validate a web corpus ?

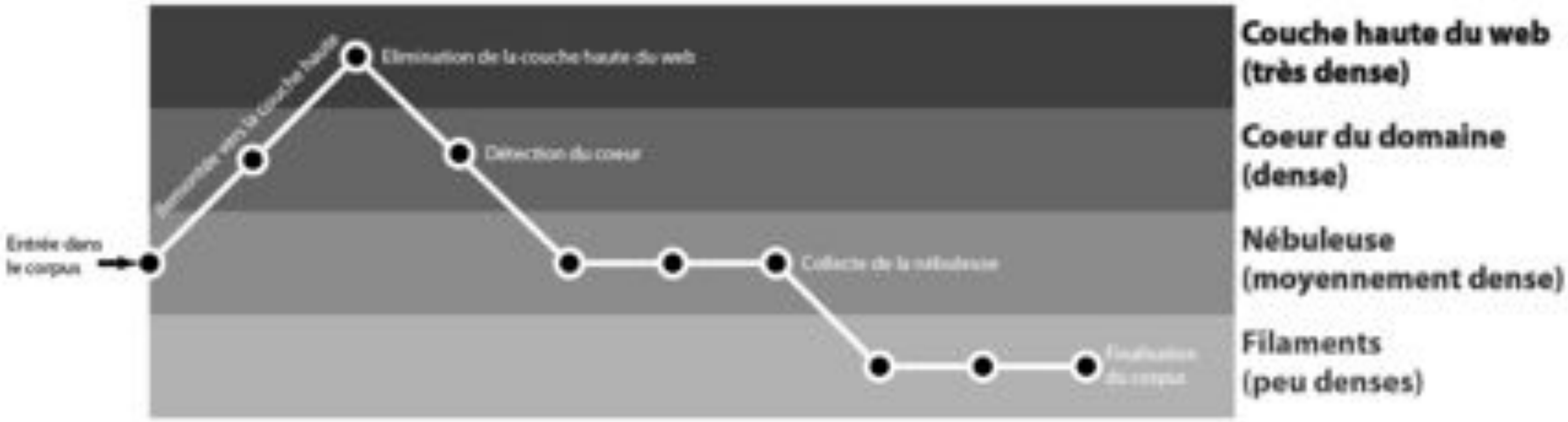
By following the medium...

Extending the corpus following links, peer to peer.

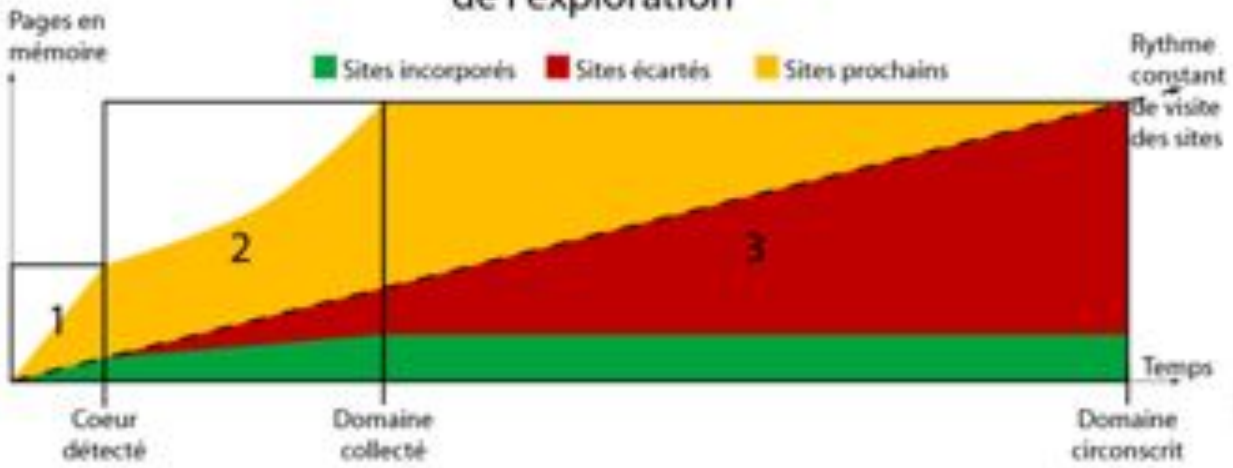
With prospective crawls.

Controlled by the researcher.

Explore and limit the corpus



Courbe d'avancement de l'exploration



Which crawler ?

WebAtlas Navicrawler

navicrawler,
to build a corpus manually

issuecrawler

issuecrawler,
to build a corpus with automatic crawling

a difficult choice

tweezers

? or

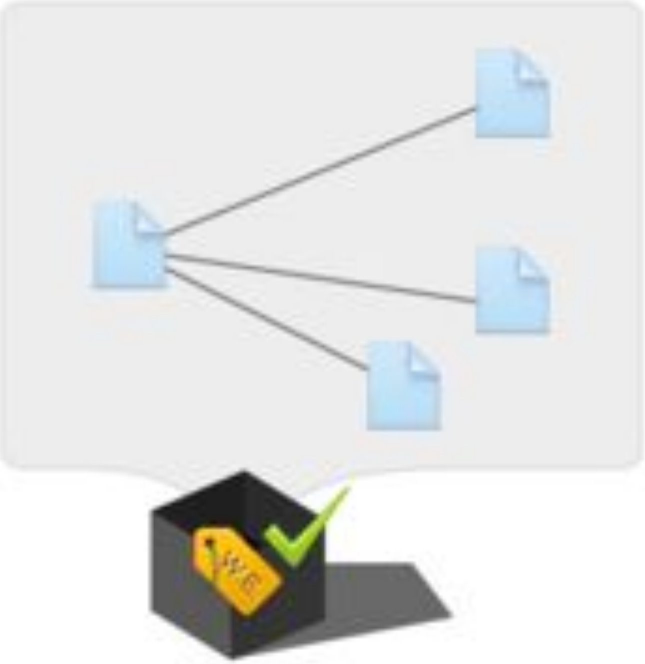
caterpillars

Research driven crawling

topic focused corpus?

