



IIPC activity: standards & tools for domain scale archiving

Catherine Lupovici

Program Officer

Bibliothèque nationale de France





International Internet Preservation Consortium

- IIPC launched in July 2003 <http://netpreserve.org>
 - 11 national libraries and Internet Archive
- Extension January 2007
- Technical standardisation at the domain scale: functional architecture and standard APIs, archival format (WARC format), metadata, permanent identification
- Strong basic tools for all the chain from acquisition to access processes
 - Open source, free licence tools
 - Generic tools that can be adapted to local policies
- Provide a forum for sharing knowledge about Internet content archiving both within the Consortium and beyond



IPPC content management approach

- Viewing the web as a whole at least at the domain level is the only way to record the real web
 - Broad extensive harvesting, focused intensive selection and harvesting, and deep web deposit are complementary techniques
- The web scale can only be handled by automatic processes using appropriate specific tools for acquisition and access
 - Cannot be achieved by scaling up item per item classical approach
- Archiving at large scale will allow future users to apply smart mining tools on historical archives.



Full set of tools for all the chain

- Focused selection and verification : curator tool development, BL, NZNL
 - First release expected by end of August 2006
- Acquisition
 - Heritrix crawler first released in January 2004
 - Current version 1.8.0 released 5 May 2006
 - Expecting version 1.10 delivering Warc format, summer 2006
- Indexing and access
 - Nutchwax: a Nutch-Lucene extension for Arc format and extended capacity to 100 million pages. Current version 0.6.0 released May 2006
 - Open source Wayback machine. First release Dec 2005. Current version 0.4.0 released March 2006
 - Wera interface : navigation principles. Version 0.4.1 released January 2006
- Download <http://netpreserve.org/software/downloads.php>



Users access to large scale collections

- IIPC simple access tool: WERA
 - search by URI, by date of harvest
 - full text indexing and search
 - navigation through the archive by URLs and over time
- <http://nwa.nb.no/wera>



WERA - Microsoft Internet Explorer

Fichier Edition Affichage Favoris Outils ?

Précédente Revenir Arrêter Rechercher Favoris Média Imprimer

Adresse <http://nwa.nb.no/wera/result.php?time=20050113115209&url=httpINDEX3AINDEX2FwwwINDEXDOTbnfINDEXDOTfrINDEX2FpagesINDEX2FcultpublINDEX2FexpositionINDEXUNC> OK Liens

Google wera demo Rechercher PageRank 18 bloquée(s) Orthographe Options wera demo

Uri: http://www.bnf.fr/pages/cultpubl/exposition_288.htm Go

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

Des. 29th Jan. 28th

Search: sartre Go

Resolution: Days Auto: ☒ Help


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

Démarrer Catherine LUPO... WERA - Micros... VALA Microsoft Power... Diaporama Pow... Microsoft Word FR 15:30



