

Web archiving: what shall we preserve and how to make it usable

Catherine Lupovici

Head of the Digital Library Department

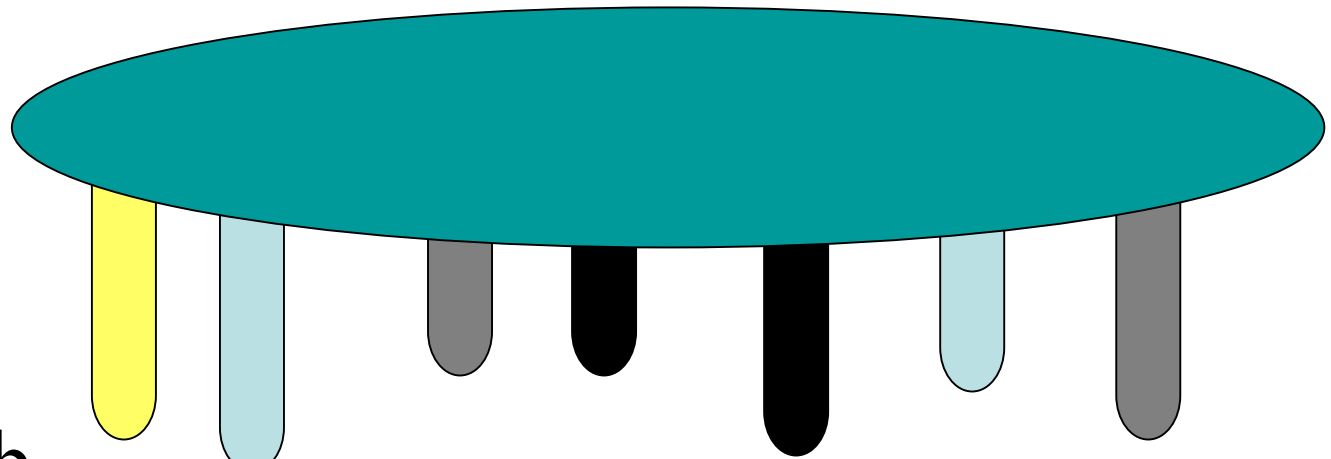
Bibliothèque nationale de France

The web content

- Web content typology
 - Classical publications (including self publication) & grey literature (text, images, sound and video)
 - Records of corporate bodies and Institutional Content
 - Radio & TV programs, broadcast and on demand
 - Services, Interpersonal communication
 - Emerging types: blogs and social web
- Web content technical characteristics & challenges
 - Mass, continuing resources and dynamic content
 - Interlinking
 - Surface web or visible web (open web content) and deep web (restricted access for the robots)

