World Library and Information Congress:
71th IFLA General Conference and Council

"Libraries - A voyage of discovery"

August 14th - 18th 2005, Oslo, Norway

Conference Programme:
http://www.ifla.org/IV/ifla71/Programme.htm

septembre 28, 2005

Code Number: **194-E**
Meeting: 150 SI - ICABS (IFLA/CDNL Alliance for Bibliographic Standards)

Web archives long term access and interoperability: the International Internet Preservation consortium activity

Catherine Lupovici
IIPC Program Officer
Bibliothèque nationale de France
catherine.lupovici@bnf.fr

Abstract: *The International Internet Preservation Consortium was created in July 2003 by twelve pioneer institutions already involved in Web archiving and including eleven national libraries. The objectives of the consortium are to provide a forum for sharing knowledge about Internet content archiving and to develop interoperable tools, methods and standards to acquire, archive and provide access to the archived web sites. At the end of the first three years of IIPC a full set of open source free tools for the whole processing chain will be released. The IIPC standardization activity covers the web archives format, the preservation metadata, and the permanent identification aspects focussing on very large archive specificity.*

I would like first to thank ICABS for the invitation to present the activity of the International Internet Preservation Consortium, which has convergent objectives with ICABS. This presentation will first provide a general information on the organization, objectives and results of the consortium then it will focus on the work on standardization for web archives long-term access and interoperability.

1. IIPC (http://netpreserve.org)

The IIPC was created by twelve members already involved in web archiving with the objectives to:
- 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 the archived web sites
- Rise awareness of Internet preservation issues and initiatives through conferences, workshops, training events, publications, …

1.1. IIPC organization

The consortium was launched in July 2003, and the members for the first three years phase are:
- Bibliothèque nationale de France, leading the project
- National Library of Australia
- Library and Archives Canada
- National Library of Denmark
- National Library of Finland
- National Library of Iceland
- National Library of Italy, Firenze
- National Library of Norway
- National Library of Sweden
- The British Library (UK)
- Library of Congress (USA)
- Internet Archive

The idea of the consortium comes from discussions between national libraries and research institutions already involved in Web archiving who exchanged their view and experiments during specific workshops organized by BnF during ECDL conferences in 2001 and 2002[1]. The need for the development of a smart crawler was identified during those exchanges and BnF proposed to COBRA and Bibliotheca Universalis members to build a consortium in January 2003. Twelve partners agreed to put resources on such a project and the agreement was finalized and signed in July 2003 for a first phase of three years. The partners decided to limit the first phase to the initial members in order to build the minimum set of tools with active pioneer partners within the very short time of the first phase. By the end of 2005, new members application will be publicized for extension of the consortium not only to other national libraries but also to other institutions like archives or research institutions.

1.2. IIPC work program

The consortium took a very pragmatic approach defining two levels of work for the partners. Six working groups were created involving the interested partners and leaded by one of them, each partner contributing with its own resources. In addition specific projects are accepted by the consortium for support including possible partial funding. They have to involve at least two IIPC members and of course to contribute to the consortium objectives.
The deliverables expected include tools released under open source free license, and recommendations on methodologies, processes and standards.
The consortium steering committee approves the strategic objectives of the working groups, the support of Projects and the type and level of support the Consortium will provide to them. It reviews the progress of the Working Groups and the Supported Projects.

[1] http://bibnum.bnf.fr/ecdl/index.html

1.2.1. Working groups on methods and standards

- **The framework working group**

The framework working group objectives are to define the general framework for all the technical consortium activity in order to ensure cross interoperability for archives in the future and to prepare the appropriate international specific standards that will be needed. This working group prepared, at the very beginning stage of the consortium, the general high level architecture on which all the other working groups and projects are building. The functional modules defined are focused to web archiving activity and will have to exchange via APIs with each institution system depending on local implementation.

- **The Metrics and Test Bed working group**

The objective of the Metrics & Test Bed working group is to define and implement a test ground for crawlers in order to assess a specific crawler against the different harvesting problems that can happen due to content formats and gateways to content inside web sites. Those technical limitations where identified from previous existing works and a survey made by the Library of Congress and are published on the consortium web site[2].

