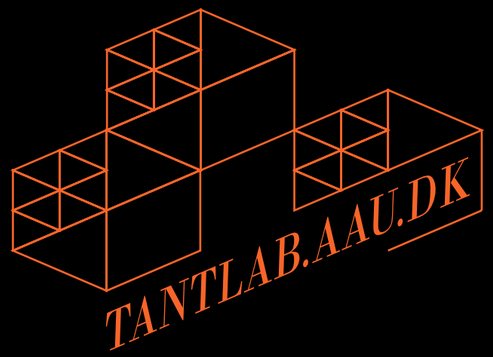
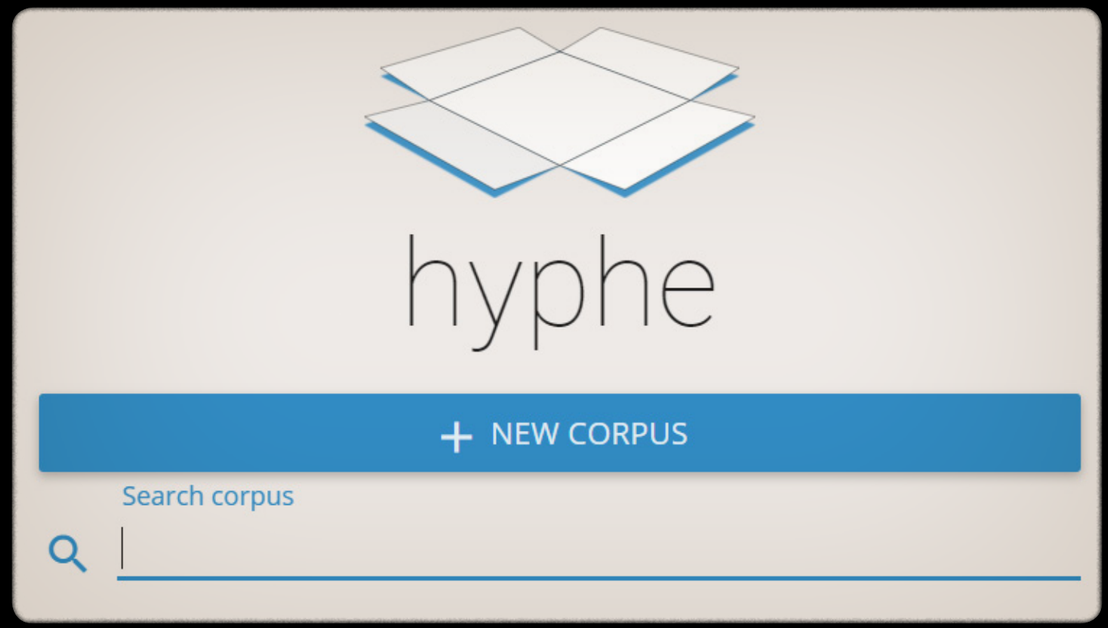


Crawling the web

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/

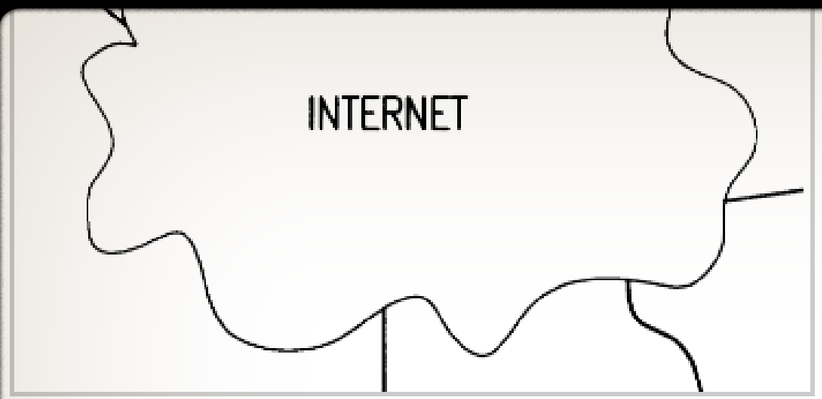
Source code and install
https://github.com/medialab/hyphe

