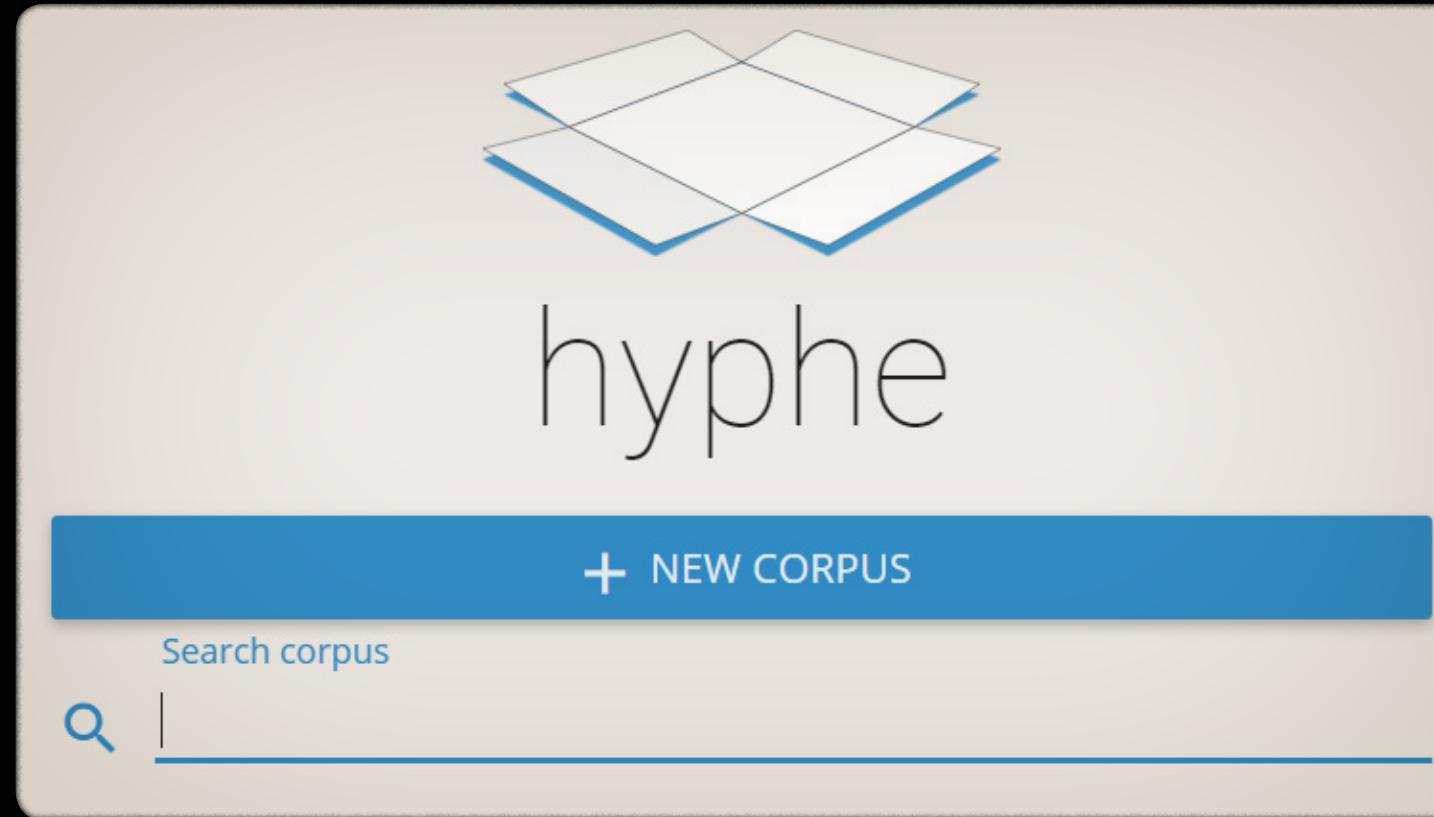


Crawling the web

What is the purpose of Hyphe?

Mathieu Jacomy
Aalborg University TANTLab

What is Hyphe?



A free, libre, open source web crawler
designed for the social sciences

What is Hyphe?