1

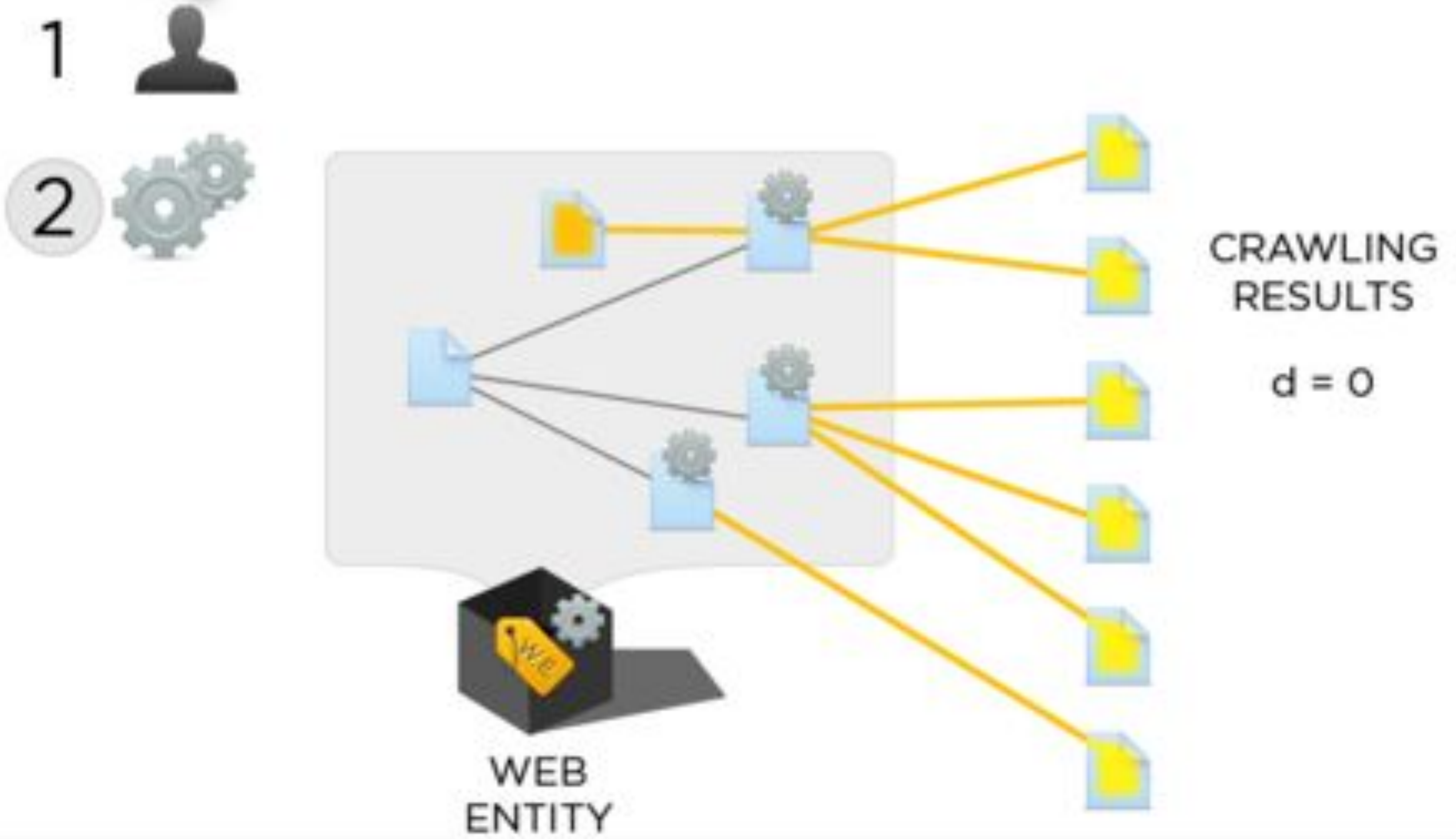
RESEARCHER SELECT *STARTING ENTITIES*



WEB ENTITY