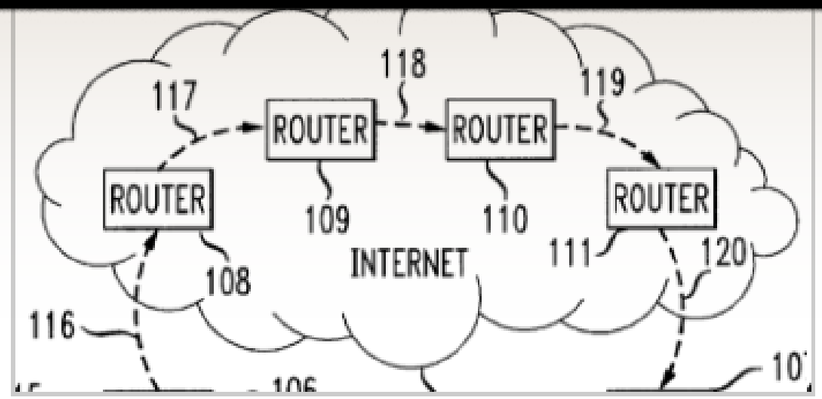
Paper
http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13051/12797

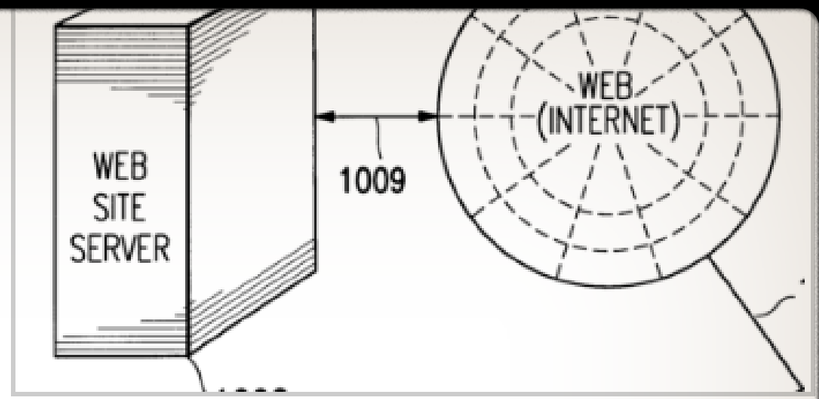
ICWSM poster
http://www.medialab.sciences-po.fr/wp-content/uploads/2016/05/Hyphe-ICWSM-A3.pdf

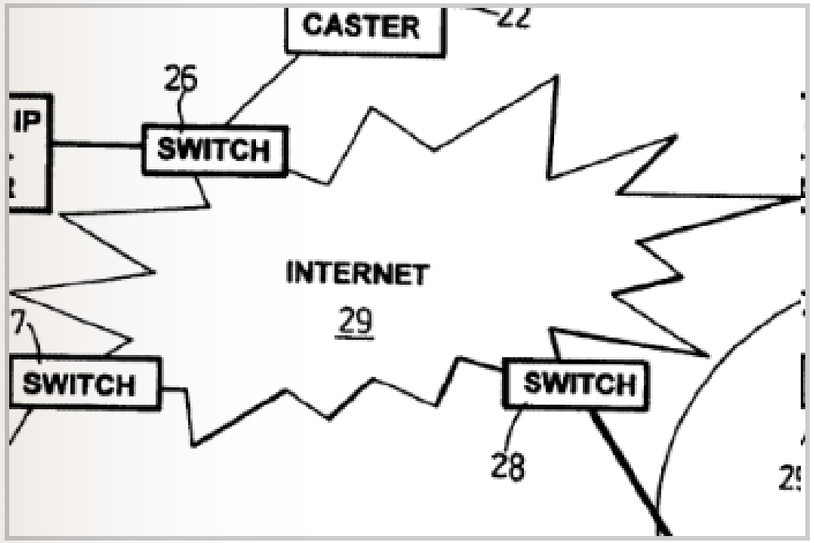
Bug reporting
https://github.com/medialab/hyphe/issues

