L'Archivage du Web

Julien Masanès
Internet Memory Foundation

Collège de France
Mars 2012

Introduction

- Centralité du web, application de publication de l'internet
- Premier artefact culturel, source pour l'histoire et la science du future
- Ce que la problématique de sa préservation nous apprend de ce média

L’objet

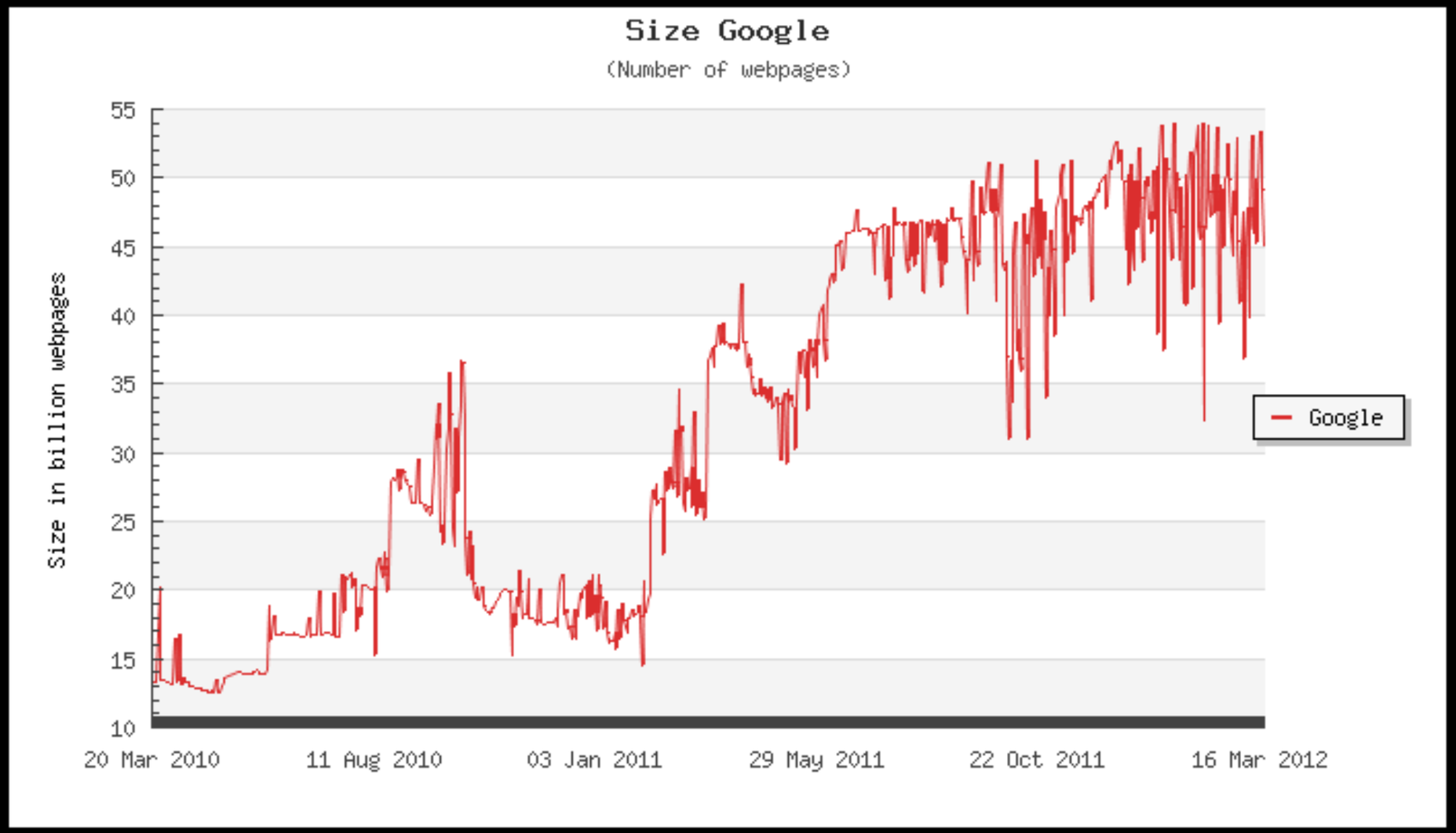
Mesure

- infini (génération à la demande)
- cela dépend de l'outil de mesure (crawler)

Mesure

- 555 million millions sites web (Décembre 2011).
- 200 millions nouveaux sites en 2011
- 152 millions blogs (2010 BlogPulse).
- 250 millions tweets par jour sur Twitter en (Oct-2011)
- 30 milliards d'éléments de contenus (liens, notes, photos, etc.) partagés sur Facebook chaque mois (2010)

Mesure


Size Google
(Number of webpages)
Size in billion webpages
55
50
45
40
35
30
25
20
15
10
20 Mar 2010
11 Aug 2010
03 Jan 2011
29 May 2011
22 Oct 2011
16 Mar 2012
Google

Mesure

- 1 million livres/an
- (Unesco)

- imprimé : 10^9 pages

- web : 10^{15} pages

x 1 million

Structuré ou non ?

