International Internet Preservation Consortium (IIPC)

A short introduction

Þorsteinn Hallgrímsson
National and University Library of Iceland
IIPC Mission

To acquire, PRESERVE and make accessible knowledge and information from the Internet for future generations everywhere promoting global exchange and international relations
IIPC Goals

PRESERVATION of the WEB

- Collect and preserve a rich body of Internet content from around the world
- To foster the development and use of common tools, techniques and standards that enable the creation of international archives
- To encourage and support national libraries everywhere to address Internet collecting and preservation
IIPC Development

- July 2003 – IIPC Established (12 members)
- 2007 – 2009 – Phase 2: (38 members)
- 2010 – 2012 - Phase 3 (in preparation)

IIPC working groups – focus shift from phase 1 to phase 2

**Phase 1 (2003-2006)**
- Access Tools
- Content Management
- Deep web
- Framework
- Metrics and Test-bed
- Researchers Requirements

**Phase 2 (2007-2010)**
- Access
- Harvesting
- Preservation
- Standards
IIPC Projects - Accomplishments

Enhancements to the Heritrix crawler

WARC standard
  Currently in Draft International Standard approval process
  ISO standard next month

WARC tools

Web Curator Tool (New Zealand and British Library)
Netarchive Curator Tool Suite (Denmark)

Access tools
  NutchWAX for indexing
  Open Source Wayback for access and display
Collection building by harvesting the Web

Three Main Approaches / Criteria:

• Bulk
  • National domain, (.dk, .fr, .is)

• Selective
  • Legal constraints
  • Institution policy (philosophy)
  • Resources
  • Technology

• Event based
  • Election
  • Major sports event
  • Royal marriage
  • Hurricane Kathrina
Access and Preservation

Access

- Use same methods as in life Web
- Indexing
- Registration / Cataloguing does not work

Preservation

- Volumes
  - Billions of documents
- “All” existing formats
Web Archiving in Iceland

New legal deposit law on 1.1. 2003

- National Library shall collect and preserve the .is domain
- and Icelandica (no permission required)

Publicly accessible web sites requiring a password must allow the library to harvest the web site

Access to the web archive is not specified
Web Archiving in Iceland

Collection building, i.e. Harvesting
- Total .is domain – 3 times a year
- Selective – 40 websites weekly
- Events – elections 2006 and 2007

Key figures: 8 TB data, 400 million URL, 0.3 FTE

Public access is planned on December 1, 2008
- Elections 2006 and 2007
- Weekly collections 2006 and 2007
- 2-3 total harvests
Focus and challenges

Quality Assurance
  - Limited Resources

Full text indexing
  - Improved access (relevancy)

Preservation
  - Let others do it!
Archiving the UK Web

Helen Hockx-Yu
Web Archiving Programme Manager
British Library
Overview

- UK Web Archiving Consortium (UKWAC) initiative since 2004 to build a collective national web archive.
- Permission-based selective archive.
- Underwent major system / data migration.
- Archive contains over 3,700 unique websites and over 11,400 instances, measuring approximately 2TB of data.
- BL the largest collector: to date archived 1,853 unique websites, 5,264 instances, or 1TB of data
- Ongoing Web Archiving Programme: BL as the point of first resort for a comprehensive archive of material from the UK Web domain
The issue: lack of national legislation

- National legislation is the most effective solution to the legal problems faced by web archiving
- Legal Deposit Libraries Act 2003 and extension of legal deposit to non-print publications
- LDAP Web Archiving Sub-committee advising the Secretary of State on implementation of the Act: regulation-based harvesting and archiving of freely available online publications.
- Slow process with delays; earliest legislation expected April 2010
- Low response rate to the permission requests (25% success rate)
- Only a small fraction of the UK domain is being collected; valuable websites disappearing
Preserving web archives

- Digital preservation team responsible for long-term preservation and ongoing access for all digital content

- Web archive as content stream in BL’s Digital Library System (DLS): stores and preserves any type of digital material in perpetuity

