Web Harvesting and Archiving

João Miranda

Instituto Superior Técnico

Technical University of Lisbon

jmcmi@mega.ist.utl.pt

Keywords: *web harvesting, digital archiving, internet preservation*

Abstract. This paper describes the main topics related to web archiving. From harvesting to preservation, several issues must be taken in concern. A few organizations soon realized the importance of preserving the web for future generations and launched distinct programs like PANDORA and the Internet Archive. An international endeavour in this field led to the creation of the International Internet Preservation Consortium.

Introduction

Internet plays a major role in our current society. A substantial part of intellectual content produced at the present time is born digital. If this material is not collected, it may be lost forever. Some organizations started initiatives for preserving web material. Despite its good will, collecting the web is not as easy as one might think. Its dynamic nature and heterogeneity character put a challenge when aiming at such a task.

The Internet

Internet combines the attributes of many different traditional media. With the growth of the World Wide Web (WWW) in the 1990's the amount of contents has enormously expanded, resulting in large quantity information available worldwide.

The volatile condition of the web was soon acknowledged and the lifespan of published information was sometimes not as long as expected.

A significant part of the information published on the internet has no parallel in print or other supports, being created originally in digital format only. If this information is not archived in any way, it may never become available again in the future.

One may consider three different types of content: contents created for internet only; contents designed for a digital and conventional distribution; contents designed having in mind only conventional distribution, such as paper print, but where digital processes are used in a part of the production process [1].

The web is a unique source of information at the present time. Access to its collections is not merely suitable for researchers, historians and scholars: it stands as a part of present society's culture. Access to literature and information sources is considered essential to education and maintenance of a developed society.

The WWW eases global communication and is an important medium for scientific communication and publishing.

Formerly, exclusively librarians and archivists managed long-term preservation and archiving of print information. At present, this task is aided by computer science experts that manage the development of systems able to support such a mission. In fact, archiving the web is halfway between the library and the information technology lines as it can be seen as a library collection which can not be done without the help of information technology expertise.

Despite having valuable information and resources, such as results of scholarly and scientific research, there is much content lacking quality that does not necessarily deserve to be archived [2].

Digital information has different weaknesses in comparison with traditional media, such as paper. It is more easily changed or corrupted without recognition [3].

Internet's dynamic nature results in continuous evolution of web resources, with its elements being changed or deleted. This may lead to unavailability of precious scholarly, cultural and scientific resources to future generations. This also happens with web sites of major events as, for example, the Sydney Olympic Games, in which a significant part of its web presence disappeared in few time, had not been the efforts of the National Library of Australia and PANDORA archive [4].

As pages and sites often change or disappear without a trace, web archiving initiatives are needed to help preserve the cultural, informational and evidential value of the Internet.

Initiatives for web archiving

Most countries seek to preserve and provide access to cultural and intellectual heritage by collecting and storing it in libraries, archives and museums. As a part of our culture relies nowadays in the Internet, and considering its ephemeral character, information may be lost forever if not collected.

Provided that the web is a worldwide phenomenon, no single initiative can expect to have a total coverage of the web. Therefore, a close cooperation between different archiving initiatives is extremely important.

In 1996 the Internet Archive [5] and the national libraries of Australia and Sweden started initiatives for collecting web material. In 1997 and 1998 the national libraries of Denmark, Finland, Iceland and Norway started similar initiatives. Several other countries such as France, UK, Austria, Japan, China, Greece and USA (Library of Congress) have followed the same steps since the year 2000. However, these projects had different purposes. Whereas in Australia the centre of interest was on a limited number of chosen sites, in Sweden the focus was on the national web. The Internet Archive aims at collecting as much as possible of the world web.

Legal Deposit has to follow the advances of modern publication media and an extension to the web has to be undertaken. Some countries have changed their legal deposit laws, allowing the collection of web published material.

In Denmark [6] the legal deposit law has been revised in 1997 so as to allow the legal deposit of static documents on the internet. Work has been developed and a new revision came into force on July 2005. The State and University Library and the Royal Library are now allowed to automatically gather all Danish websites. These comprise, for example, the .dk top level domain, websites minded on a Danish audience, written in Danish or about Danish notable people. The legal deposit law covers all public available material. The two libraries developed, in collaboration, the strategies and software for the collection, archiving, preservation and access of material, which is held at netarchive.dk.

In 2001 the first International Web Archiving Workshop (IWAW) took place in Germany. Since then, the IWAW is held in association with the European Conferences on Digital Libraries (ECDL). The purpose of these workshops is to bring together researchers and IT developers interested in web archives issues.