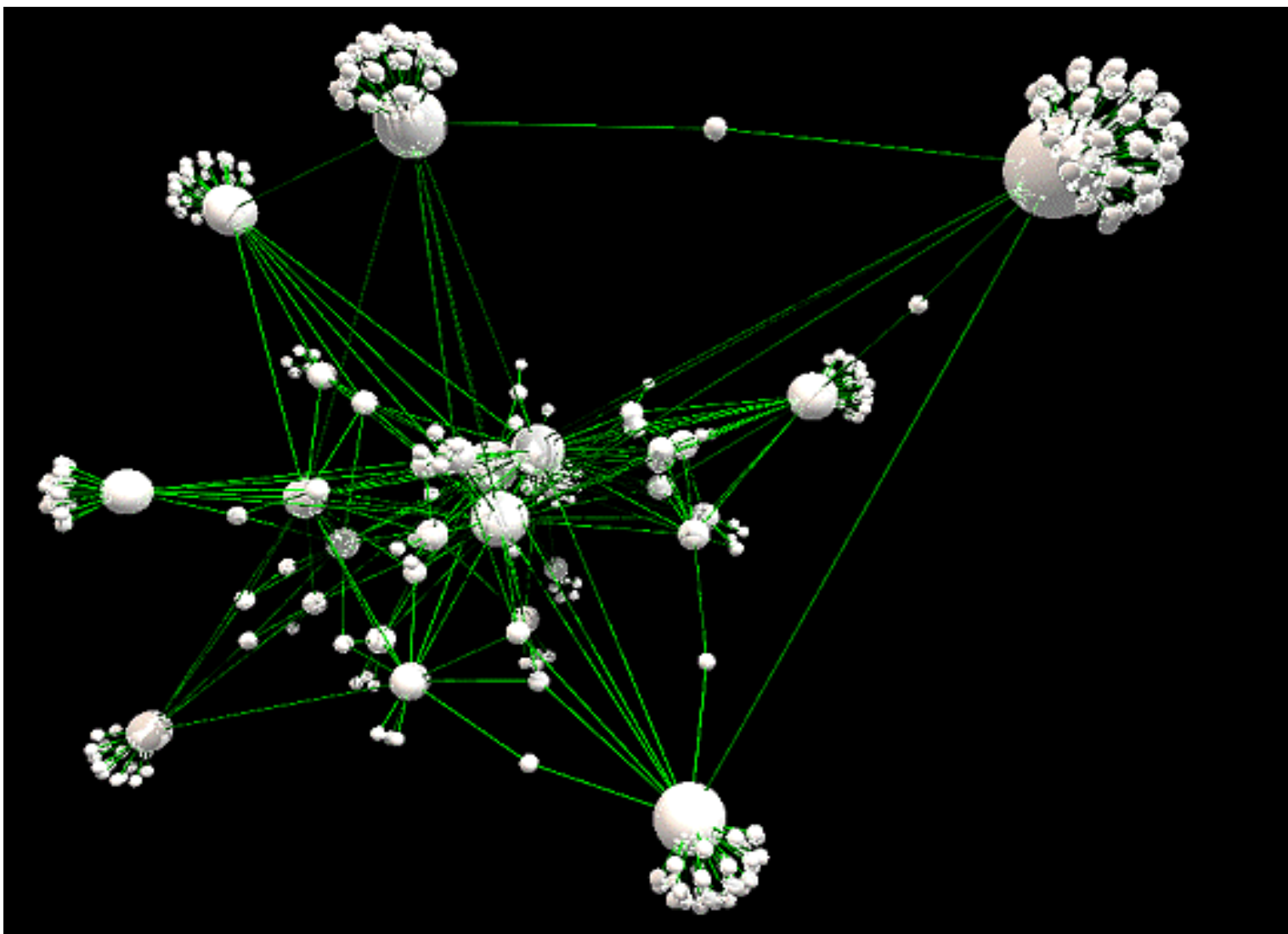
Surface web and deep web

- Surface open web



- Deep web

Links are part of the web content



Web size estimates

- OCLC Web Characterization study:
<http://wcp.oclc.org>

	1998	2000	2002
Unique sites	2 636 000	7 128 000	8 712 000
Public sites	1 457 000	2 942 000	3 080 000

- Indexable Web
 - 200 million pages in 1997 (K.Bharat and A.Broder)
 - 11,5 billion pages end January 2005 (A.Gulli and A.Signorini)
- A snapshot of .fr domain Dec 2004-Jan 2005
 - 121 million files, 500 000 hosts, 3 Tb

National libraries memory missions

- Preserve and provide access to what is made available publicly in the country. Legal deposit concept
 - Along with the legal deposit law possibility for exceptions to copyright for long term preservation (authorisation to bypass the technical protections)
 - Access authorised generally locally on site
- Legal deposit philosophy
 - Memory mission for future users: no selection when collecting or statistical sampling representing a class of resources
 - The opposite of collection development policy philosophy to fulfil the requirements of a current community of users

Collection policy

- Opposite historical approaches in early experimentation
 - Canada (1994) and Australia(1996) started with deposit of only digital resources (e.g. e-journals). As classical catalogued resources
 - Sweden with periodical automatic harvesting of the national web domain (1997). Domain centric policy. Not catalogued
 - Library of Congress thematic harvesting for presidential elections 2000 and 11 September 2001. Topic centric policy, some descriptive metadata
- Complementary approaches are requested
 - Harvesting is preserving the web inter-linking and navigation feature. Context has to be recorded at harvest time. WARC format
 - Deposit allows to get deep web in relationship with producers. The deposit must includes preservation metadata. WARC format

