The International Internet Preservation Consortium (IIPC)

Catherine Lupovici

Program Officer

Bibliothèque nationale de France

IIPC objectives

- Provide a forum for sharing knowledge about Internet content archiving both within the Consortium and beyond
- Develop and recommend standards
- Develop interoperable tools and techniques to acquire, archive and provide access to web sites
- Raise awareness of Internet preservation issues and initiatives through conferences, workshops, training events, publications,...

IIPC

- Launched in July 2003
- Members :
  - Bibliothèque nationale de France, leader
  - National library of Australia,
  - Library and Archives Canada
  - National library of Denmark,
  - National library of Finland,
  - National library of Iceland,
  - National library of Italy,
  - National library of Norway ,
  - National library of Sweden,
  - British Library (UK),
  - Library of Congress (USA)
  - Internet Archive

Membership

- 12 members currently for the duration of the agreement (-> July 2006)
- New members application (end of 2005) for the second phase of the IIPC

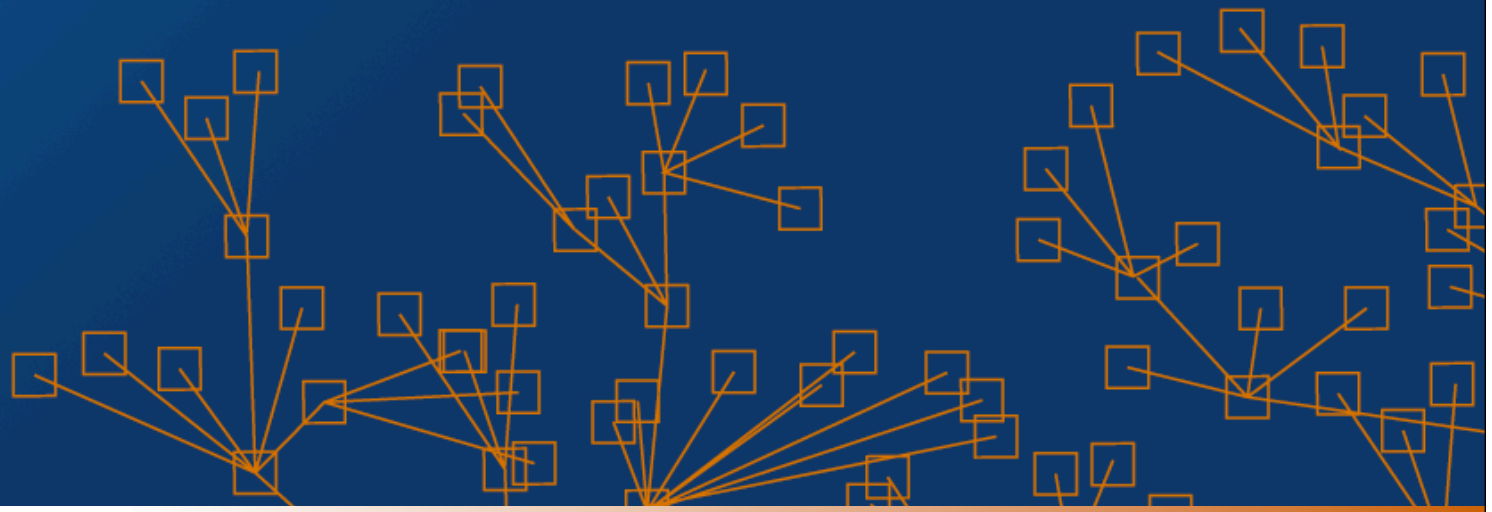
IIPC: a pragmatic approach

- Two levels of works :
  - working groups
  - projects accepted by the steering committee
- Deliverables expected:
  - tools released under open source free license
  - recommendations (methodologies, processes, standards,…)

Working Groups

- Framework: architecture and standards
- Metrics and Test Bed: defining and implementing a test ground for crawlers
- Access tools: development of a toolset
- Deep Web: development of tools for deposit and access of database-driven document gateway
- Content management: common vision of collections coverage and complementarity. Statistics.
- Researchers requirements: comments and advice on content and access

IIPC
Web Archiving Toolset



Full set of tools for all the chain

- Focused selection and verification
- Acquisition
- Collection storage and maintenance
- Access and indexing
- Download http://netpreserve.org/software/downloads.php
- See other presentations for Heritrix, WERA

Acquisition Chain (1)

- Smart Archiving Crawler Project
  - Specification in early 2003
  - Joint call for tender by BL and BnF
- Goal: to implement large scale, automatically focused crawls
- Priority based on citation linking and thematic assessment
- Call in October 2004, first prototype fall 2006

Acquisition Chain (2)

- Deep Arc
  - Specification in early 2002
  - Developed by bnF
- Goal: to allow site producer to easily extract DB to XML flat files
- Available for test http://deeparc.sourceforge.net/

Access tools

- DB query interface generator for databases stored as XML
- Developed by NLA with partial IIPC funding
- Xinq (XML INQuiry) http://www.nla.gov.au/xinq/

IIPC toolkit ready before mid-2006:

- Robust & scalable up the to the global web
- Implement IIPC standards (ARC 3.0, metadata, API…)
- Easy to install and use for advanced user (web archiving engineers)
- Open source and available for the community of web archives

Standards for web archiving and preservation

Standardization actions

- Functional modules and architecture with standard APIs
  - To allow interoperability with each institution system
  - Modular approach for including new capabilities
- Format for Web archiving and interchange
  - From ARC format to WARC, intended to be introduced as an ISO TC 46/SC4 standard
- Metadata for long term preservation
- Permanent identification

Conclusion

- The first three years of the consortium have been dedicated to the basic toolset creation along with standardization 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 also be a key part of the future consortium activities