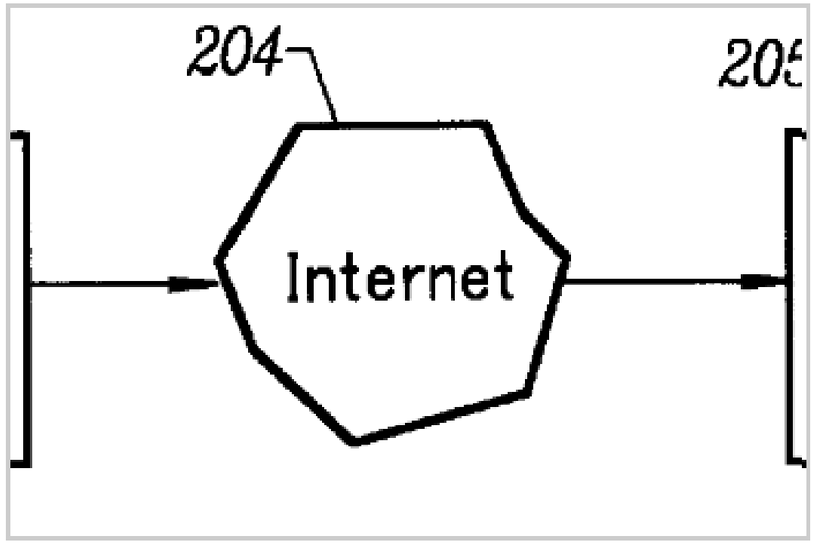
Depictions of internet in academic papers