Official website

<http://hyphe.medialab.sciences-po.fr/>

Demo

<http://hyphe.medialab.sciences-po.fr/demo/>

Source code and install

<https://github.com/medialab/hyphe>

Bug reporting

<https://github.com/medialab/hyphe/issues>

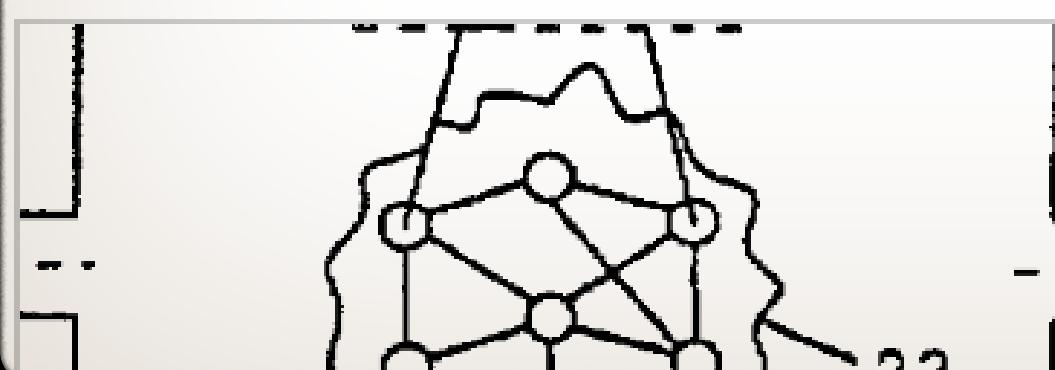
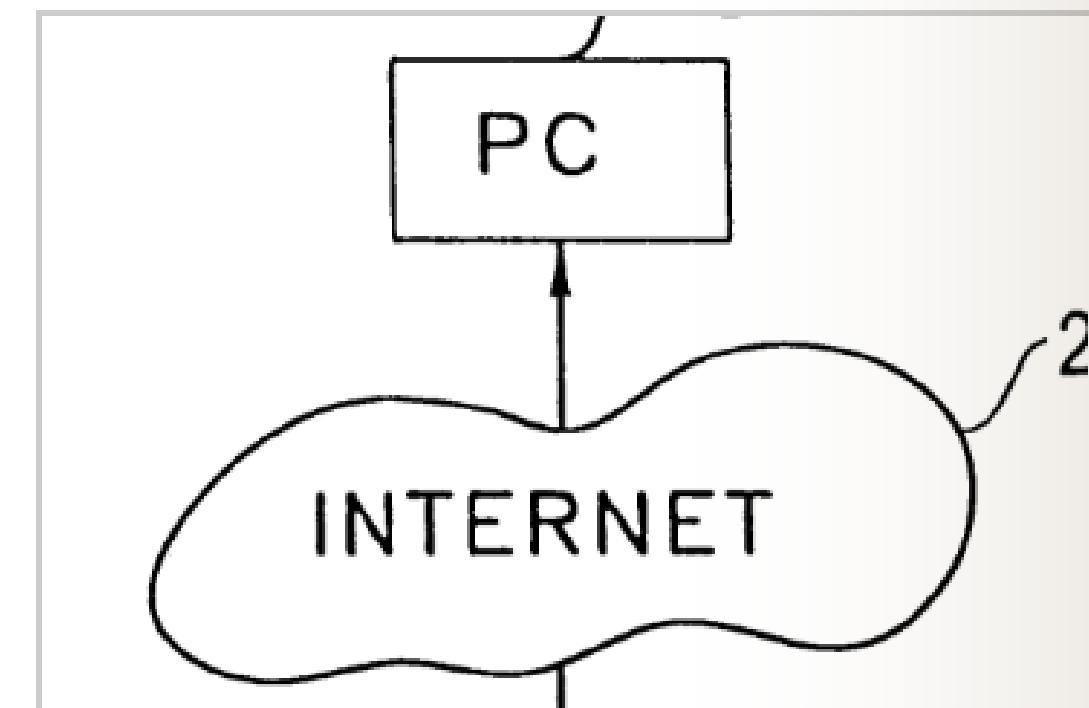
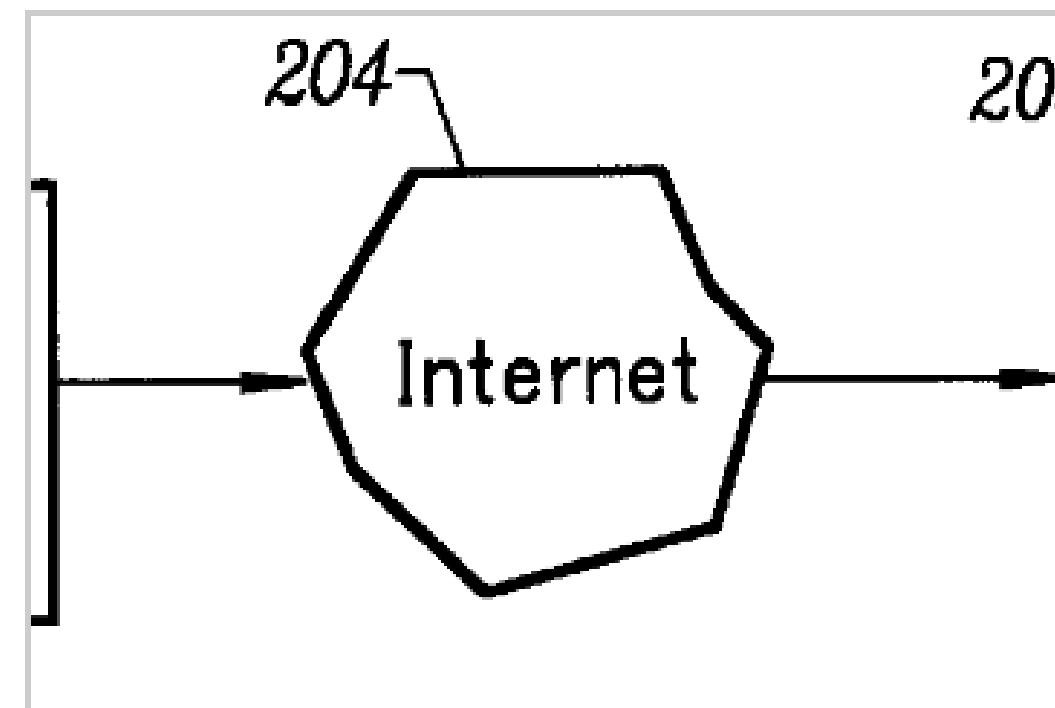