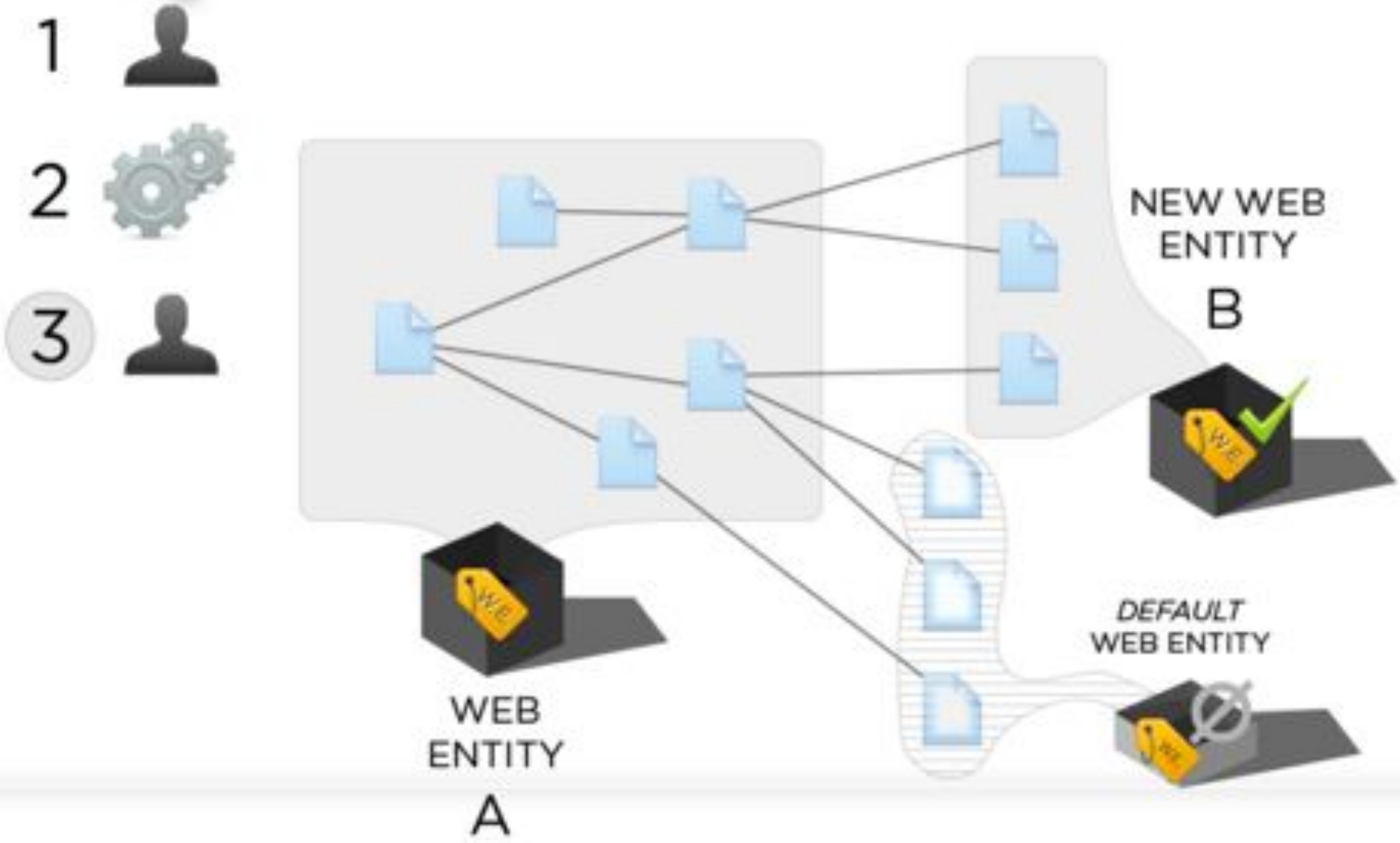
Research driven crawling

topic focused corpus?