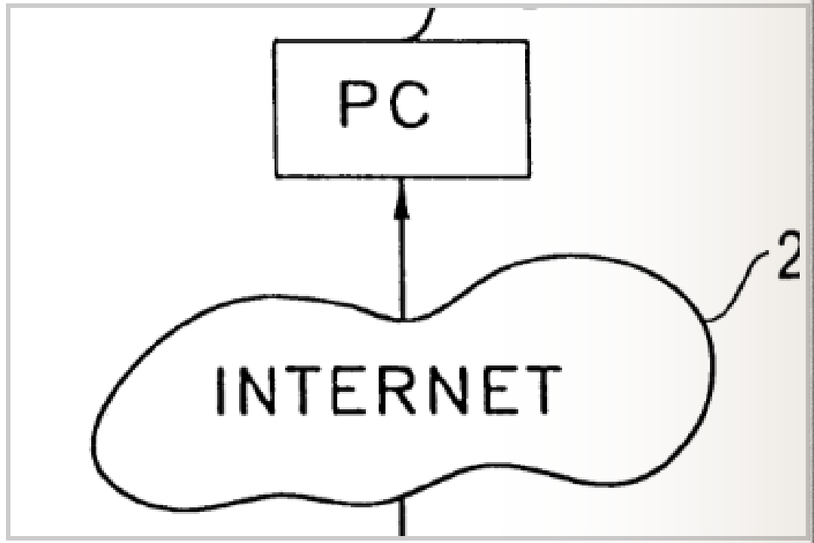


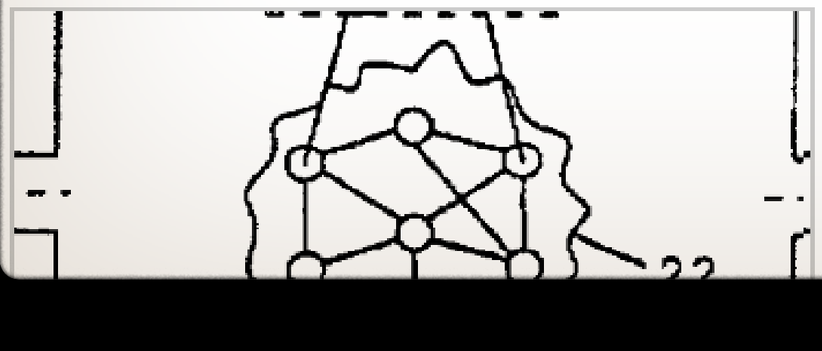






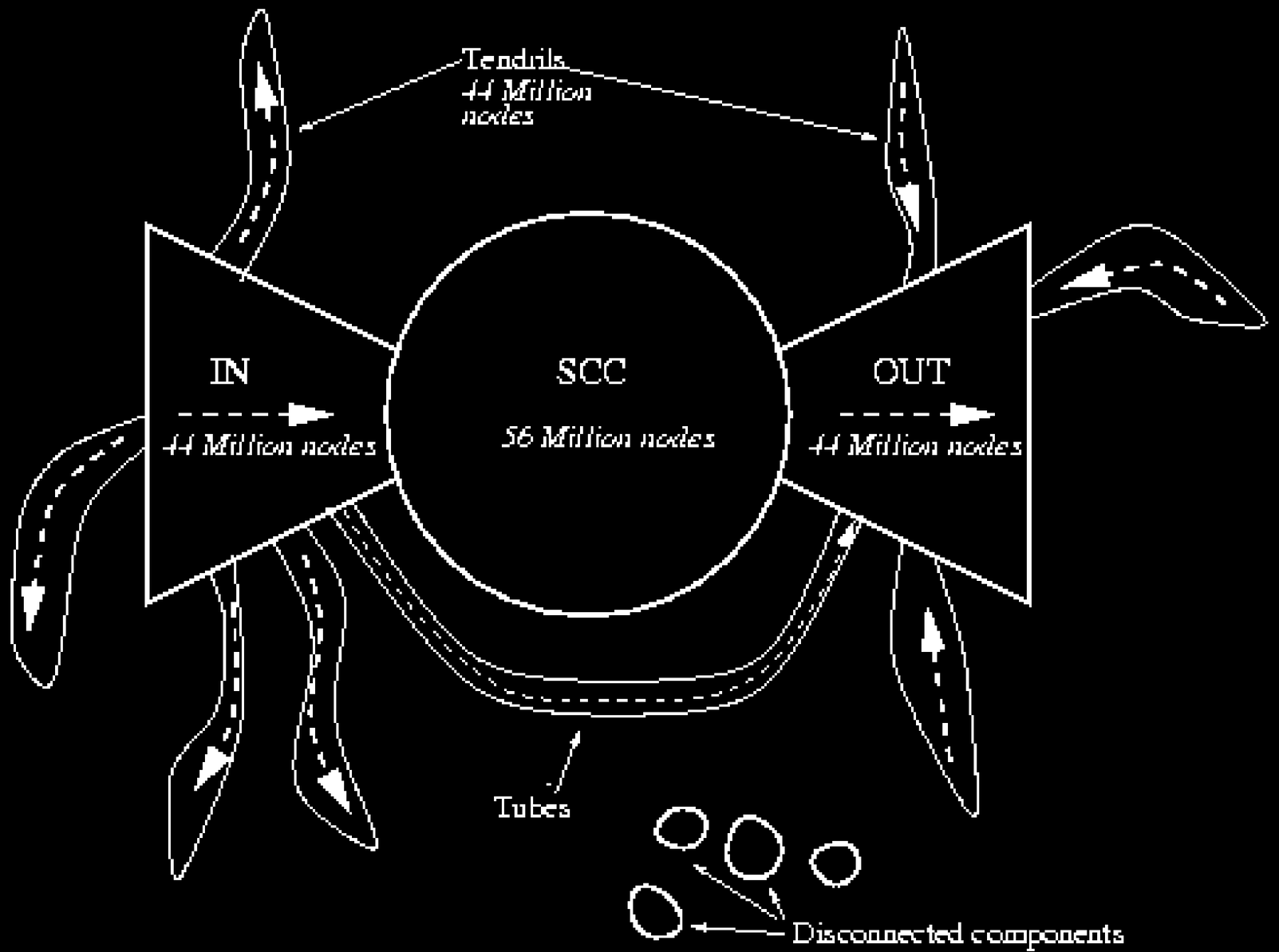