- HTML URLs parsé 1,486,186,868
- Domains with Triples 65,408,946
- URLs with Triples 302,809,140
- Typed Entities 1,222,563,749
- Triples 3,294,248,652

Web Data Commons, http://webdatacommons.org/

Un système de publication actif

- Web Information Systems
- Contrôle par le producteur
- Publication continue (y compris pages anciennes ‘archivées’)
- Frontières de l’objet visé sont flou (un site?)

Conserver implique exactement l’ opposé

Le Web comme artefact culturel

- Multimédia, convergence de tous les type de contenus numériques
- Hypertexte actionnable
- Edité globalement par des centaine de millions de personnes

Conservation sans le filtrage traditionnel de l'édition

Cardinalité

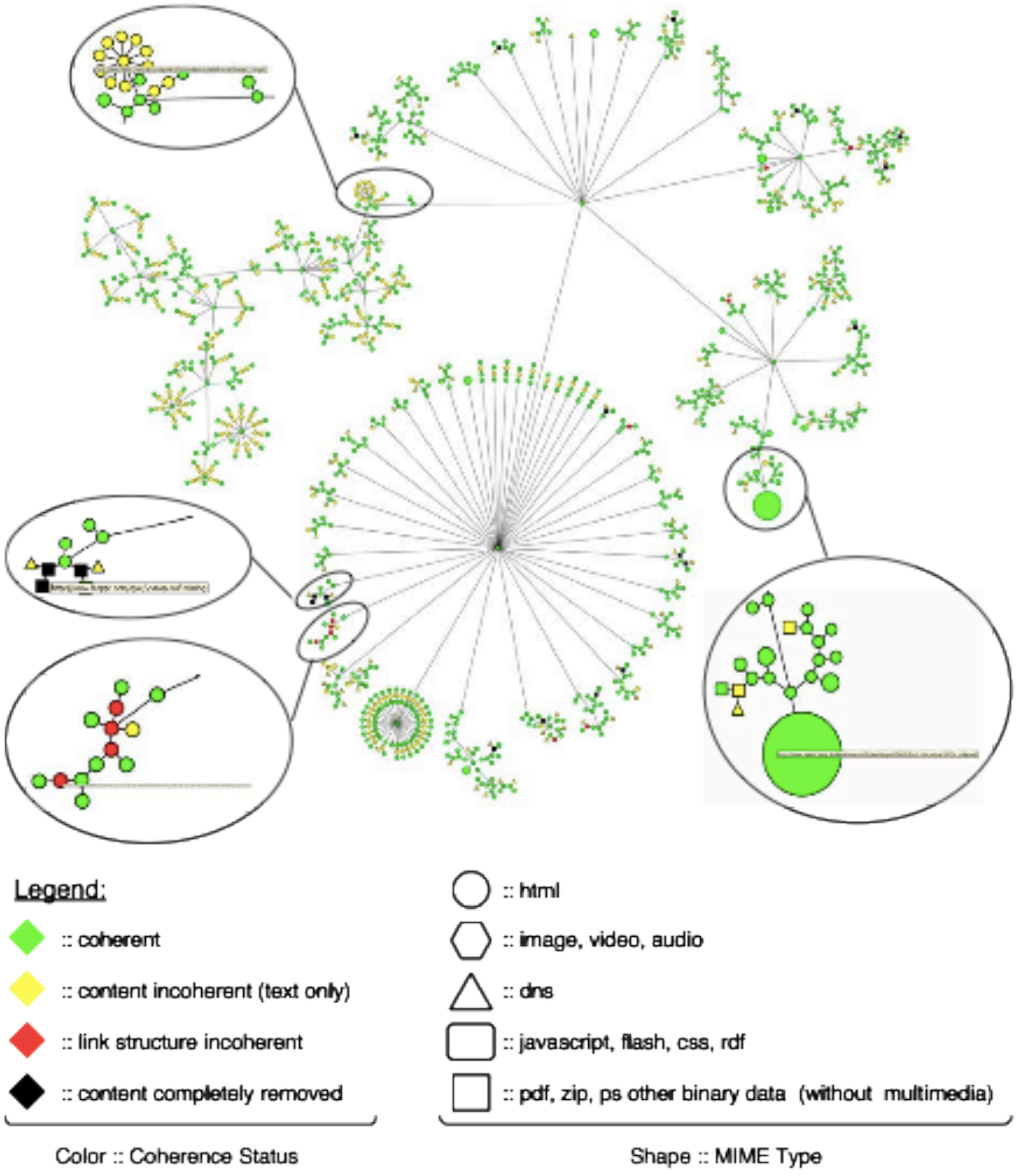
- Différent selon les institutions (musées, archives, bibliothèques)
- Cardinalité des incunables
  - 20 millions de livres
  - 30 000 éditions
  - 650
- Une cardinalité élevée donne deux avantages pour la conservation : la redondance et le temps