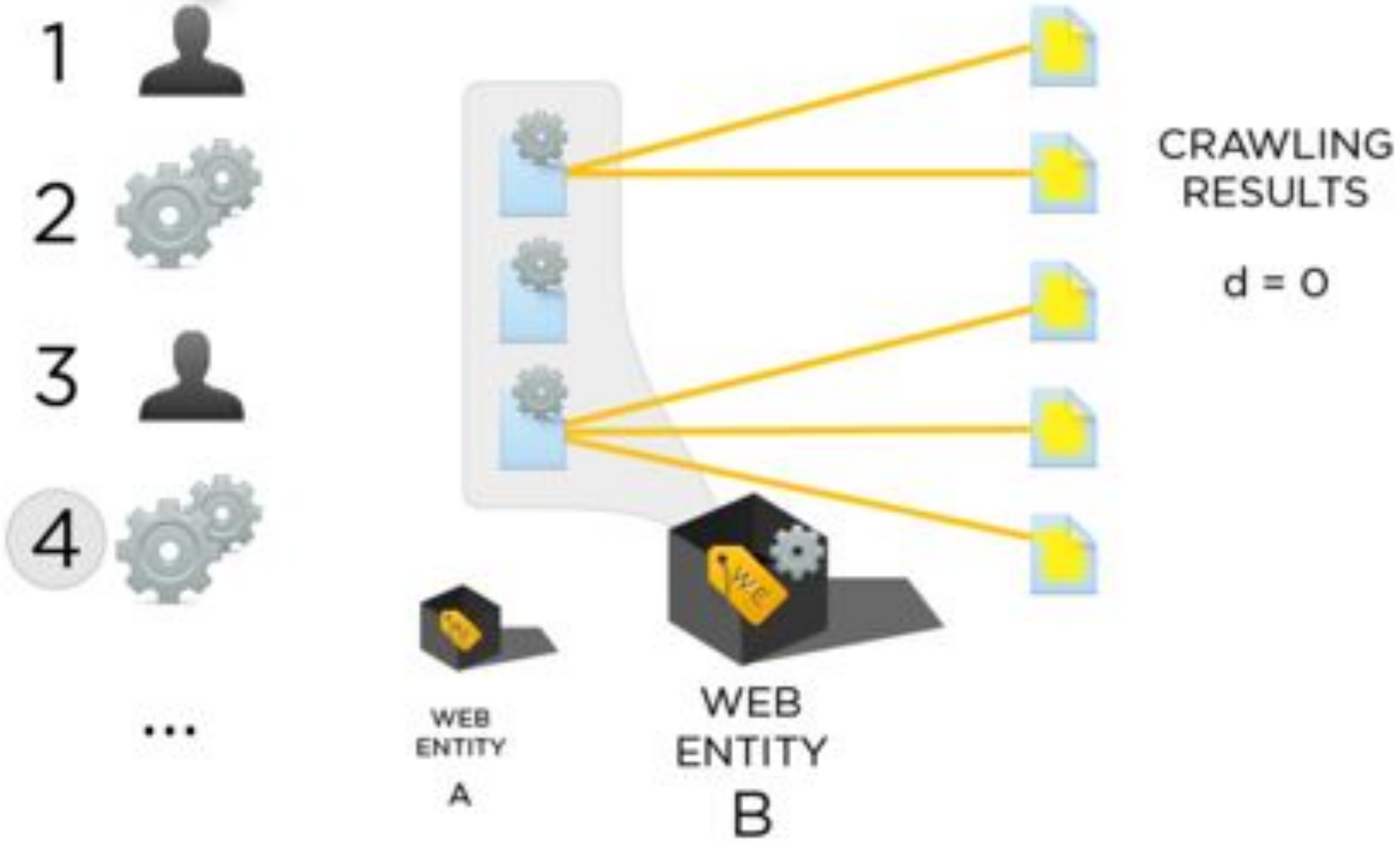
Research driven crawling

topic focused corpus?



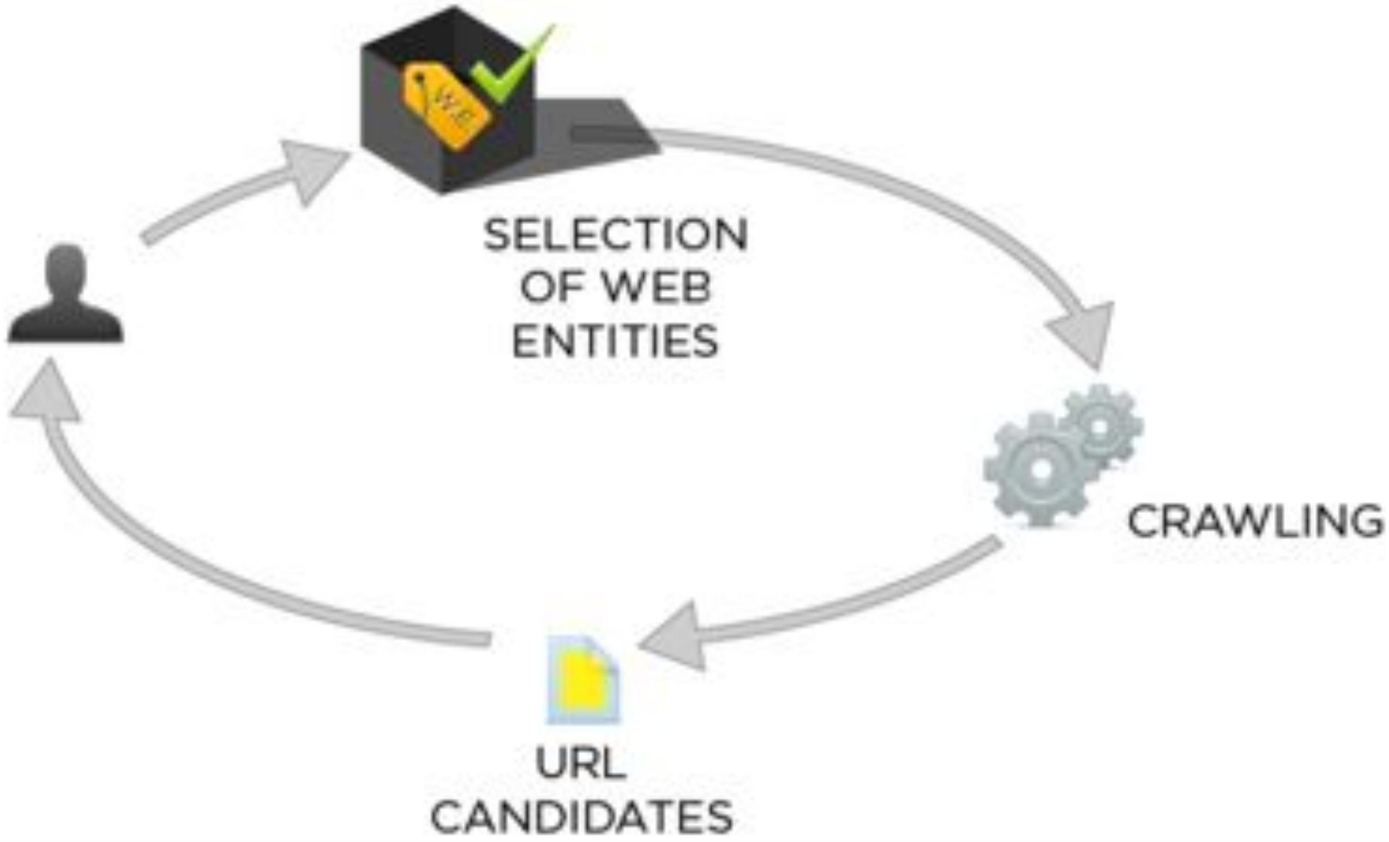
Research driven crawling

topic focused corpus?



Research driven crawling

topic focused corpus?