1.2.2. Working groups on tools development

- **The Access tools working group**

The Access tools working group started with use cases in order to define the different requirements for accesses. Access for curators during the selection phase, access for content management purposes, access for digital archive management and last but not least access for the end user. For the first phase of the consortium it has been agreed to focus on the end user access and to develop a tool set covering the large scale indexing aspect and browse capabilities to navigate inside an archive.

- **The Deep Web working group**

The Deep Web working group is focusing on identifying the deep web and on the development of tools for deposit and access of database-driven document gateway. It is important for institutions planning to handle the deposit of deep web content to preserve the descriptions stored in databases with the digital objects they describe. This working group has already finished his initial work. It focuses now on exploring and assessing the technological evolution on general web search engines for indexing deep web.

1.2.3. Working groups on policy and requirements

- **The Content management working group**

The Content Management working group was created to enable the partners to share a common vision of collection coverage and facilitate each repository to complement the others. The collection coverage can be understood differently depending of the collection policy, which can be intensive in depth collection of a small number of sites or extensive transversal collection process of a very large number of sites. It can also be topic centric (like for instance a specific election) or domain centric (for instance a national domain like .se). The work was refocused on a part of the requirements to facilitate the development of a core set of tools used in the different approaches.

- **The researchers working group**

The researchers working group was created to provide comments and advice on what to include in web archives and how to index web archives. For instance would it be of interest to record the in-links data with the web sites as context information? The researchers will be invited to assess the end user access tools.

A group of 40 researchers represent the various Internet-related fields like media studies, literacy, political science, history, and computer science. Some of them are connected to the Association of Internet Researchers leading to synchronize the activity of the working group with the regular international conferences of the AoIR.

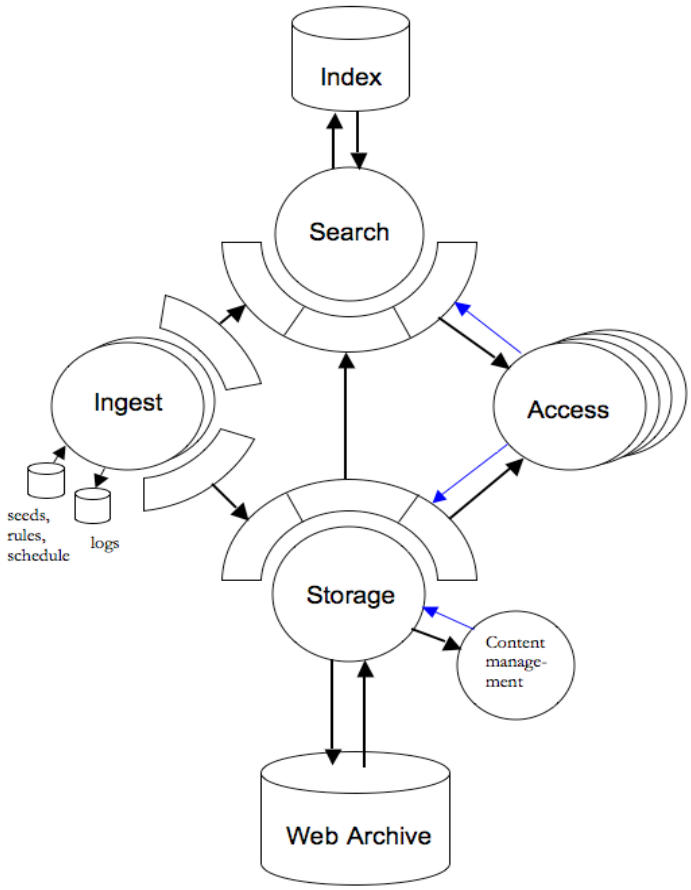
[2] http://netpreserve.org/publications/reports.php

1.3. IIPC web archiving tool set

At the mid term of the consortium, and building on the stabilized common architecture defined, the steering committee decided to put the efforts on finalizing a full set of tools to be released at the end of the first phase of the consortium. The objective is to cover the core functions of web archiving on the four following aspects:

- Focussed selection and verification to feed the acquisition process
- Acquisition by harvesting or deposit
- Collection storage and maintenance
- Access