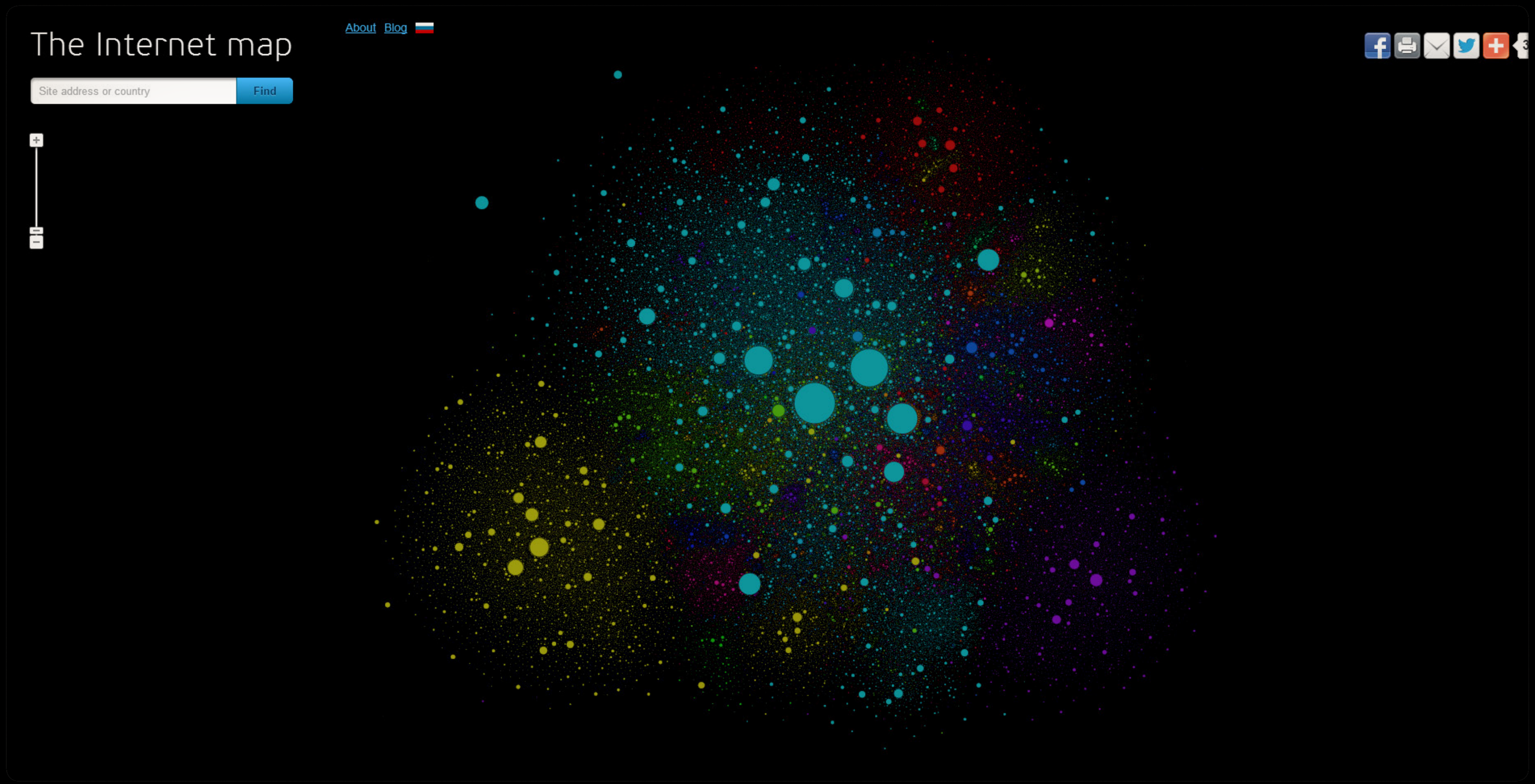
http://noahveltman.com/internet-shape/