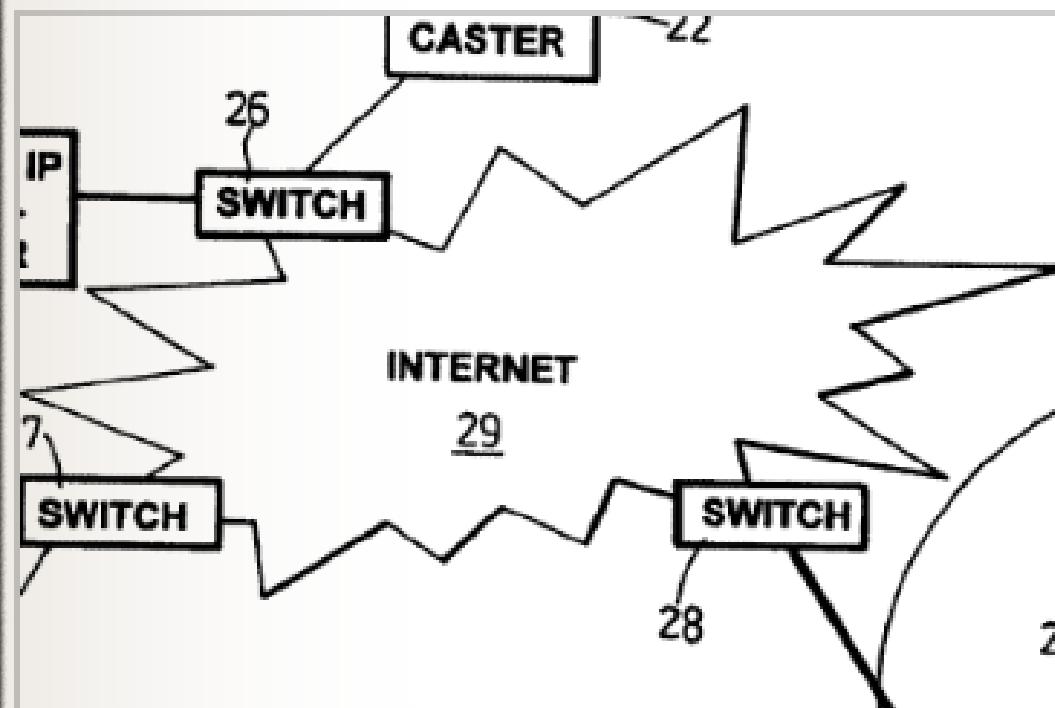
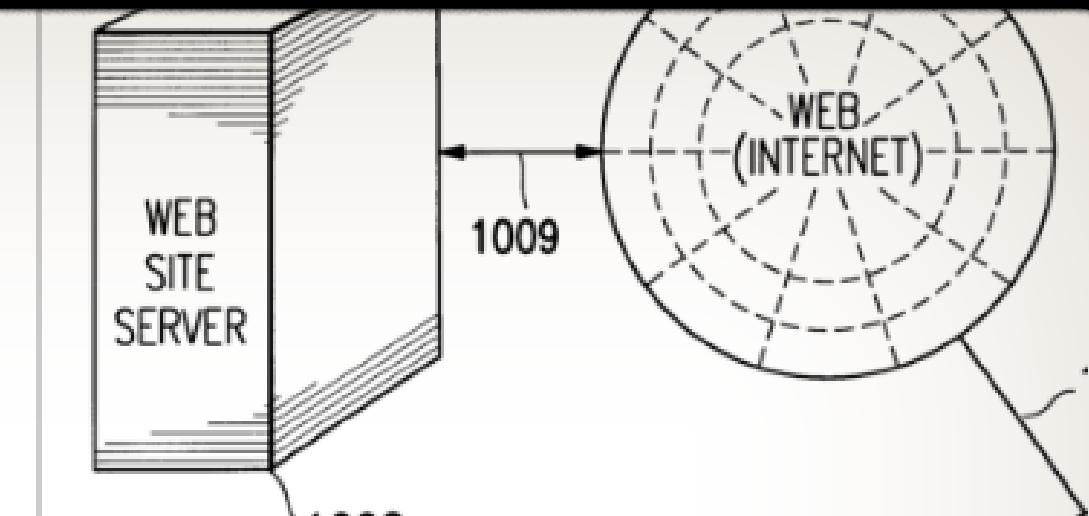
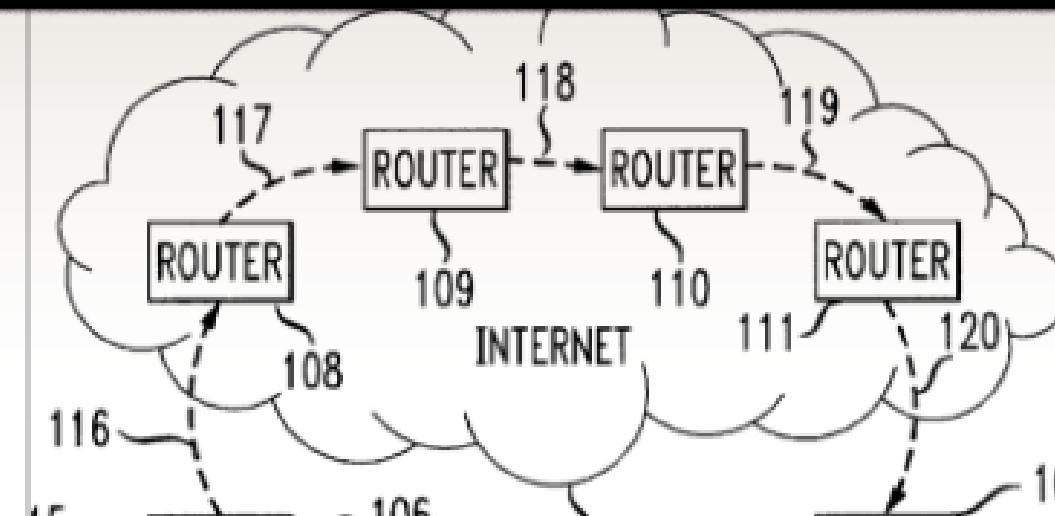
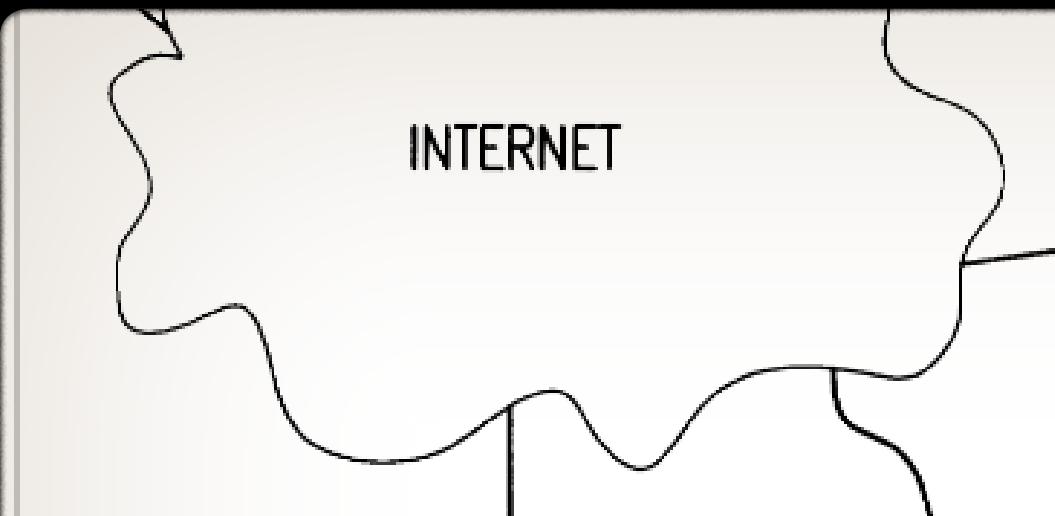
Paper