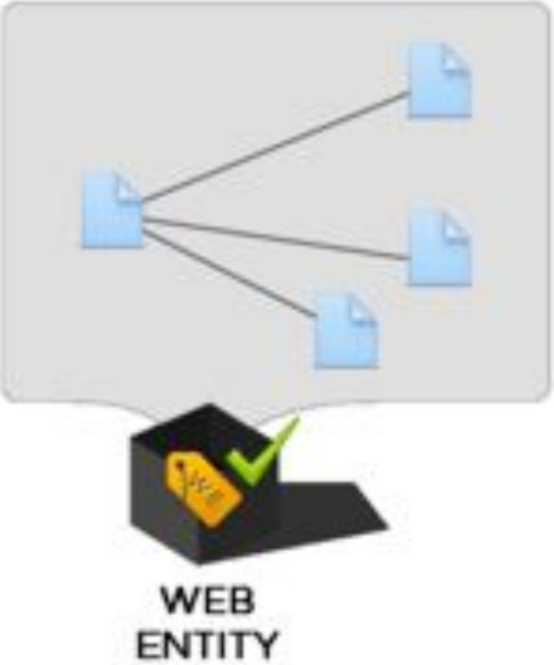
SELECTION OF WEB ENTITIES

CRAWLING

URL CANDIDATES

From sources to corpus

Define



Select

status

✓ included

✗ excluded

? undefined

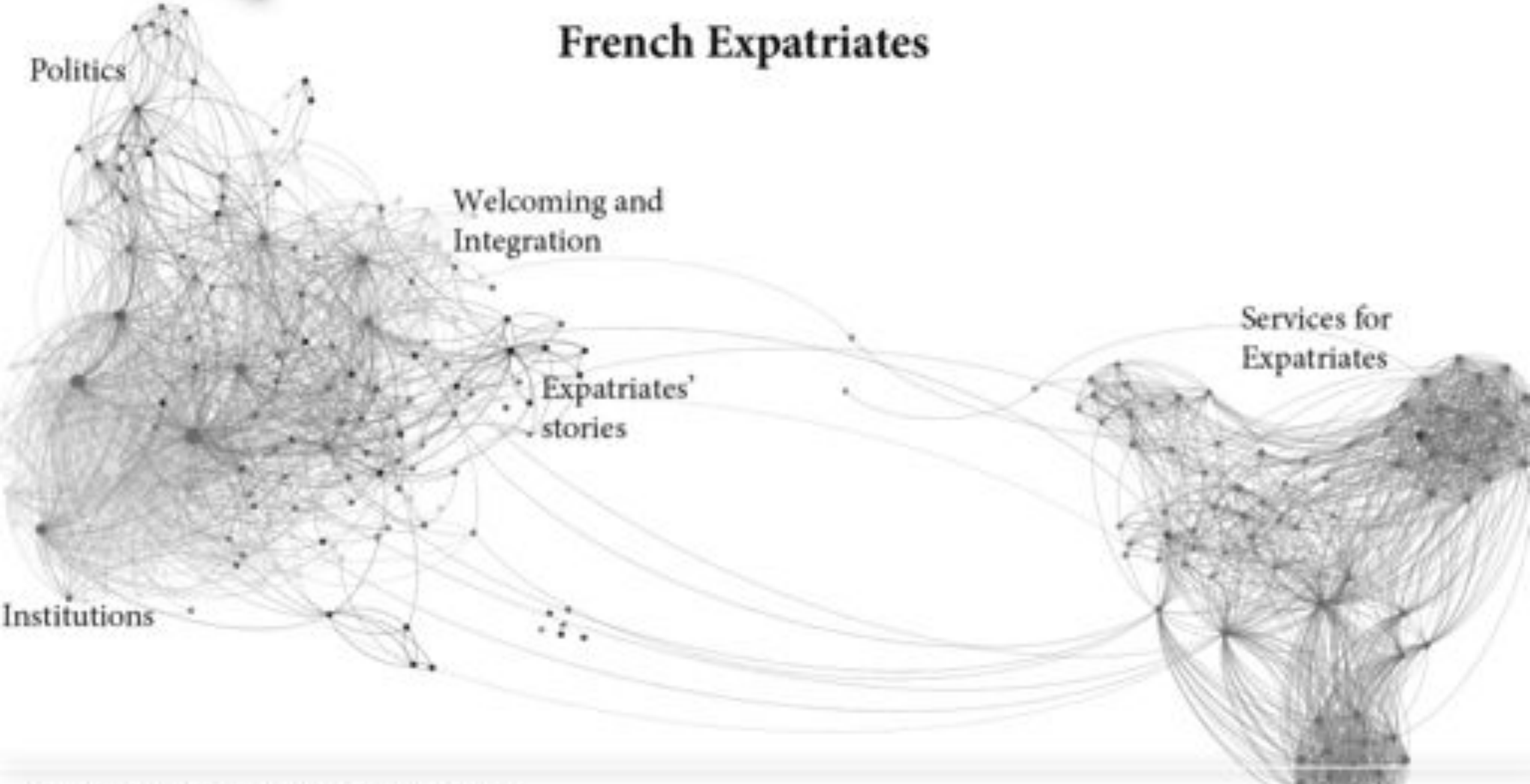
∅ unknown