The first scientific image of the web



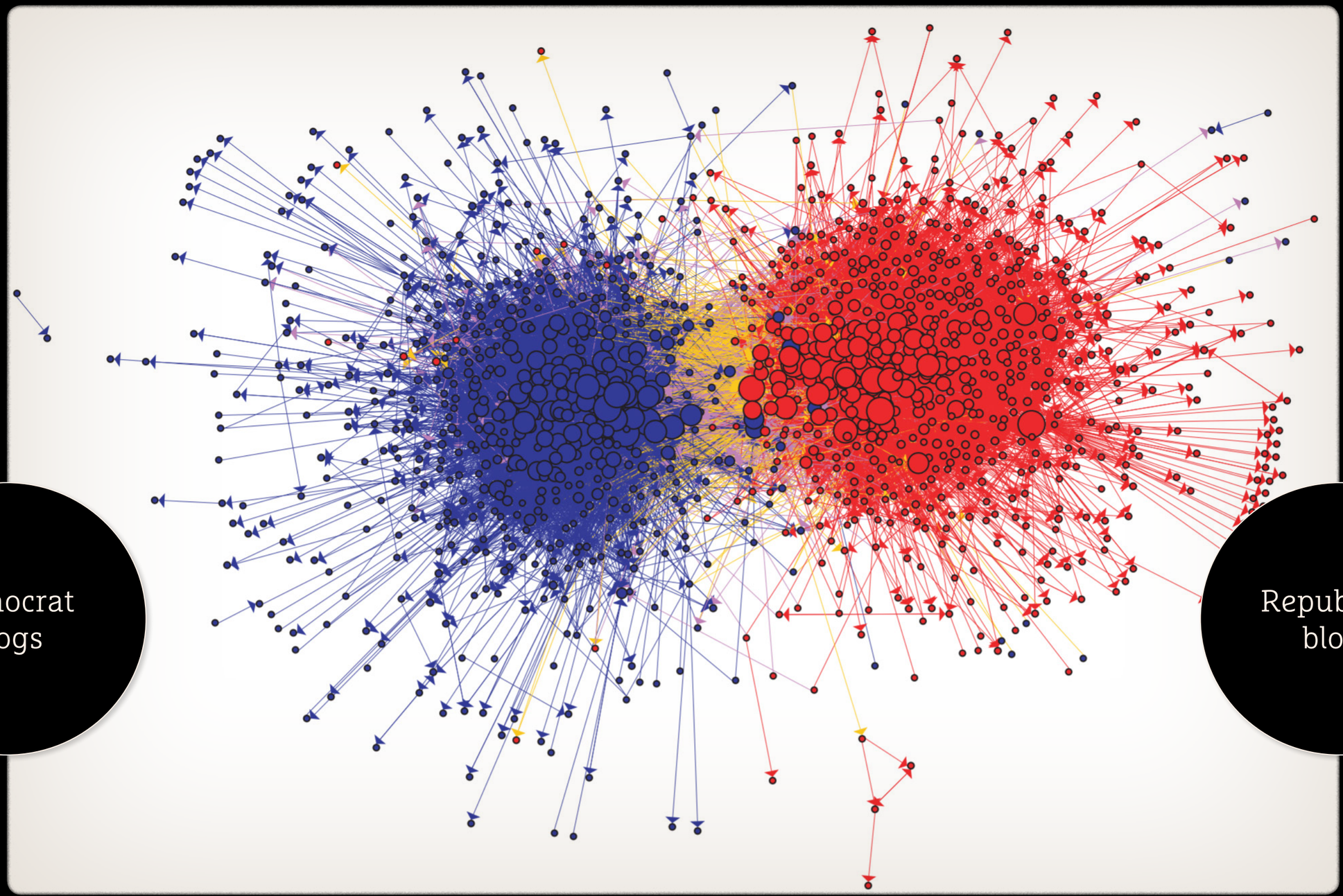
The bow tie,
IBM's Almaden Research, 2000