The high level architecture of the tools is the following:



The IIPC toolkit covering the whole processing chain will be ready before mid-2006. The tools will offer the basic functions and will be robust and scalable up to the global web. They will implement the IIPC standards for the archival format, the metadata, the APIs. They will be easy to install and use for advanced users as they are designed for web archiving engineers. They will be open source and available freely for the community of web archives. Some of the tools are already available for test and

validation. The already available acquisition tools that allow the IIPC standards implementation for long term preservation and interoperability are:

- **Heritrix** is a large-scale crawler developed jointly by Internet Archive and the Nordic libraries on specifications written in early 2003. Heritrix brings to the web archiving community the capability to improve the finding of paths to content beyond the HTML link encoding, like CSS or Flash. It is basically configured to broad crawl but also allows setting parameters to focus the crawl in different way including site priority[3]. Heritrix generate directly files in the Arc format adopted by IIPC. Tools allowing the management of the files stored in the Arc format are also available on the Heritrix wiki site[4],

- As part of the contents that cannot be harvested because of crawlers limitation the Deep Web working group identified the databases acting as gateways to the contents as important descriptive metadata to archive together with the related contents. In order to provide a long-term preservation and access to the databases, they need to be migrated in an open format before ingesting them into the repository. BnF developed the **DeepArc** tool in 2003 for an experiment on voluntary deposit. It was discussed in the consortium and accepted as an element of the toolset. It is available for test as free license software on Source Forge[5]. The tool offers a graphic user interface to write the mapping between the database and a target XML Schema and the migration of the database into a flat XML file appropriate for preservation and possible reuse in the archive access context.

2. Standardization activity

The IIPC standardization actions are focused on:
- Standard APIs between the functional modules
- Format for web archiving and interchange
- Metadata for long term preservation
- Permanent identification

2.1. Standard APIs

IIPC is working on standard APIs between the functional modules in order to build an open modular architecture able to inter-operate with each institution existing system if it is itself open enough. One main area already identified is the description aspect and the relationship with the Library Management System for libraries and specifically with the acquisition, cataloguing and OPAC modules. But in the future with the extension to other institutions like Archives or Research institutions it will have to interact with other systems.

Standard APIs are also very important in this period of very fast technological evolution to allow replacing separately modules by more sophisticated ones as soon as new tools will become available.

The standard APIs will be released on the consortium web site. The first ones will concern the access function and all the interactions between the access, search and storage modules for end user full text search and browse capabilities.

[3] http://sourceforge.net/projects/archive-crawler
[4] http://crawler.archive.org/cgi-bin/wiki.pl?BnfArcTools
[5] http://deeparc.sourceforge.net

2.2. IIPC format for Web archiving and interchange

The Heritrix harvester has been developed to produce directly harvest results in the Arc storage format adopted by IIPC as a storage and exchange format. The access tool under development and test by the consortium is indexing Arc files.

The Arc format has been used by Internet Archive to store their archives since the beginning in 1996[6]. The IIPC members using the Heritrix crawler currently use it.
The Arc file format consists of a sequence of URL records. Each record starts with a header containing metadata about the harvesting technical context coming from the HTTP protocol exchange between the crawler and the host, followed by the file corresponding to the harvested URL. The size of an Arc file is between 100 to 600 MB and allows in the context of Web archives to minimize the disk access and to avoid handling specifically a large amount of very small files.
A DAT file is associated with each Arc container providing metadata (dates, checksum…) and mainly the list of URLs included in the container.
A CDX file is associated which each Arc container providing an index of all the URLs with the position of the beginning of the record an its length inside the container. It allows retrieving directly a specific URL for passing it to the access interface for display and to record and preserve the real URLs without need to transform them to comply with any file system naming rule.
The current format features make it very effective for storage and retrieving files over a very large archive. For instance a 2004 BnF snapshot of .fr domain contains around 25 000 Arc files corresponding to 500 000 different hosts and more then 121million URLs.

The consortium is working with other interested communities on an extension of the Arc format to accommodate larger requirements. This format called Warc is intended to be introduced to ISO TC46/SC4 as a format for web archives as the custodian institutions are interested by having an international standard format on which they can build trusted repository and trusted exchange of data as they have been used to with other types of data.