<http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13051/12797>

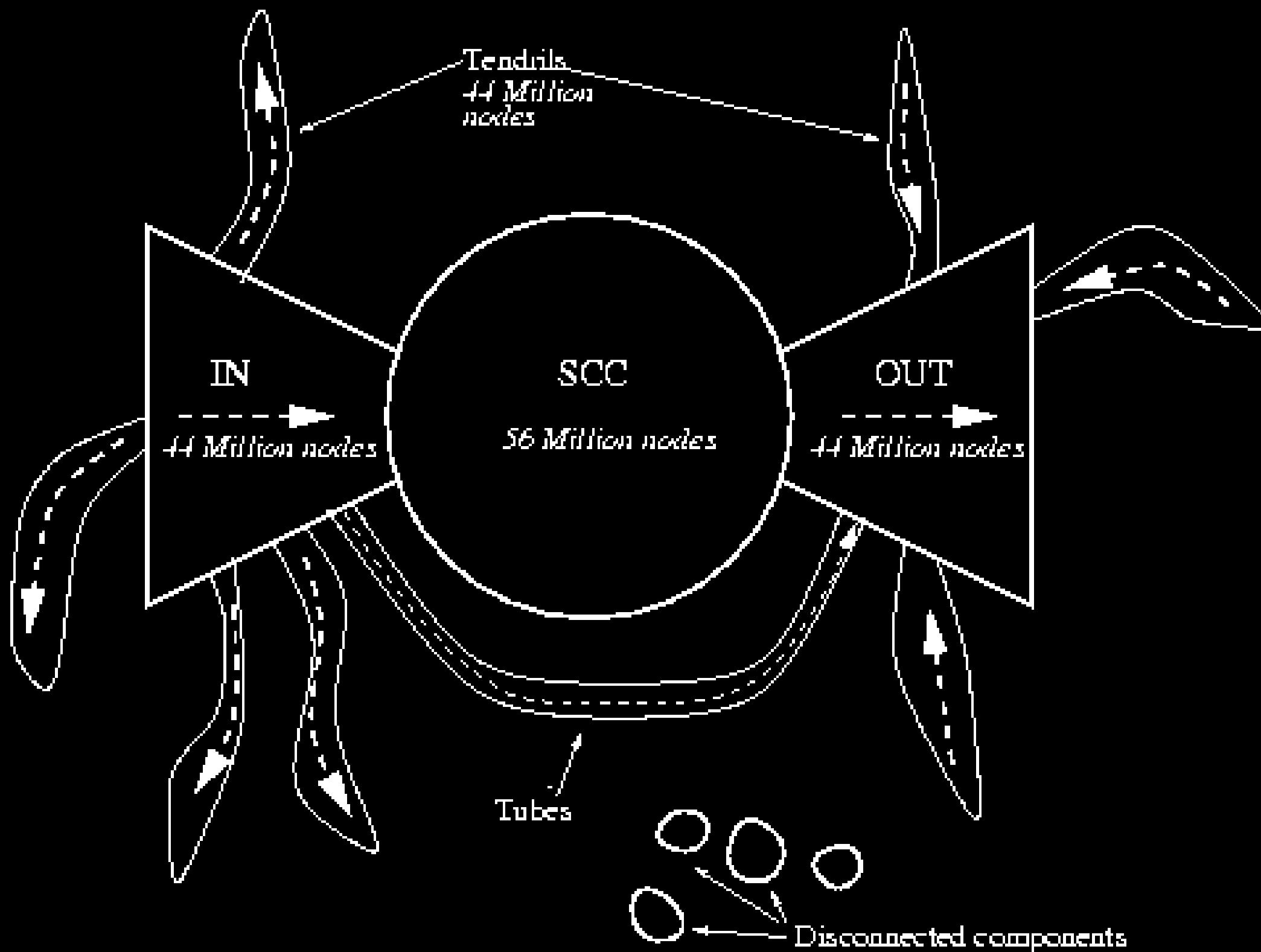
ICWSM poster

<http://www.medialab.sciences-po.fr/wp-content/uploads/2016/05/Hyphe-ICWSM-A3.pdf>

