The International Internet Preservation Consortium (IIPC)

The International Internet Preservation Consortium (IIPC) was formally chartered in Paris on July 24, 2003 with 12 participating institutions, i.e. the national libraries of Australia, Canada, Denmark, Finland, France, Iceland, Italy, Norway, Sweden, The British Library (UK), The Library of Congress (USA) and the Internet Archive (USA). The National Library of France is the administrator of the consortium.

The members acknowledged the importance of international collaboration for preserving Internet content for future generations and the mission of the IIPC is to acquire, preserve and make accessible knowledge and information from the Internet for future generations everywhere, promoting global exchange and international relations. In order to achieve its mission, the IIPC is working to accomplish the following goals:

- To enable the collection of a rich body of Internet content from around the world to be preserved in a way that it can be archived, secured and accessed over time.
- 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 archiving and preservation.

The initial agreement is in effect for three years, i.e. until July 2006. During that period the membership is limited to charter institutions, but the IIPC seeks to involve national libraries everywhere and will in the summer of 2005 welcome inquiries about future membership. The members agreed jointly to fund and participate in projects and working groups to accomplish the goals of the IIPC.

Background

The Internet is a specific medium combining many of the attributes of books, journals, radio, images and video and the World Wide web (web), that has grown exponentially since 1994, incorporates all those elements. What mainly distincts it from other media is its proliferation, enormity and accessibility. Almost anybody with a minimum of skill and effort can publish documents on the web and this results in billions of documents that can be accessed by most people from anywhere in the world.

The rational for archiving the web is basically the same as used by most countries to preserve and provide access to their cultural and intellectual heritage by collecting it and storing in museums, archives and libraries. It is obvious that currently, and increasingly in the future, a large and significant part of our culture will only exist on the Internet. If the traditional axiom of the Legal Deposit laws and other collection activity holds true it is therefore an absolute necessity to extend this concept to the web of the Internet. If this is not addressed now an important part of our culture, together with most documentation of the cultural change involved, will be lost.

In most countries the national libraries have traditionally collected manuscripts and published printed material. As publishing technology has progressed the libraries extended their collection activity by including physical electronic media like CD-ROM'S and some electronic publications like electronic journals. The main instrument enabling the national libraries to do this economically and comprehensively has been the legal deposit law of each country. Until now this has served the libraries very well but the emergence of the Internet with its widespread connectivity and the web interface has changed this.

In order to reflect this a few countries have changed their legal deposit law, thereby enabling the collection of material published on the web and several are in the process of changing their legal deposit laws. Access to this material will depend on the copyright laws in each country.

A couple of years after the web started to gain popularity it was discovered that it is very volatile and that a lot of documents and information in the web is short-lived and disappears after a short time. As a result initiatives for collecting and storing web material were started in 1996 by the national libraries of Australia and Sweden and in the USA by the Internet Archive (IA) a private not for profit company. In 1997/1998 the Nordic national libraries of Denmark, Finland, Iceland and Norway started web archiving initiatives and they have together with Sweden cooperated in several web archiving projects. Since 2000 many new initiatives have appeared all over the world: USA (Library of Congress), France, United Kingdom, Slovenia, Check Republic, Austria, Japan, Lithuania, New Zealand and recently China and Greece.

The first web archiving initiatives had very different priorities and focus. The IA was (and still is) the most ambitious with the goal of collecting as much as possible of the global web, in Sweden the focus was on the national web and in Australia the focus was on a limited number of selected web sites.

In 2000 the Library of Congress did a cooperative project with the IA and as a result the IA realised that it shared common interests with those national libraries that were interested in collecting and archiving the Internet: IA had the technical know-how and the national libraries had expertise in collection building. At the IFLA conference in Glasgow in 2000, Brewster Kahle, director of Internet Archive held a half-day workshop, organized jointly with the Section on Information Technology, on "Capturing the web: learning from experience in the National libraries". Brewster Kahle's very engaging presentation of pioneer work in web archiving and his presence at a meeting with some of the largest national libraries launched an initiative that became the IIPC. Very soon after the IFLA meeting the IA proposed a project with a few libraries, amongst them the BnF, the BL, the LoC and the KB in Denmark, to explore the possibility for cooperative efforts in archiving the web. Those parties plus people from Italy and the Nordic national libraries met in Rome in September during the 2002 European Conference on Research and Advanced Technology for Digital Libraries (ECDL). In the first half of 2003 the parties met in Paris, Washington DC and Stockholm and in July the IIPC was chartered at a meeting in Paris.

IIPC goals and organisation

The IIPC members acknowledge that in order to preserve and provide access to essential material accessible through the Internet, documenting history, culture and civilization, it is necessary to invent a new way of collaborating among national heritage institutions. The nature of Internet publication creates a huge web of ephemeral and linked documents, ignoring borderlines, that make it necessary to build non-isolated collections. On the other hand, the huge amount and the complex nature of content at stake, makes it impossible for a single institution to ensure long term preservation of this material. Therefore, IIPC members intend to create a network of heritage institutions that will have the challenging task to create collections shaped as part of a virtual global distributed collection to ensure that the distributed and linked nature of the original web material is not lost forever.