Describe

tagging

Links : topological analysis

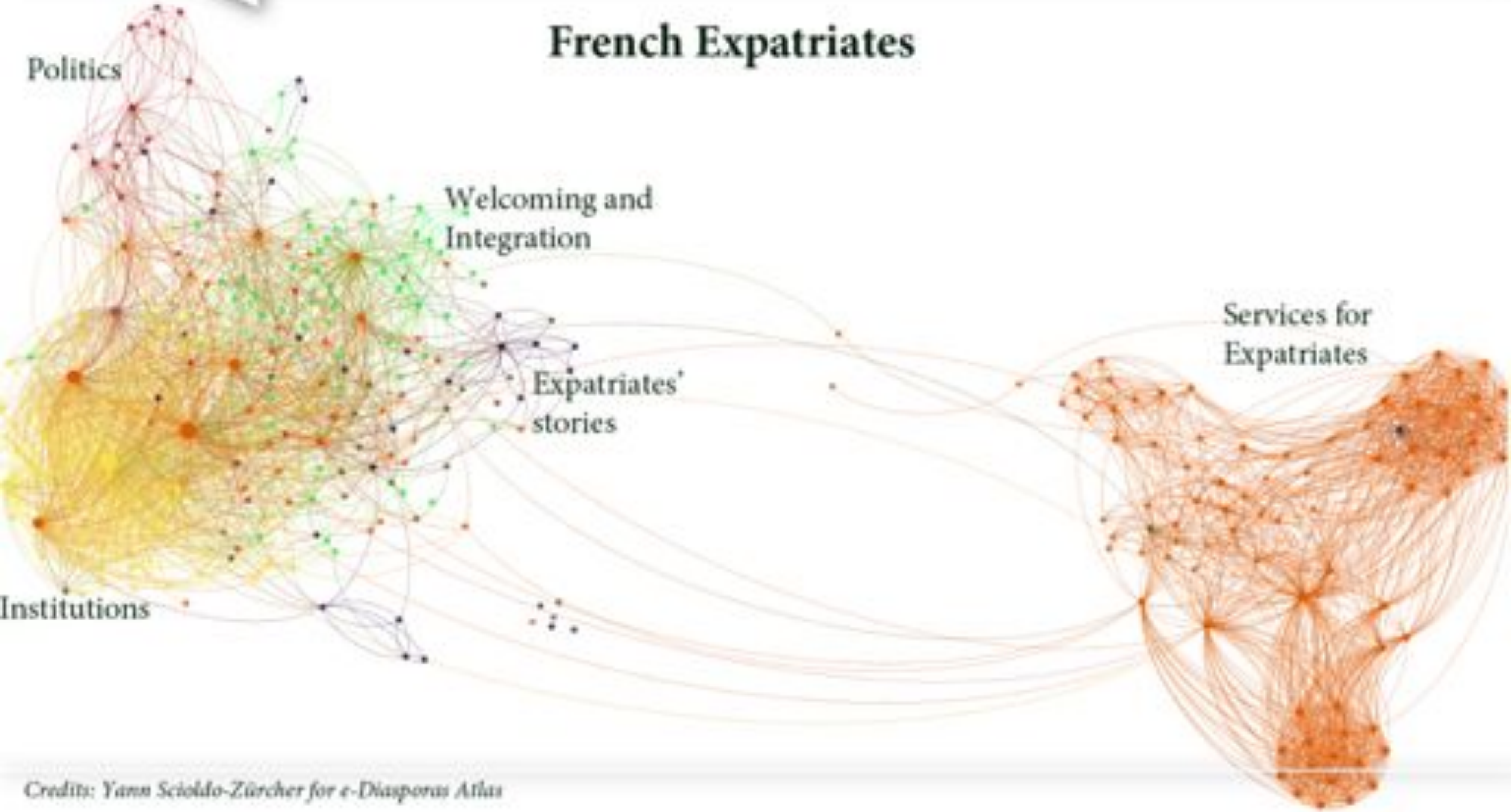
E-Diaspora Atlas



Credits: Yann Scioldo-Zürcher for e-Diasporas Atlas

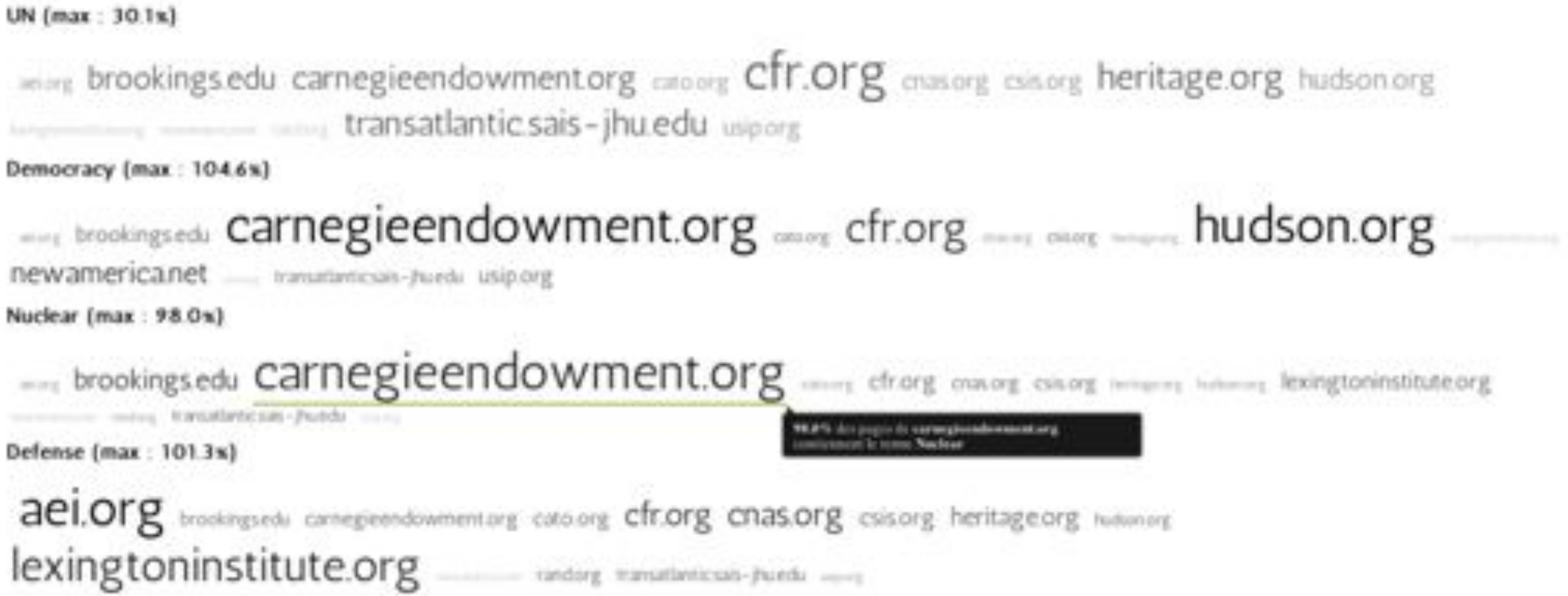
Qualification rules !