Depictions of internet in academic papers

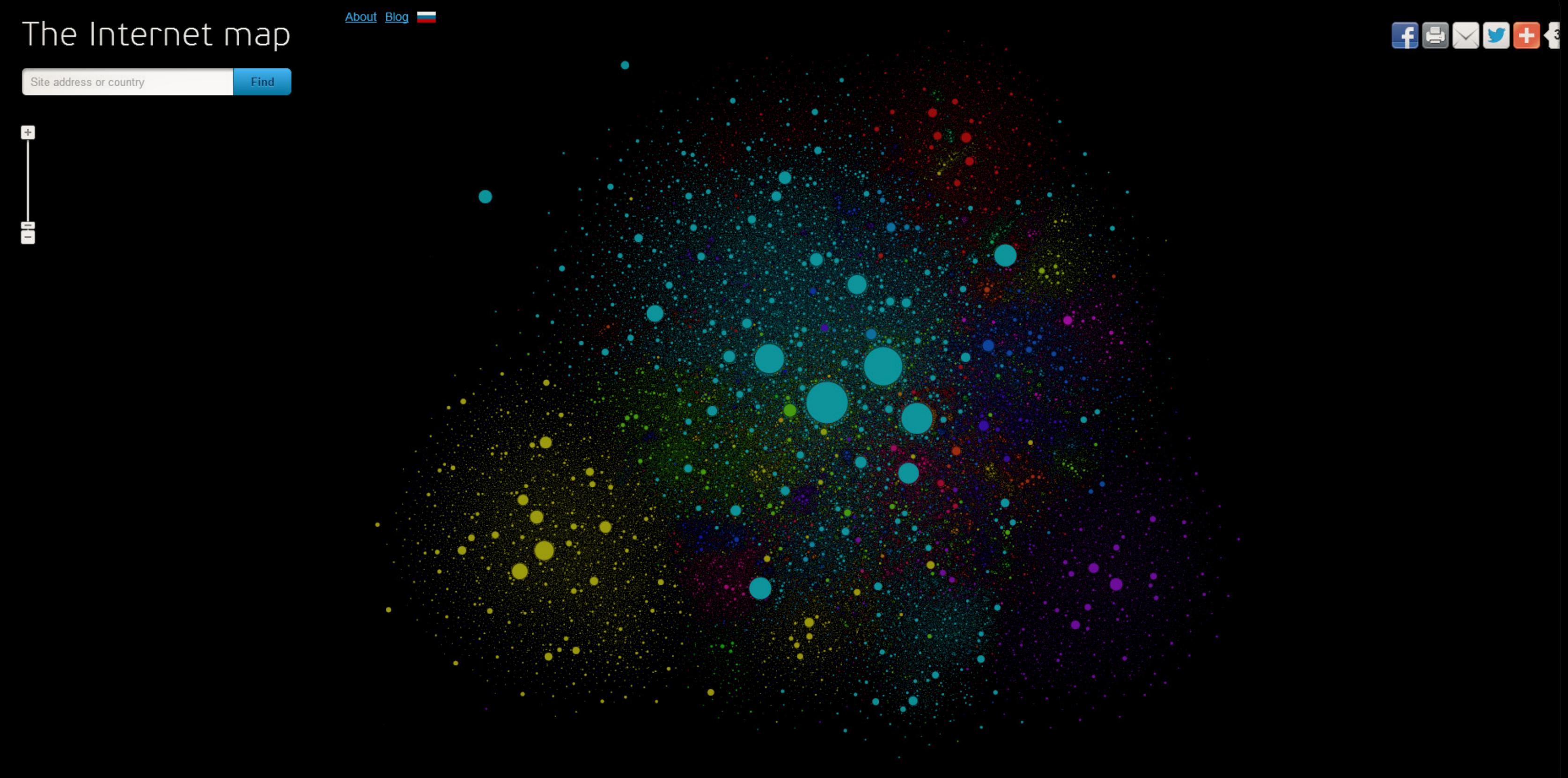


The first scientific image of the web

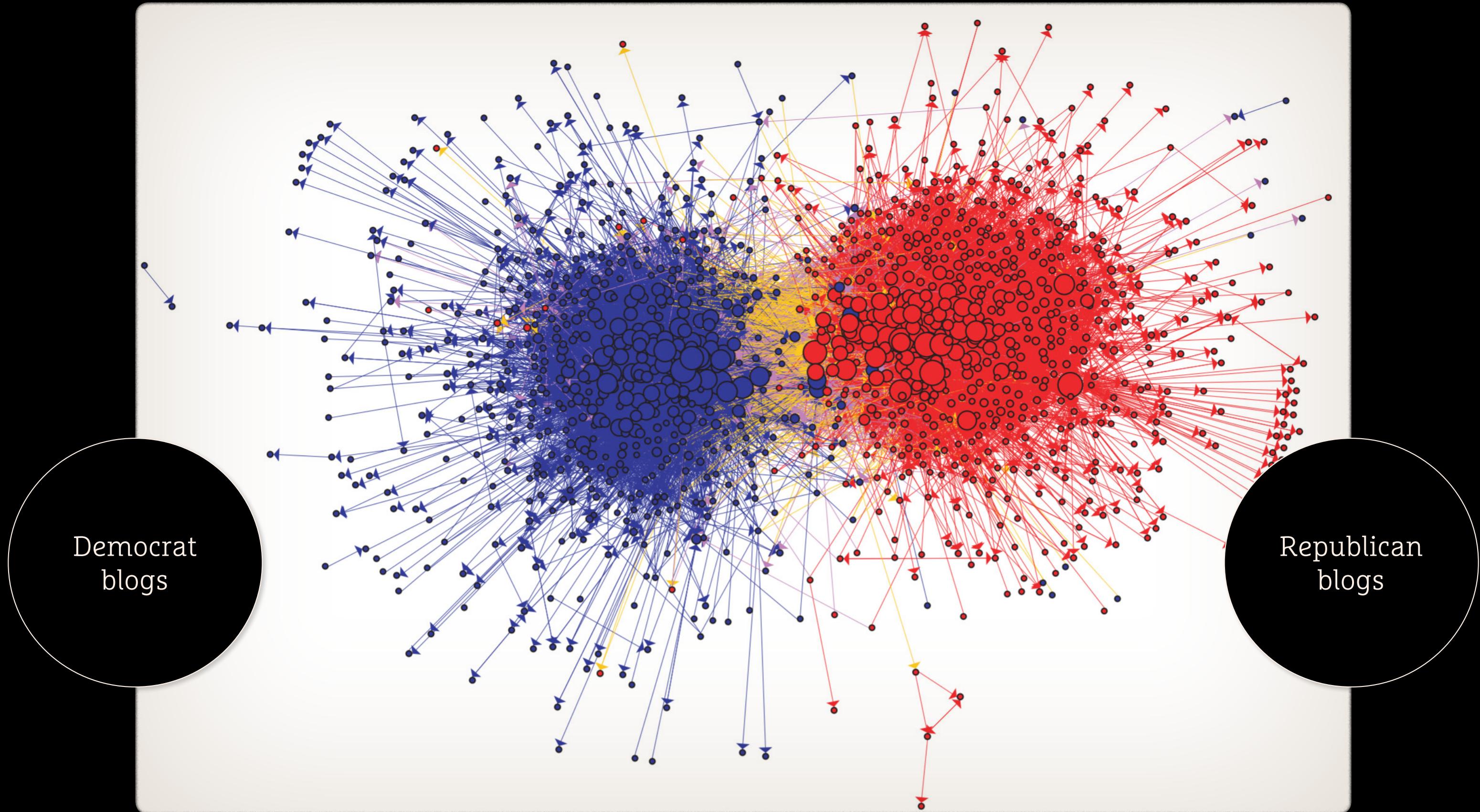


*The bow tie,
IBM's Almaden Research, 2000*

A more recent depiction of the top 10,000 websites



A network view of a subcorpus of the web



Adamic, L. A., & Glance, N. (2005, August). *The political blogosphere and the 2004 US election: divided they blog*. In *Proceedings of the 3rd international workshop on Link discovery* (pp. 36-43). ACM.