A more recent depiction of the top 10,000 websites



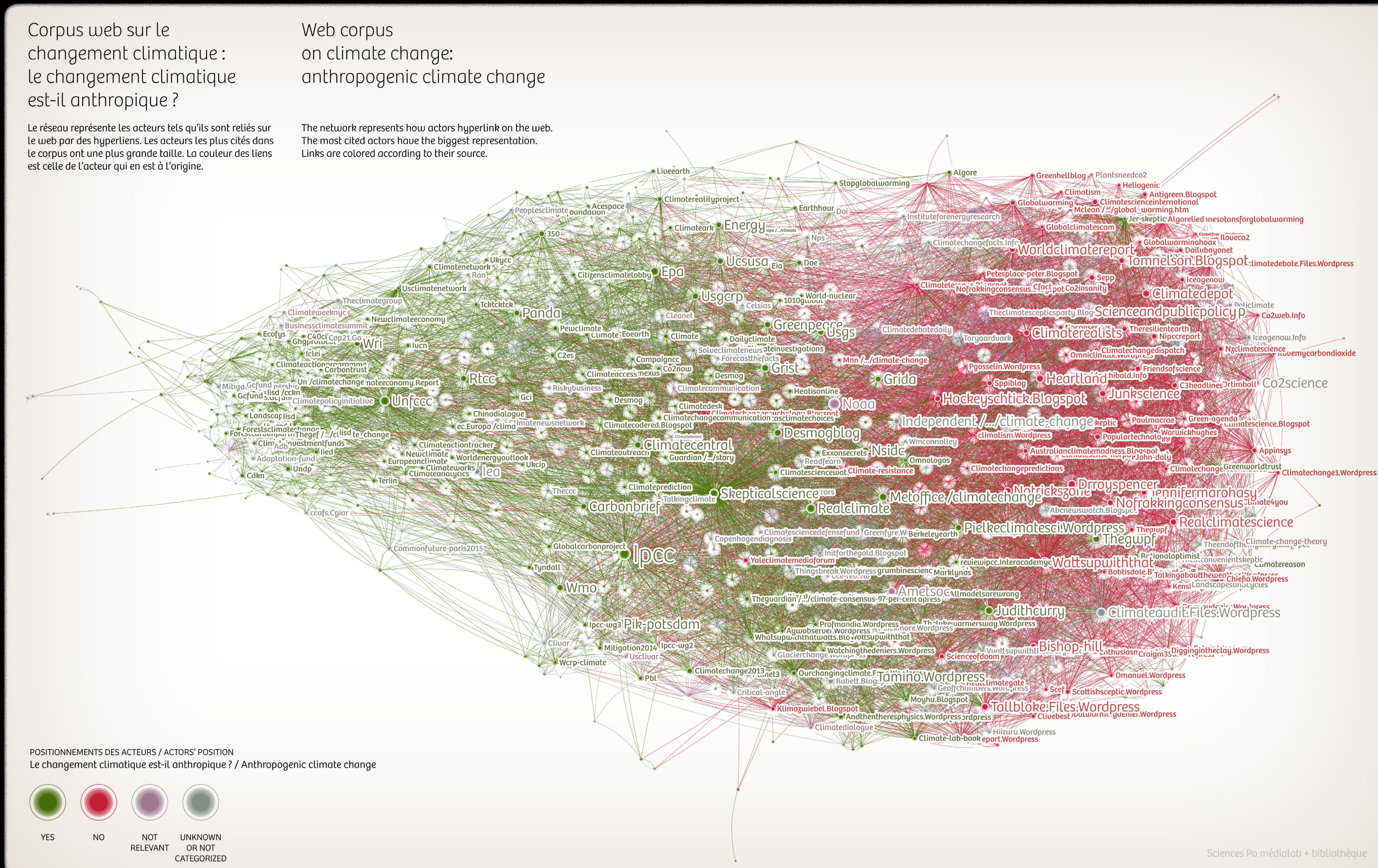
http://internet-map.net

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



Climate change debate on the web
Sciences Po médialab (2015)

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

Hyphe's web crawling process

1. Sourcing
   Define your field a priori
   and gather starting points

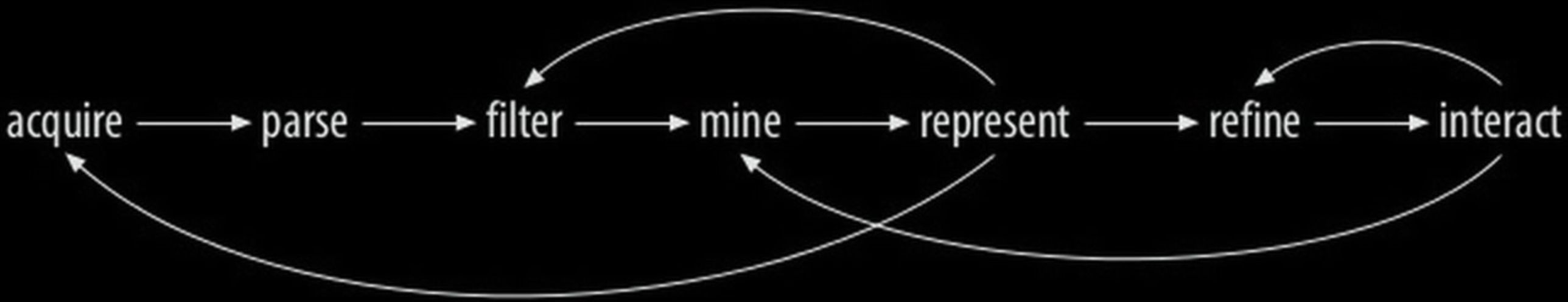
2. Harvesting (crawl)
   Download the data
   with a crawler

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



The chain of data mining
by Ben Fry

Thank you for your attention

@jacomyma
reticular.hypotheses.org
Mathieu.Jacomy@gmail.com