IIPC members will define accordingly requirements for a distributed stewardship and cross-access services to this content and, starting from there, lay the ground for a new collaboration mode among heritage institutions. IIPC members consider that this evolution is needed to achieve in the Archived Information Space a level of interconnectedness and navigability that has been achieved in the Publishing Space with the emergence of the Internet infrastructure.

In short, the counterpart of the global information space provided by the Internet has to be a Global Distributed Collection preserved by long term sustainable institutions. Preparing the evolution of national heritage institutions to achieve this is the objective of the IIPC.

Key goals

Building the global distributed collection

- Identify, assess and when needed develop appropriate standards and tools. Particular emphasis will be devoted on selection policy, structuring of material and description standards to ensure future cross accessibility to the collections.
- Develop a global view or map of collection's custody across the world.
- Define and test new collaboration model between institutions to share operational tasks at any level deemed appropriate.

Preserving the global distributed collection

- Develop risk assessment procedure and action plan for collection at risk. Identify, assess and when needed develop collaboratively appropriate tools to secure long term access to the material.
- Define and test new collaboration models between institutions to increase the safety of collections including define requirements and assess collections swapping models.

Key objectives

- Collaborative working, within each country's legislative framework, to identify, develop and facilitate implementation of solutions for selecting, collecting, preserving and providing access to internet content.
- Facilitating international coverage of internet content archive collections within national legal frameworks and in accordance with individual national collection development policies.
- International advocacy for initiatives that encourage the collection, preservation and access to internet content.
- Providing a forum for sharing knowledge about internet content archiving both within the Consortium and beyond.
- Developing and recommending standards.
- Developing interoperable tools and techniques to acquire, archive and provide access to web sites.
- Raising awareness of internet preservation issues and initiatives through conferences, workshops, training events and publications.

Organisation

The IIPC is governed by a Steering Committee where each member has a seat. The BnF member is the chair. The detailed work is carried out through working groups to define Policy, Requirements, Methods, Standards and Tools for Internet archiving. By this means projects will be developed and defined and will ultimately lead to the creation and provision of the necessary tools to fulfil the vision of universal coverage of internet archive collections. Each member must participate actively in at least one working group. The IIPC has established 6 Working groups: Access Tools, Content Management, Deep web, Framework, Metrics and Tested and Researcher Requirements. They have met five times, in San Francisco, Florence, London, Ottawa and Reykjavik, with a meeting planned in Washington DC in October 2005. Following is a short description for each group.

The ***Access Tools Working Group*** focuses on initiatives, procedures and tools required to provide immediate access, and to preserve the future access, to Internet material in a web archive. This includes identifying and specifying the various requirements the access tools must support.

The ***Content Management Working Group*** on one hand focuses on developing and maintaining a model for identification, description and assessment of content stewardship, and

on the other hand on providing a curator tool for controlling and scheduling the collection of web content.

The ***Deep web Working*** **Group** focuses on identifying strategies and producing tools for archiving web content which is inaccessible to web harvesters or has been deposited at the archiving institution.

The ***Framework Working Group*** focuses on creating a shared technical basis for web archiving activities. This common framework will take the form of architectural guidelines, standard formats and terms, and system interface specifications. The goals include enabling technical interoperability between member libraries, and reusing existing open standards and models, such as the OAIS model wherever practical.

The ***Metrics and Test bed Working Group*** focuses on defining metrics for web archiving and defining and characterizing an evaluation process for the coverage and performance of web archiving tools and processes. This includes identifying a test bed to use for evaluating web crawlers and to define the set of metrics for use in that evaluation.

The R***esearchers Requirements Working Group*** focuses on taking advice and discussing with researchers working in the area of Internet research in order to elaborate a common vision on what web archives should contain, their scope, extent, updating, structuring and associated information.

Results

The consortium has, when this is written, existed for almost 2 years and about 25 people have been involved in the working groups. Most of them have participated from the beginning with the result that the meetings are very effective. With more people involved this would not be the case.

The results are both tangible and intangible. The intangible ones are somewhat difficult to measure but they have proven very important. The IIPC forum now has a common understanding of the many complex issues that pertain to web archiving although the members do not agree on all of them. Most of the groups have delivered papers that either clarify or explain certain issues or propose best practices or standards to be used. They will provide the foundation for building future software tools that are needed to work efficiently with the web archives. Some are for the time being internal documents, but others will be made public and hopefully adapted by the international community. Producing those detailed descriptions often has the hidden value of highlighting the problems by forcing those involved to define the details.

Following is a list of some of the documents that have been produced:

- Use cases illustrating common understanding of the functionality of a web archive.
- Overview of state-of-the-art web archive access tools.
- Identification and requirement specification of a number of new access tools.
- A curator tool for controlling and scheduling the collection of web content.
- A definition of the overall architecture for web archiving.
- A definition of the the WARC (web ARChive) file format that is a revision and generalization of the ARC format used by the Internet Archive to store information blocks harvested by web crawlers.
- A definition of the IIPC metadata set.
- A defintion of system interface specifications
- Alternative techniques for deep web acquisition .