- Newly recruited Web Archive Digital Preservation Project Manager to focus on preservation and long term accessibility of web archives:
  - Identify and embed preservation workflow
  - Document dependencies
  - Metadata
  - Preservability of formats
  - Participate in and contribute to IIPC digital preservation work
Web Archiving in Denmark

Birgit Nordsmark Henriksen,
Email: bnh@kb.dk
The Royal Library, Denmark
Web Archiving in Denmark

- Legal Deposit: Static net publications, 1998-2005
- Legal Deposit: Material published in (open) electronic communication networks for a Danish audience, 2005ff
  - 2008: 71 TByte of data; 2.2 billion digital objects from 800,000 active, Danish related domains; 5 FT staff


- Challenge: Access only for research or statistic purposes to all harvested material (Directive 95/46/EC protection of individuals w. regard to the processing of personal data)
Collection Policy in Netarchive.dk

Events:
- Creates a debate among the population and is expected to be of importance to Danish history or have an impact on the development of Danish society
- Causes the appearance of new web sites devoted to the event
- Is dealt with extensively on existing web sites

- Bulk – Quarterly - 56TByte
- Selective – 80 domaines – 9 TByte
- Event based - 6 TByte
Preservation Efforts in Netarchive.dk

- Bit Preservation in NetarchiveSuite
  - In Denmark configured with redundancy:
    - Geography
    - Hardware architecture and vendor
    - Storage media
    - Software (OS)
  - Active Bit Preservation based on checksum comparison

- Next: ARC => WARC migration & Characterisation of all digital objects w. Jhove
Overview of Web archiving at BNF

- French legal deposit officially extended to the Web in 2006. No permission required.
- BnF chose a blended strategy combining bulk and selective harvesting.
- Key figures: 120 TB data, 12 billion URL, 7 FT staff + 100 curators and partners involved.
- In-house access to the Web archives since 2008
Archives de l'Internet

Outils : Recherche par URL  Recherche par mot  Parcours guidés

Recherche par URL
Retrouver un site, une page ou un fichier en indiquant son adresse internet (exemple : http://www.inventaire-invention.com).
http://www.cofardoxme.com
Remonter le temps  Recherche avancée
Option
Limitier la recherche à cette année

Recherche par mot
Retrouver ces mots dans la partie indexée des archives (environ 5%, documents archivés en nov-déc 2006 et 2007).

Parcours guidés
Découvrez le contenu des archives et se familiariser avec les outils de recherche et de consultation.
Cliquer, voter : l'Internet électoral

Tous les parcours
Focus: the challenge of change

• Does Web archiving involve new skills and job profiles?
• How to combine and to scale Web technical expertise and collection expertise?
• Need for new, daily coordination between IT and collections
• Need to implement Web archiving innovation in Library organization and find best dissemination scenario
Role distribution at BNF

COLLECTIONS

- Reference, Audiovisual, Literature and Arts, Sciences, Law, Social Sciences, Philosophy, History, Maps, Music, Photographs, Performing Arts...

LEGAL DEPOSIT

- Legal deposit Board
- Digital Legal Deposit (4)
- Legal deposit of prints

IT

- Digital & IT Steering Committee
- Software Development (2)
- Hardware Operations (1)

Web curators Leads (15)
Collection Board

Subject or media Web curators (ca. 100): tools, workshops, tutorials, guidelines

Input from external partners and end users
Storage for access
Long term preservation repository
Preservation strategy

- Format migration from ARC to WARC
- Large scale data migration issues
- A necessary step before proper archiving and long term preservation strategies within BnF global digital Repository
Australia
Web Archiving Case Study – National Library of Australia

Colin Web(b)
Director, Web Archiving & Digital Preservation
National Library of Australia
cwebb@nla.gov.au
Country overview - Australia

• National approach, led by NLA

• Selective since 1996 (April Fools Day), with negotiated permissions, quality control, access (PANDORA)