Some other European national policies

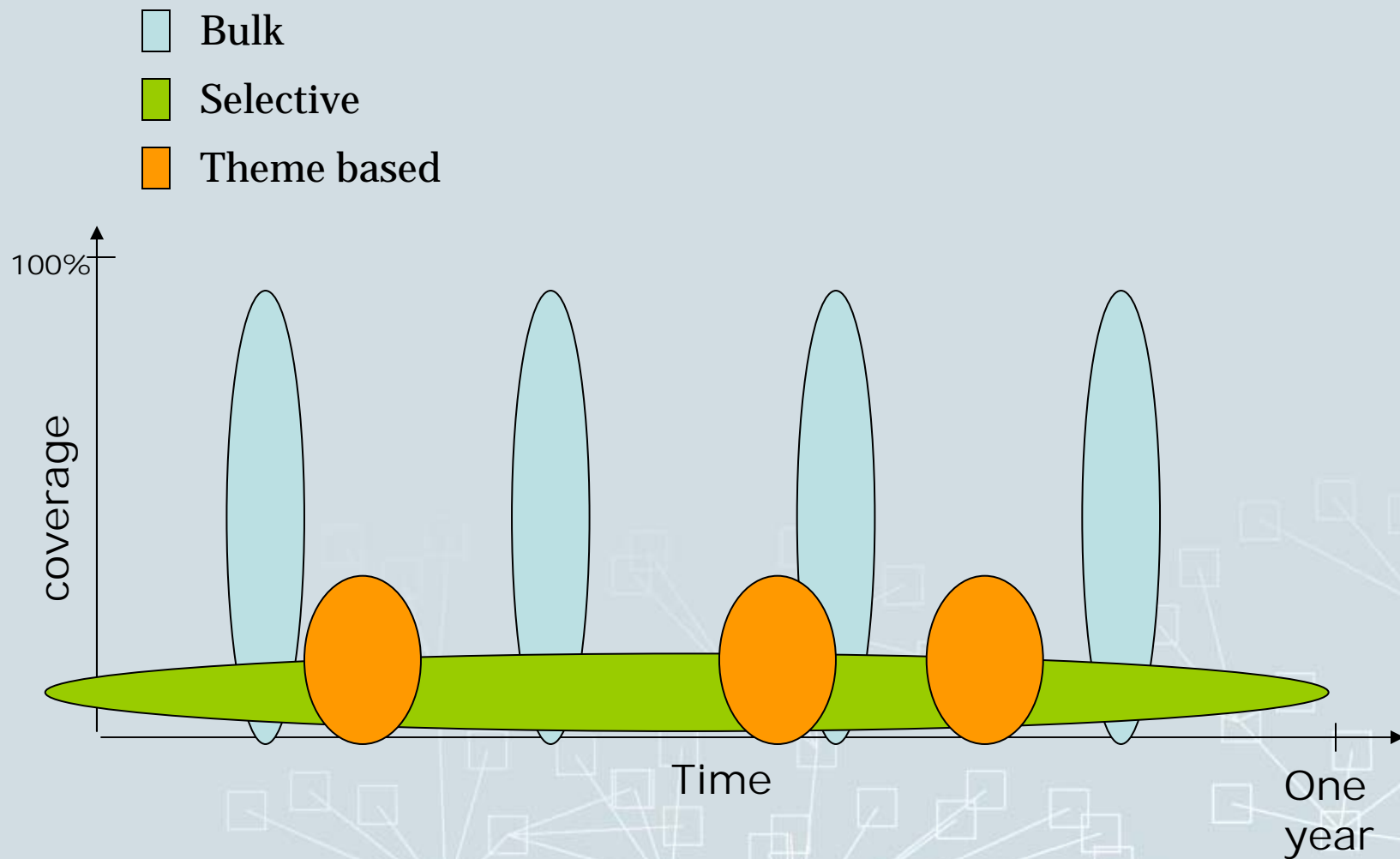
→ Denmark

- Extension of legal deposit law December 2004
- Beginning national domain crawl by the Royal Library/Aarhus University in July 2005. Harvesting only
- No users' access at the moment .
 - Request by the law to filter personal data
 - Internet access of few sites by contract with the producer

→ France

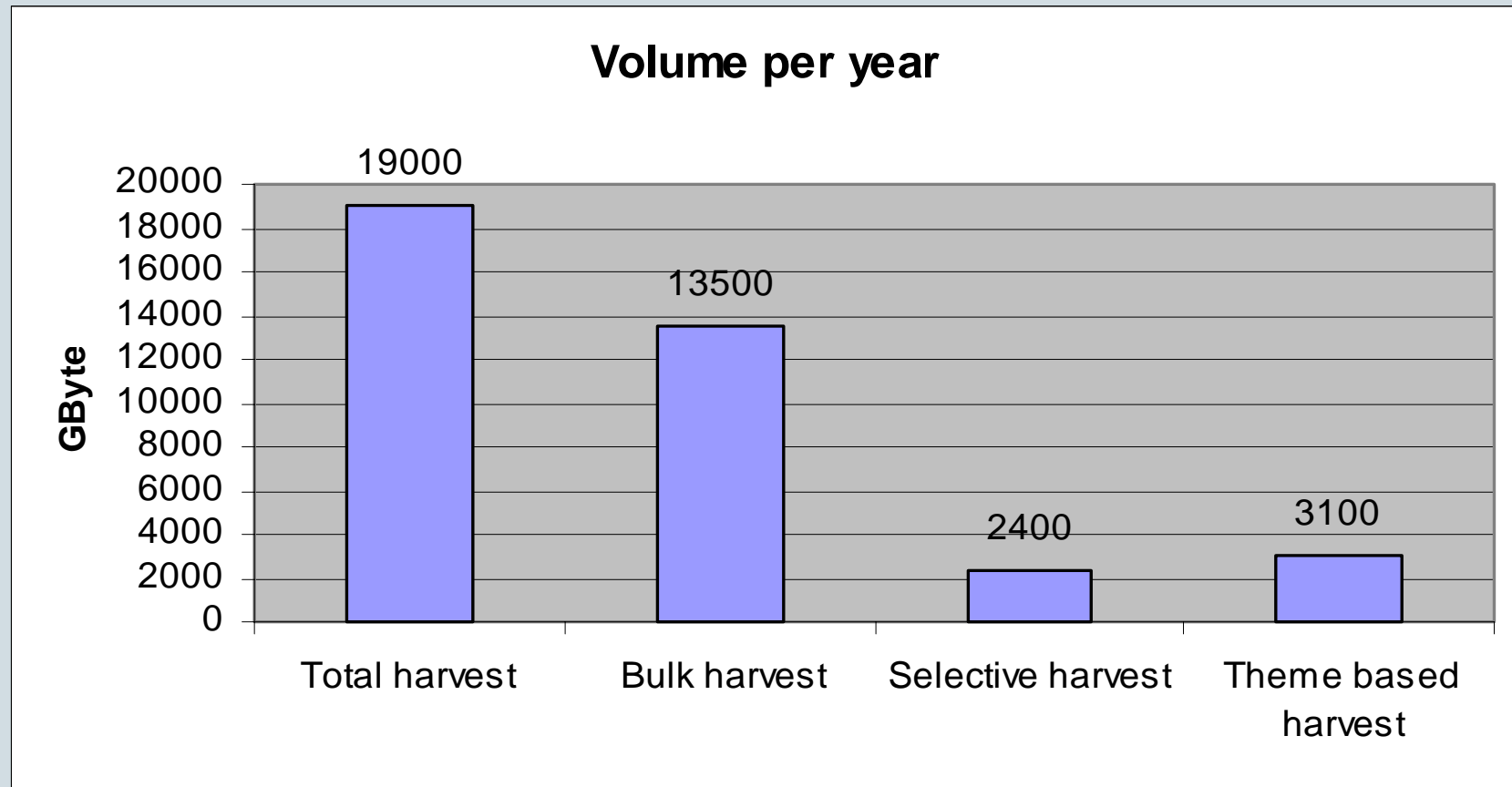
- Extension of legal deposit law under process
- Harvesting and deposit
- Intranet access only
- Broad crawl 2004, 2005, 2006 + events (elections 2002, 2004 and starting 2007 elections in next September)

Example of Denmark legal deposit harvest strategies





Volume of netarchive.dk





IIPC next phase programme

- The first three years of the consortium have been dedicated to the basic toolset creation along with standardisation activity
- The next IIPC phase will build on this first layer of tools for more sophisticated ones for acquisition and access.
- The work on digital preservation already initiated will be a key part of the future consortium activities
- The consortium will also work on archives interoperability