

# ***Access Working Group Progress Report***

**IIPC General Assembly**

**Singapore, May 2010**

**Helen Hockx-Yu**

**Kris Carpenter**

# Web archives registry

- Registry bringing together web archives of IIPC member organisations
- A simple schema to describe essential characteristics of web archives
- Launched on 22 December 2009
- Press release issued by IIPC
- 24 archives featured in the registry
- Synchronised with members information database
- Work finished but need to ensure institutional and web archive records are up to date

# IIPC members archive

mission  
who we are  
member archives  
for members  
join the iipc  
working groups  
press releases

**publications:**  
reports

**events:**  
conferences and  
workshops

**software:**  
toolkit  
downloads

## About:

## Member Archives

### Archive List

**Alphabetically** (All archives are open onsite and offsite (via the Internet) unless otherwise noted after the archive name.)

**Bibliothèque nationale de France - Archives de l'Internet (Bibliothèque nationale de France Web Archives)** - (Access is open to researchers onsite at the institution)

- **Collecting institution:** Bibliothèque nationale de France (National Library of France) - <http://www.bnf.fr>
- **Start Date:** 2002
- **Archive interface language(s):** French
- **Access methods:** URL Search, Keyword Search, Full-Text Search, Topical Collections
- **Harvesting methods:** National Domain, Bulk, Selective, Event, Thematic
- **Description:** Since 2006, BnF shares with INA responsibility for the legal deposit of the French online publications and web material at large. BnF web archiving program started in 2002 with election websites first snapshots, then continued from 2004 with a 5-years partnership with the Internet Archive, which included performing annual domain crawls of the French domain and acquisition of historical collections. Today, BnF is running both domain and selective crawls internally.

In 2010, BnF Archive consists of ca. 180 TB of data (13 billion files) from 1996 until now. The scope of this collection is the French web (.fr and beyond) and combines domain, thematic and event harvests. Special collections include a range of national, local and European Elections harvests, along with topical collections such as online diaries, blogs and literary websites or activist websites documenting the social history of the Web. 85 curators contribute to the selection of seeds, forming a collections in most areas of knowledge, in line with BnF encyclopedic heritage.

Due to legal restrictions, BnF web archives can only be searched and browsed by researchers within the library premises in Paris.



# Researcher use case

- Document / update researcher use case of web archives
- Several institutions contributing to this ongoing key strand of work
  - INA and CDL reviewing guidelines defined by a a NDIIPP-funded project on digital research requirements and will provide summary of recommendations and next steps
  - BnF plans for interviews and focus groups with current or potential users of their web archive, focusing on researchers' expectations of the content of the web archives (collections). A report is expected at end 2010/early 2011
  - BL is involved in a 3-year study on the research behaviour, tracking young doctoral students' (born after 1982) information-seeking behaviour, assessing their usage of library and information sources on and off line. Interim report now available
  - IA has committed to document past and present researcher use cases funded to date by JISC, NSF, and NEH by end 2010.

# NutchWax

- Commonly used by IIPC members to facilitate full-text search of archival web collections
- A number of strands
  - Backward compatibility with NutchWax 0.10 supported with the 0.12.9 release and up
  - The Diet Library of Japan developing support for multilingualization in Nutch. Localisation patches for Japanese, Korean and Simplified Chinese released for Nutch 1.0
  - Initial investigation of SOLR/Lucene and alternatives has begun due to issues of NutchWAX as a full text search solution. A report by the Internet Archive expected end 2010.

# Collaborative IIPC collection on Olympics 2010

- Goal of the project is to gather a collection of websites on Olympics as shared R&D resources for IIPC members, eg, to experiment with federated access. Also to highlight issues for the Olympics 2012 collection.
- 12 institutions contributed seeds for the project
- North Texas University provided a nomination tool for participating institutions to submit seeds
- IA did the crawls and will make accessible all materials harvested including extracted metadata and crawl reports by June 2010.
- BnF reported on their experience and contribution to the Olympics 2010 project, highlighting issues which they would like to see addressed for the 2012 collaborative collection.
- Planning for Olympics 2012 is expected to commence in Vienna.

# Goals for the GA

- Review progress and lessons-learnt
- Discussion of common issues and areas of interest
- Make recommendations and set priorities - work plan for the next 6-12 months