• Domain harvests each year since 2005 (large – expect 1 billion files in 2008 crawl)
<table>
<thead>
<tr>
<th>Domain Harvest</th>
<th>2005 (4 weeks)</th>
<th>2006 (5 weeks)</th>
<th>2007 (5 weeks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique files</td>
<td>185,549,662</td>
<td>596,238,990</td>
<td>516,064,820</td>
</tr>
<tr>
<td>Hosts crawled</td>
<td>811,523</td>
<td>1,046,038</td>
<td>1,247,614</td>
</tr>
<tr>
<td>Size</td>
<td>6.69 TB</td>
<td>19.04</td>
<td>18.47 TB</td>
</tr>
</tbody>
</table>

**PANDORA**

<table>
<thead>
<tr>
<th>Files:</th>
<th>43 million</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size:</td>
<td>1.73 TB</td>
</tr>
</tbody>
</table>

**Domain Harvests**

<table>
<thead>
<tr>
<th>Files:</th>
<th>1,297 million</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size:</td>
<td>44.2 TB</td>
</tr>
</tbody>
</table>
PANDORA cf Domain Harvesting

Size in Terabytes

- PANDORA: 1.73 T
- AusCraw l05: 6.69 T
- AusCraw l06: 18.47 T
- AusCraw l07: 19.04 T
Country overview - Australia

• National approach, led by NLA
• Selective since 1996 (April Fools Day), with negotiated permissions, quality control, access (PANDORA)
• Domain harvests each year since 2005 (large)
• No legal deposit
• Desire for more curatorial ‘shaping’ and community input.
The challenges are interconnected

Sorting out -

• What do we **want** to collect & preserve?
• What are we **allowed** to collect & preserve?
• What are we **able** to collect & preserve?
• What can we **afford** to collect & preserve?
**Archiving the web?**

The Web is like ...a web, a net ...

...Spread in all directions and dimensions ...

...Growing in all directions constantly ...

...Consisting of bits that change all the time ...

...Including many parts we have no current means of capturing

...Of a size that takes many weeks for the most efficient harvesting tools to download even what we can currently copy from just the Australian domain ...

Are we “archiving the web”, or doing something else?
“Single biggest issue”

• Balancing breadth, depth, timeliness, accessibility – from a small and uncertain resource base
  (eg Online newspapers)
Preservation strategy, now and in the future

• Knowing what we have
• Understanding our dependencies and being able to recreate technical environment
• Collaborative development of linked tools
IIPC Preservation Working Group
Preservation Working Group - some context

- IIPC history and focus
- Ready for some focus on long term preservation
- San Francisco SC meeting – Jan 2007
- Face to face meetings, teleconferences, email discussion of papers, reports on tools and approaches
- Sub-groups on bit pres, access pres, organisational issues?
Preservation Working Group - brief from Steering Committee

To identify preservation standards and practices that appear to be applicable to web archives.
Preservation Working Group - some questions of interest

- Do web archives need different preservation approaches?
- What are the key risks for web archives?
- Are there existing standards & approaches we can use?
- What is vision of a preservation web archive?
- Impacts of scalability and diversity?
Preservation Working Group - some questions of interest (2)

- Do needs of massive archives match those of small scale selective archives?
- Can we propose preservation workflows for ingest?
- What supporting infrastructure do we need to manage preservation of web archives?
- Balancing a preservation focus with other IIPC concerns – should we draw boundaries?
Preservation Working Group — work plan priorities

1. Annual survey to document technical environment for web access
2. WARC issues – What pres specifications? What issues in converting to WARC?
3. Sorting out metadata issues
4. Work on preservation tools – evaluating, influencing, identifying gaps, developing
5. Progressing policy discussion – When is action needed? What losses are acceptable? …
Preservation Working Group — work plan priorities (2)

6. Sharing benchmarks for auditing our capability to sustain access

7. Workflows – proposing some generic and specific preservation workflows

8. Skills – strategies for skills development – IIPC fellowship? Staff exchanges in preservation?

9. Planning – what do we need to know to plan and take effective preservation action?
Preservation Working Group — work plan priorities — ways forward

► Real projects
► Discussion groups with deliverable targets
► Frequent interaction with Technical Committee and the preservation community.