The tangible results of the IIPC are few but very important. One of the missions of the IIPC is to foster the development and use of common tools that enable the creation of international

archives. The IIPC itself is not responsible for projects of this kind but it approves them and supports some of them with funds. The projects are either carried out by individual members or by a few members in co-operation. Although not specified the consensus has been reached, that tools developed shall be available as Open Source tools. This is important because they are available for all free of charge and it will encourage continued development and maintenance of the tools by those interested in web archiving.

The tool with the greatest impact is the Heritrix Crawler/Harvester for collecting web material that was developed in co-operation by the IA and the Nordic national libraries. The Heritrix is already heavily used by some of the members and it appears to be the tool of choice for adding new features and capabilities needed by members. Some members are looking for development of extensions that would enable selective harvesting of a national domain according to predefined criteria, so called “smart crawling”. Another project that will be finished later this year is a fusion of a Full Text Indexer/Search Engine, provided through the IA, with a User Interface provided by the Nordic national libraries. This will allow full text indexing, search and retrieval of a web archive similar to what is provided in the real web. The National Library of Australia developed a tool called eXplore for searching/browsing the content of a an archived database, stored as XML and BnF developed a tool called Deep Arc to extract data rom a database. BnF has also developed an Arc files manipulation tool

Future of IIPC

The current IIPC charter ends in July 2006 and it is clear that the consortium can not be continued unchanged. At the steering committee meeting in June 2005 this was discussed and proposals for continuation will be discussed at the next meeting in late October 2005. The issue is unresolved. Many institutions will want to join the consortium and participate in its work and that is a good evolution. The challenge is to keep the work focused and effective because if the members become too many this will not be the case. There are many unsolved problems and hopefully new members can help with this. A notable but very conscious omission of IIPC tasks is long time preservation of web archive. The IIPC members believe that this is best solved as part of the broad scope of long time preservation of digital data. Ultimately when web archiving reaches maturity and becomes fairly common the IIPC may not be needed anymore.

Concluding remarks

Web archiving is still in its infancy and only a small part of the global web is being harvested and preserved. It has proven virtually impossible to enforce a law that would require all who publish a work on the Internet to send a copy to a Legal Deposit library. The only practical solution is to use Internet- or web-harvesting to collect web documents, and accordingly the legal deposit laws in each country must be changed allowing the national libraries to do this. This takes time even in countries where the national library has been pushing for this law.

The pioneer in web archiving is the Internet Archive of San Francisco, and it has collected by far the largest archive of web documents by harvesting all over the globe. Still it contains only a part of what is available. It is very difficult to establish national boundaries for the web and therefore the IIPC is working on defining and developing both the necessary technology components and the procedures and standards that will enable the building of national and even global web archives with the necessary access.

Currently it seems that the national libraries will be permitted to collect what is published on their national domain and store the information in a web archive.

Two different collection (harvesting) methods are in use: Either the aim is to have as broad a scope as possible in order to have a representative collection of web material, or the aim is to have a narrow scope and collect in depth to have as precise as possible idea of what a given

web site contains at the time of harvest. Those methods are mixed in the three most prevalent collection policies currently used:

1. Cross section harvesting aims to provide a snapshot of the web domain, and uses broad harvesting.
2. Selective harvesting aims to harvest a very high percentage of a limited number of chosen web sites, and uses in depth harvesting.
3. Event or thematic based harvesting aims to cover very thoroughly a certain event or theme, and uses a combination of broad and narrow scope harvesting.

Archiving the web is on the borderline between the library and the information technology (IT) profession and the methods above reflect that. It is a library collection but it requires substantial involvement of IT resources. Reflecting this on one side is the Australian Pandora project with traditional librarian values where “quality” web sites are selected and catalogued by librarians, and access is by structured search. On the other side is the Internet Archive using a cross section harvesting of the web material from all over the world, regardless of national legislation, and access is provided by computer generated indexing.

Access (navigation, indexing, searching) to the web archives is not as self evident and the rules for general access differ very much between countries, but in most cases it is secured for research purposes. This must change in the future if the web archives are to be of optimum use.

The web Archives contain multimedia documents and currently it is only possible to index text. For other types of documents like images, indexing is limited to harvester generated metadata plus textual information in the file header. However a lot of research is being done with the purpose to use computer programming to index sound, images and video/movies, and when this becomes practical these parts of the web-Archives can be indexed.

Despite the ongoing work it is important to note that from the point of view of collecting and preserving the Internet our understanding of the medium and its contents leaves much to be desired. Every current tool needs to be improved and new tools must be developed. Hopefully the work of the IIPC will increase awareness and activity in web archiving and coupled with advances in technology and international cooperation, this will in the near future enable the national libraries of the world to collect and preserve the Internet like the printed collections of today. The National Libraries should consider web-archiving and indexing as a tremendous, albeit challenging, opportunity instead of looking at it as a problem or liability.

Acknowledgements
This article is based on published and unpublished IIPC documents and an article by Birgit N. Henriksen available at:
http://www.netarkivet.dk/website/publications/webarkivering-webarchiving.pdf