La cardinalité ‘paradoxale’ du Web

- Un nombre virtuellement infini de copies
- Mais une très forte dépendance à un serveur unique

Capture et cohérence

- extension temporelle incompressible des capture
- en contradiction avec la publication permanente
- risque d'incohérence temporelle au sein même de l'archive


Legend:
:: coherent
:: content incoherent (text only)
:: link structure incoherent
:: content completely removed
Color :: Coherence Status
:: html
:: image, video, audio
:: dns
:: javascript, flash, css, rdf
:: pdf, zip, ps other binary data (without multimedia)
Shape :: MIME Type

L’archive

Une mémoire de la toile

- Echantillonnage automatique raisonné et documenté
- Saisie d'un état
- Construction de séries temporelles pertinentes
- Inclusion dans l'internet

Une infrastructure pour la science

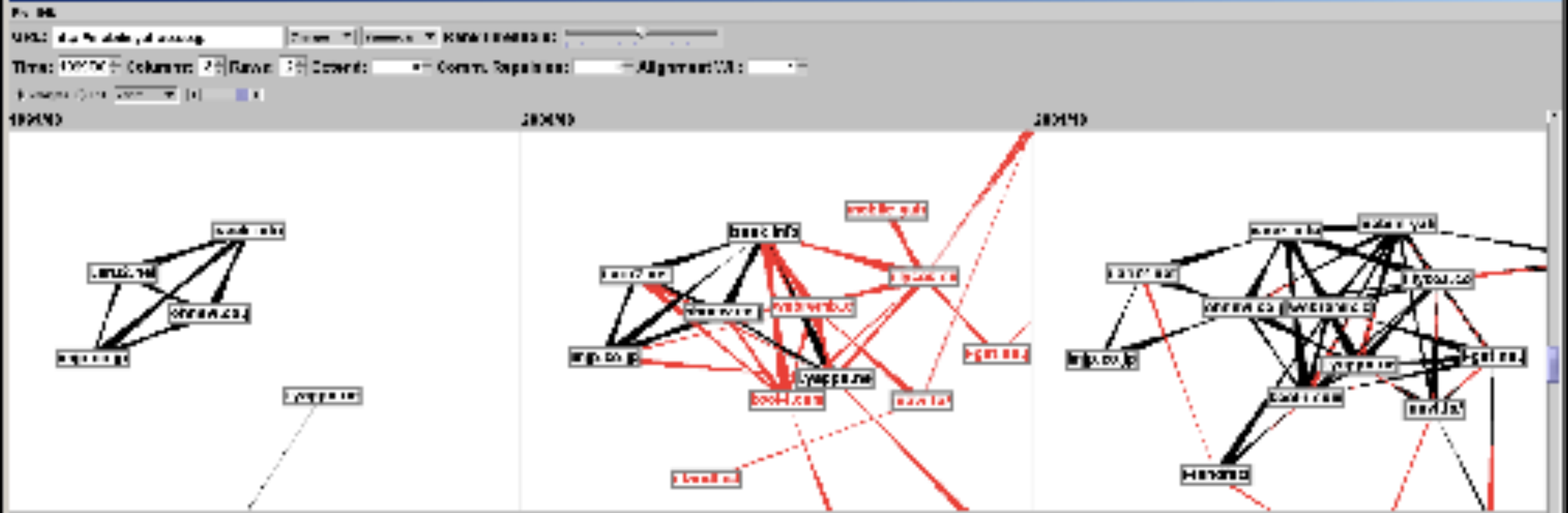
- rôle dans la construction du savoir
  - quel sera l'équivalent des bibliothèque et des archives pour le web ?
    - CERN de la Web Science
    - Inclusion dans l'internet

Internet Archive: http://archive.org/

Internet Memory : http://internetmemory.org

IIPC : http://netpreserve.org/

Bibliothèque Nationale de France : http://www.bnf.fr



- *M. Toyoda et M. Kitsuregawa, A system for visualizing and analyzing the evolution of the web with a time series of graphs, Salzburg, Austria: ACM Press New York, NY, USA, 2005.*

Quel régime d'archive ?

- ce que l'on garde ce que l'on ne garde pas (valeur) ?
- droit à l'oubli ?
- vie privée
- accès (humain/machines)
- ...

Julien Masanès
Internet Memory Foundation

internetmemory.org

Aux archivistes du Web