The requirements that the Warc draft format addresses are first of all the result of the work done inside the consortium within the framework defined at the beginning:

- an extension of the current features to include deposited sites as well as harvested ones in the same storage application and with the same access requirements in order to cover the broad crawl, the focused crawl and the deposit approaches
- the capability to manage metadata and content objects in the same format in order to facilitate storage and access to incrementally harvested material which leads to manage more relationships between objects that complement others due to the fact that the web contents are not stable over the time. The access tool developed by the consortium also offers navigation through a time line and would like to offer comparison facilities between "same" objects at different time, which impacts the format requirements
- extension to provide the possibility to segment a single very large object over several Warc files

The current Warc draft also addresses preservation requirements with the provision for preservation metadata that could be generated separately of the object record and the possibility to create converted records for instance after the migration of an existing record's content.

The non-IIPC members associated to the discussion are the Los Alamos National Laboratory, who did for the Library of Congress a study on the extension of the current Arc format, and the California Digital Library. The NARA (US National Archives and Records Administration) who is involved in digital preservation and contributed to the OAIS reference model standardization also uses the current format for web archives.

[6] http://www.archive.org/web/researcher/ArcFileFormat.php

2.3. IIPC web archiving metadata set

IIPC is working on a metadata set addressing the long-term preservation requirements and does not work currently on the descriptive aspects. The work is focusing on documenting the technical dependencies having in mind the specificity of web contents manifested in large archives containing a huge number of files inter-linked. Archives containing a small number of sites have preservation metadata requirements closest to discrete electronic objects that are already discussed in other groups.

On the web different formats are often combined in single pages and almost all existing formats can be found there. But over 90 % of objects only use four main formats publicly documented (HTML, JPEG, GIF, and PDF) and most of the remaining ones are not dependent on a specific operating system. In the web context, reliable format information is the fundamental Representation Information of the OAIS reference model that is needed. It cannot be built on the MIME type provided by the server.

Web archiving even when doing broad crawls means sampling the real web, which can change very rapidly. In addition web crawlers execute a policy that can be based on various parameters from server's TLD to citation linking, textual content, detection of changes etc… All those data are part of the context information of the harvested material, which have to be recorded to help the future user to understand what the harvested content represents compared to the original content.
So the IIPC metadata set includes the selection context, the interaction and transaction context elements between the client and the server (i.e. ID, cookies etc.). It constitutes an extension of the Preservation Description Information (PDI) of the OAIS reference model adapted to Web archiving.

In addition the granularity level to which to apply the different metadata has to be finalized. The physical objects defined by the Warc format are the Warc file and the record file levels. But the Metrics working group is still discussing about logical and intellectual entities like a site, a page, and a document ore more technical ones like a host. So the discussion is not yet finished. IIPC already recognized the need to discuss with the PREMIS experts who approached the Web site intellectual granularity more in relationship with the descriptive level.

The final decision on which application level for each metadata is also related to the metadata generation cycle and what can be automated during the acquisition phase including the crawling process. This acquisition phase is discussed by IIPC Content management working group as a pre-ingest phase generating an OAIS SIP (Submission Information Package). The next step will more precisely look at the ingest phase and the AIP (Archival Information Package) generation.

2.4. Identifiers

The identification is also discussed in close connection with the web archives framework and the Warc format.

At the Warc file level IIPC members agreed to have a common framework for naming, which will allow mirroring and redundancy between different repositories to minimize the disaster risk and to facilitate cooperation on the preservation effort. The attribution of the identifier of a Warc file will be under the responsibility of the institution creating it. The structure of the identifier is not yet finalized.

At the Warc record level the current discussion is about having a globally unique identifier, but questions where raised about the coherence with the identification of other resources within institutions.

Conclusion

The first three years of IIPC have been dedicated to define a general framework and adopt the corresponding standards for the development of a full set of specific tools for the whole processing

chain of web contents from the selection to the acquisition, the storage, the indexing, and the access by the end user.

One of the possible focuses for the next phase and the extension of the consortium to other type of institutions are the long-term preservation and the trusted repository. The Content management working group is now going to consider it soon along with the metadata set generation cycle.