E-diaspora atlas



Credits: Yann Scioldo-Zürcher for e-Diasporas Atlas

Words: lexical analysis



UN (max : 30.1x)

aei.org brookings.edu carnegieendowment.org cato.org cfr.org cnas.org csis.org heritage.org hudson.org lexingtoninstitute.org newamerica.net rand.org transatlantic.sais-jhu.edu usip.org

Democracy (max : 104.6x)

aei.org brookings.edu carnegieendowment.org cato.org cfr.org cnas.org csis.org heritage.org hudson.org lexingtoninstitute.org newamerica.net rand.org transatlantic.sais-jhu.edu usip.org

Nuclear (max : 98.0x)

aei.org brookings.edu carnegieendowment.org cato.org cfr.org cnas.org csis.org heritage.org hudson.org lexingtoninstitute.org newamerica.net rand.org transatlantic.sais-jhu.edu usip.org

98.0% des pages de carnegieendowment.org contiennent le terme Nuclear

Defense (max : 101.3x)

aei.org brookings.edu carnegieendowment.org cato.org cfr.org cnas.org csis.org heritage.org hudson.org lexingtoninstitute.org newamerica.net rand.org transatlantic.sais-jhu.edu usip.org

Web Archives

Rich content in time

HCI → web archive : archive a corpus

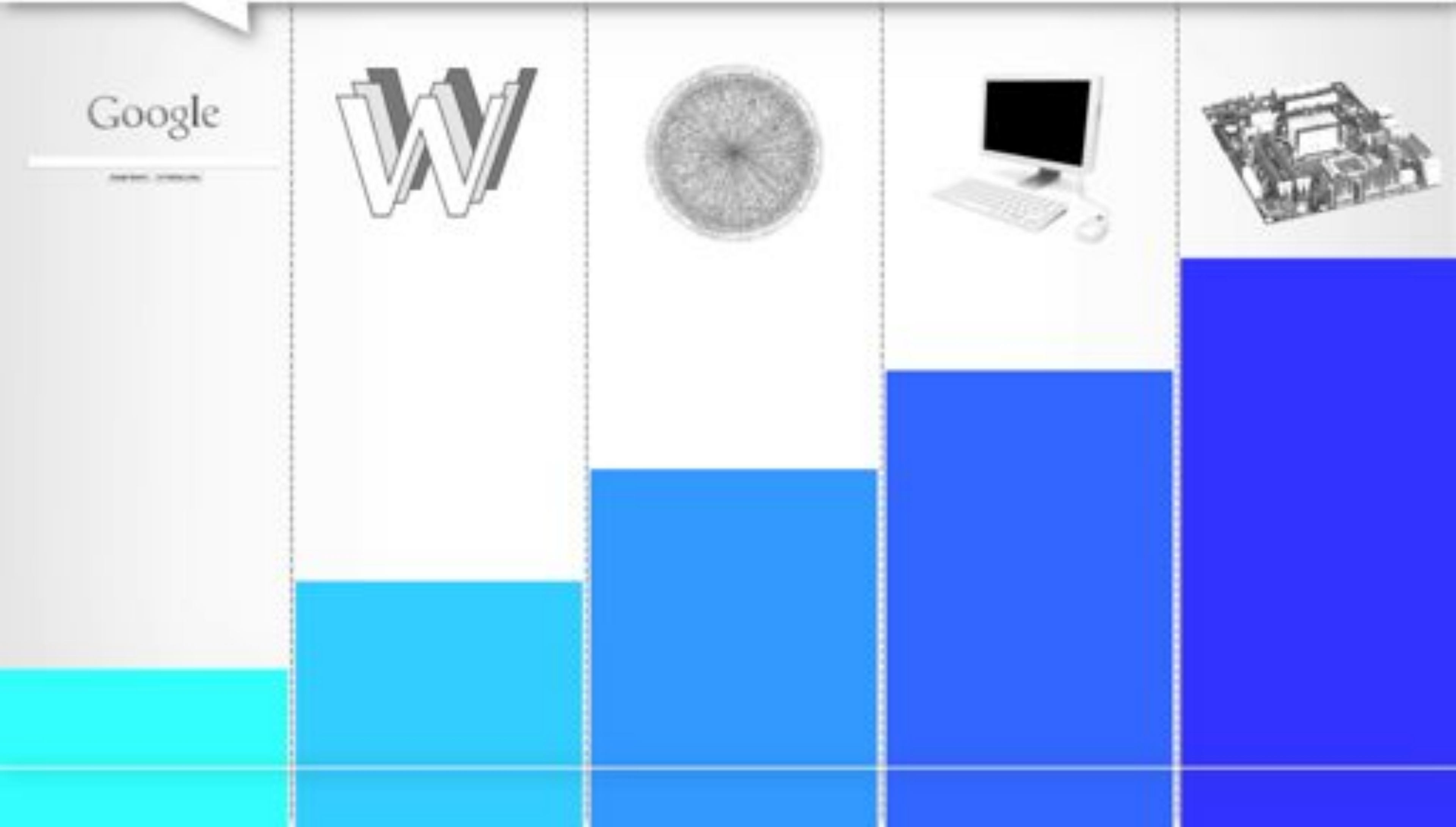
- export a corpus to be archived by a web archive institution
- harvest rich content
- regular harvesting through time