Users access to large scale collections

- Principles shared by institutions already involved in Web archiving at large scale
- Minimum requirements
 - Indexing and search by URI, by date of harvest
 - full text indexing and search
 - navigation through the archive by URLs and over time
- Need for smarter tools for automatic classification and semantic organisation of the collections

Uri: Search Viewing version 2 of 2
Jan. 13rd 2005, 11:52

Des. 29th

Jan. 28th



Resolution:

Days

Auto: He

Years

Months

Days

Hours

Minutes

WERA... External links, forms, and search boxes may not function within this collection. Uri: http://www.bnf.fr/pages/cultpubl/exposition_288.htm, time: 2005-01-13 11:52:09

[Accueil](#) > [Offre culturelle et éditions](#) > [Programme culturel](#) > Exposition

Exposition

> Sartre



09 mars 2005 - 21 août 2005

Site François-Mitterrand \ grande galerie

Philosophe, romancier, dramaturge, biographe, polémiste, journaliste et théoricien de l'esthétique, Sartre, dont on célèbre les cent ans cette année, a participé à tous les événements importants de son époque et a été de tous les combats pour la défense de l'individu ou des nations.

Tout en faisant appel à de nombreux documents audiovisuels pour recréer l'environnement quotidien et les grands événements du siècle, l'exceptionnel fonds Sartre du département des Manuscrits sera mis en valeur par des œuvres de peintres qu'il a fréquentés tels Giacometti ou Wols, et de photographes qu'il a connus comme Brassai, Cartier-Bresson et Gisèle Freund.

10h-19h mardi-samedi, 12h-19h dimanche

Tarif plein : 5.00 euros

History of the French cultural heritage constituted by legal deposit

- 1537 Creation of legal deposit in France. Printed material (books)
- 1648 Engravings including maps
- 1793 Musical scores
- 1925 Photographs and “graphic art productions of any type”
- 1941 Posters
- 1963 Sound recording
- 1975 Still images and videos (Any medium, any production technique)

- 1992 Off line electronic publications including “Software, Databases and expert systems”
- 2006 Web & possibility to ask producers to deposit the electronic files in place of the classical medium

Current web archiving in BnF

- One year snapshot of the French national domain. Broad crawl and focussed crawl
 - URL, date and full text indexing
 - Target two snapshots /year
- Thematic focussed crawls
 - Elections: presidential & legislatives, Oct. 2006-June 2007
 - URL, date, subject metadata at the selection time
 - Usage study with Institut des Sciences Politiques and restricted access in the reading rooms to the full collections
- Continuous crawl: *Journal officiel* since the first online issue, June 2004. Daily targeted crawl

Impact on preservation and access policies

- Digitisation of classical analogue collections
 - Mass digitisation will allow to use the digital reproduction for communication to users in place of the original
 - Local and distant access
 - It is reducing the pressure of preservation for fragile material
- Deposit of electronic files in place of the printed output urges to manage digital preservation of large scale digital collections
 - Trusted repositories providing risks management, preservation and access are becoming top priorities for cultural heritage institutions
 - Good access applications allowing to search and navigate through the digital collections not only through the mediation of descriptive metadata but also using smart tools for content analysis of contents and facet navigations



BnF

15 December

Cultural Heritage on line. The challenge of accessibility and preservation,

13

Collections size

- BnF, some figures
 - 10 million volumes of printed material since King François 1er
 - 80 000 volumes of printed material digitized since 1992
 - + 30 000 in 2006
 - + 100 000 in 2007 and following years
 - Legal deposit of printed material: 60 000 books + 30 000 serial titles published in France/year
 - Web archives as legal deposit extension to online public resources .fr domain harvesting
 - 2004 broad crawl: 121 million URLs, 500 000 hosts, 2,5 TB
 - 2005 broad crawl: 167 million URLs, 1 228 000 hosts, 4,1 TB
 - 2005 focused crawl zooming on blogs: 55 million URLs, 2 000 000 hosts, 1,33 TB