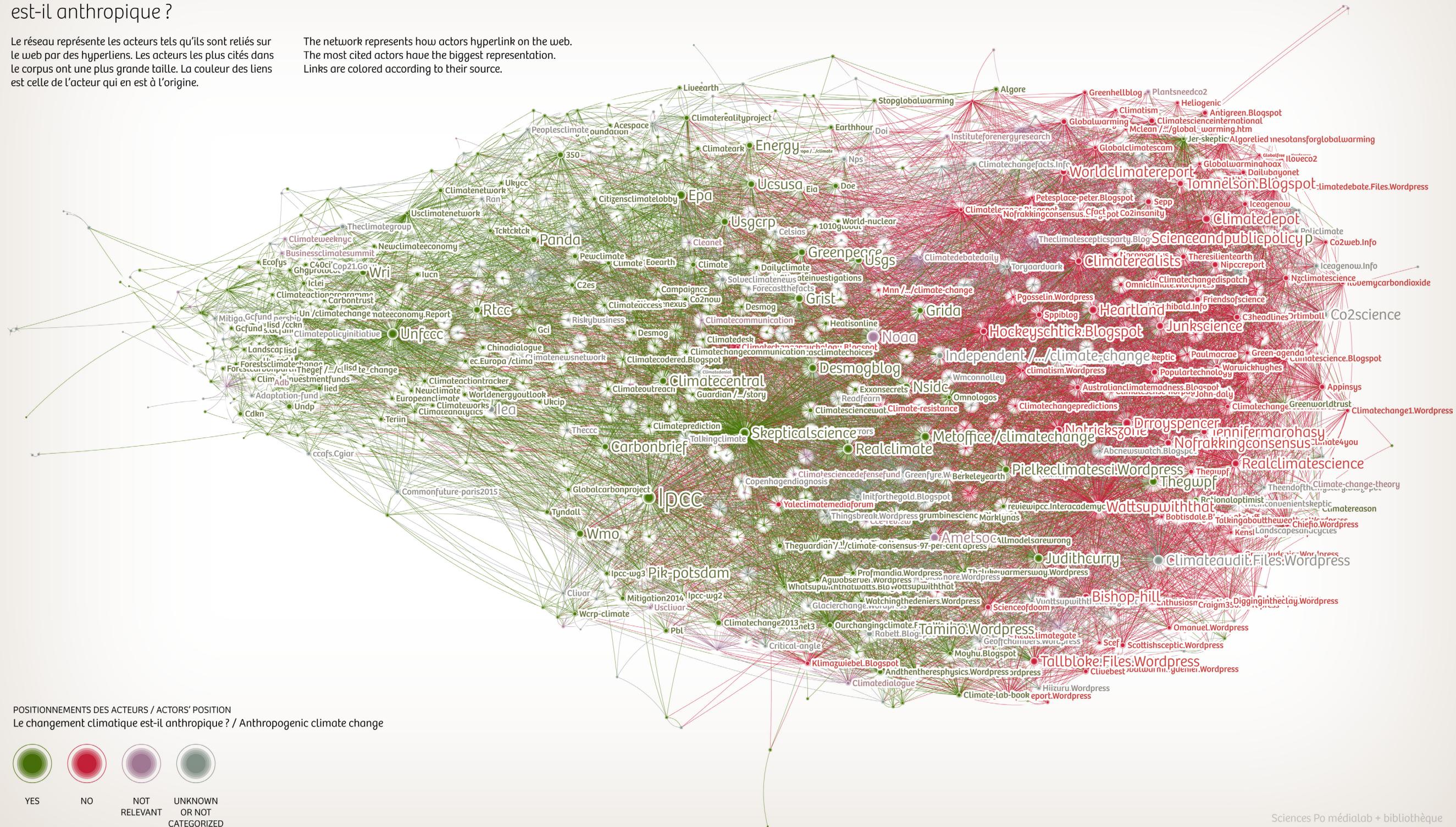
Fully analyzed corpus of a domain

Corpus web sur le
changement climatique :
le changement climatique
est-il anthropique ?

Le réseau représente les acteurs tels qu'ils sont reliés sur le web par des hyperliens. Les acteurs les plus cités dans le corpus ont une plus grande taille. La couleur des liens est celle de l'acteur qui en est à l'origine.

Web corpus
on climate change:
anthropogenic climate change

The network represents how actors hyperlink on the web.
The most cited actors have the biggest representation.
Links are colored according to their source.



Sciences Po médialab + bibliothèque

Hyphe's web crawling process

1. Sourcing

Define your field a priori
and gather starting points

3. Monitoring

Visualize corpus
and monitor its properties

2. Harvesting (crawl)

Download the data
with a crawler

4. Curation

Select documents to limit
topic drifting and adjust
corpus boundaries

5. Finalization

Validate general quality
and export corpus

Hyphe's web crawling process

1. Sourcing

Define your field a priori
and gather starting points

2. Harvesting (crawl)

Download the data
with a crawler

3. Monitoring

Visualize corpus
and monitor its properties

4. Curation

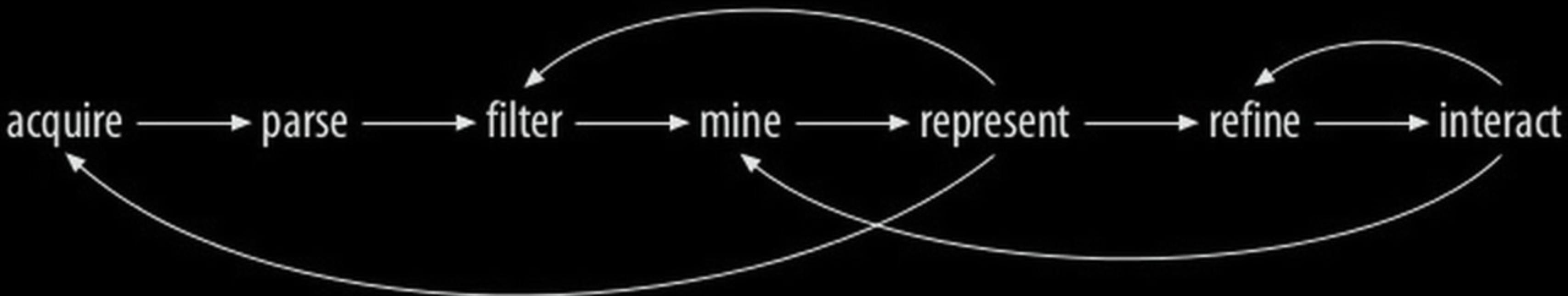
Select documents to limit
topic drifting and adjust
corpus boundaries

5. Finalization

Validate general quality
and export corpus



The data mining process has multiple loops



Thank you for your attention

*@jacomy
reticular.hypotheses.org
Mathieu.Jacomy@gmail.com*