Archive → HCI : crawl an archive

- use HCI to build a corpus from an archive and not the live web
- benefit from the anteriority of the archive

Representativity ?

limitations

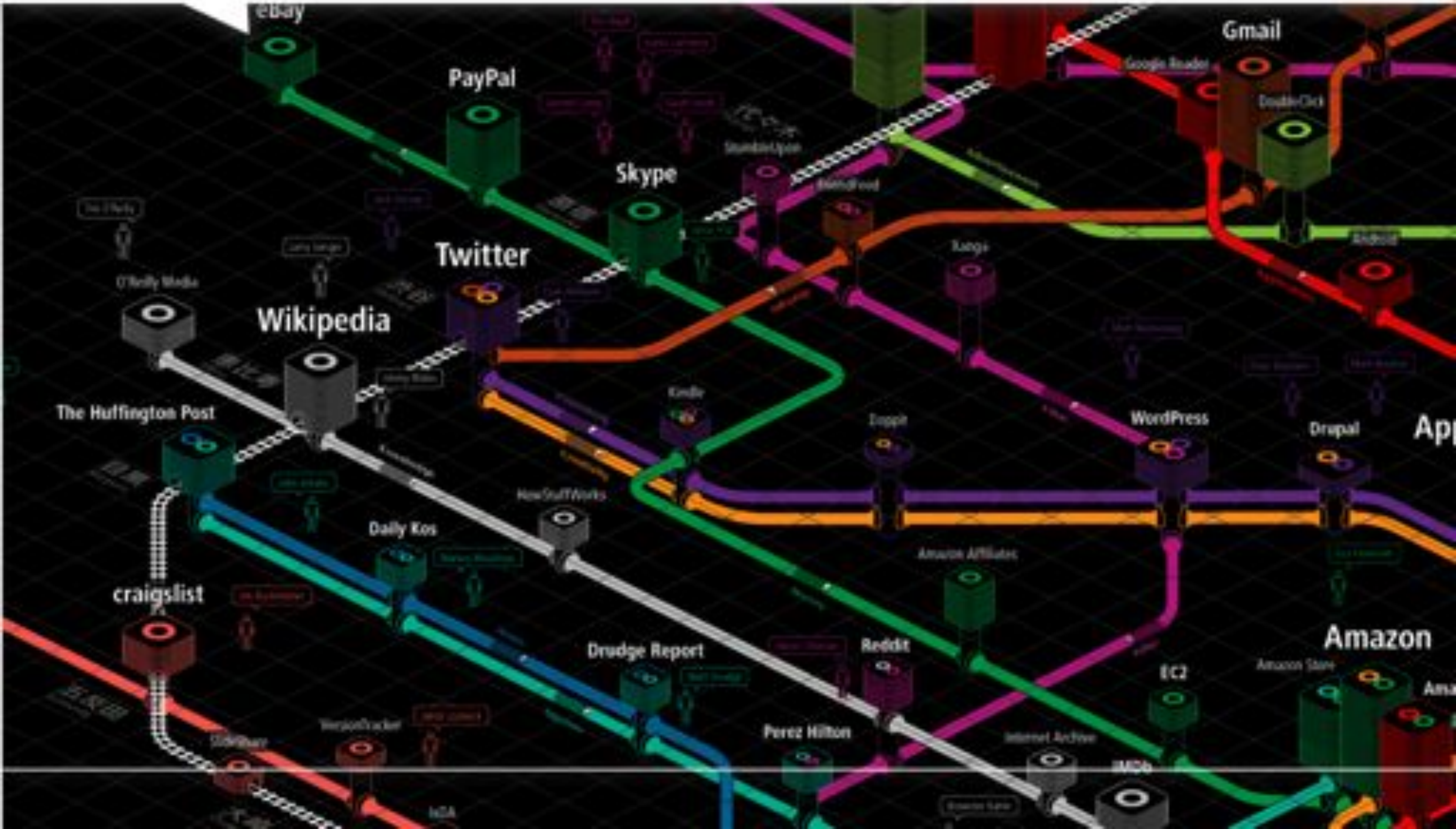


Web field as a carbon paper

The world seen from Facebook

The web is not everything
But don't forget it actually exists

Web Trend Map 4
Information Architect



Complementarity of methods

Quanti : ELLIPS panel of 6000 french people equiped with tablets

Quali : BeQuali, a qualitative survey archive

Web : use the web as a survey field

- trainings and assistance : web corpus for Social Sciences
- tools and methods : Hypertext Corpus Initiative
- a technical architecture : storage and crawling servers

thank you

medialab.sciences-po.fr