PANDORA

The PANDORA (Preserving and Accessing Networked Documentary Resources of Australia) archive is an initiative of the National Library of Australia and was established in 1996 with the development of a proof-of-concept archive and a conceptual framework for a permanent service.

Instead of attempting a whole or substantial domain archiving, and regarding its limited resources, the National Library decided to take a selective approach [8]. Collecting digital publications is a time-consuming and expensive task, so the National Library of Australia decided to focus on publications considered valuable for research in the present and in the future.

After selecting the sites, an agreement is secured with the site owners. Then the material is collected using web harvesters or directly provided by the owners.

In 15 December 2005, PANDORA contained 10.418 archived titles, representing 20.857 archived instances, 28.734.346 files and 1038 GB. It is growing at a rate of 26 GB per month and around 518.932 files per month.

PANDORA is based on an archive management system called PANDAS (PANDORA Digital Archiving System).

Kulturarw

In 1996, the Swedish Royal Library started one of the pioneer web archiving projects – Kulturarw. It used an active collection policy based on the development of a harvesting robot by the Royal Library. At first, public access was not part of the plan, but in 2002 the Royal Library allowed public access to the collection.

Kulturarw uses a harvesting robot running twice a year to collect the Sweden sites. Its archive is around 4,5 TB and is growing at an average rate of 2-3 TB per year.

The Internet Archive

The Internet Archive (IA) was one of the first attempts to preserve the web. It was founded in 1996 with the purpose of offering permanent access for researchers, historians, and scholars to historical collections existing in digital format. Although established as a non-profit organization, it collected pages that had been crawled by Alexa Internet, a commercial company developing web navigation tools. One of the main goals of IA is to prevent internet and other digital materials from disappearing into the past.

In 1999 the IA expanded to include texts, moving images and software in its collections. The IA made its collections available to the public in 2001 using the Wayback Machine, making it easy to retrieve a page from its archived resources [9].

In Wayback Machine users may enter a URL and get a list of the instances of the URL archived through time. Users start navigation by choosing one of those instances and then the system selects the links which are most close in time relatively to that date. In this way, users get a time dimension in navigation which enhances the knowledge and perception about the temporal evolution of a given website. However, Wayback Machine does not give users the full text search over archived content they are used to handle with common search engines. Instead, users must know exactly where they want to go and, given a specific URL, Wayback Machine will display the archived instances of that location. Alternatively, users may enter a domain name and Wayback Machine will display the pages archived under that domain name.

The IA has special collections dedicated to specific subjects. Recently, it created the Hurricane Katrina & Rita Web Archive, a collection of websites documenting the historic devastation and massive relief effort due to Hurricane Katrina. The sites were crawled between September and mid October 2005. This collection contains more than 25 million searchable documents and will be preserved by Internet Archive with access to historians, researchers, scholars and the general public.

Internet Archive is in close cooperation with institutions like the Library of Congress and the Smithsonian Institute.

The IA contains approximately 1 petabyte of data and currently growing at a rate of 20 terabytes per month. If one tried to place the entire contents of the archive onto floppy disks and laid them end to end, it would reach from New York, past Los Angeles, and halfway to Hawaii.

The International Internet Preservation Consortium

The International Internet Preservation Consortium (IIPC) is an international association seeking the development of efforts for the preservation of Internet content.

The IIPC was launched in Paris on July 2003. Its twelve members were the national libraries of France, which is the coordinator of the project, Italy, Finland, Sweden, Norway, Australia, the National and University Library of Iceland, Library and Archives Canada, The British Library, The Library of Congress (USA), and the Internet Archive (USA) [10].

The initial agreement is in effect until July 2006. At that date, new members, other than the founder institutions, are welcome to integrate the group.

IIPC has the mission to promote global exchange and international relations, acquiring, preserving and making accessible Internet information for future generations around the world. It aims at enabling the collection of a rich body of worldwide internet content in a way that it can be archived, secured and accessed over time. It fosters the development and use of common tools, techniques and standards that enable the creation of international archives. It encourages and supports national libraries everywhere to address Internet archiving and preservation.

IIPC has the goal of building a global distributed collection in order to guarantee that the nature of the original internet content is not eternally lost. It also provides international advocacy for initiatives that promote the collection and preservation of the internet content.

IIPC intends to develop a collaborative work to design the implementation of solutions for collecting, preserving and making internet archived content available for the public. The ephemeral nature of the internet and its worldwide structure make it difficult for a single institution to guarantee the collection and preservation of its contents. IIPC members realised that in order to preserve the vast information internet offers it was imperative to create a new way of collaboration between the national institutions responsible for national cultural heritage.

In order to achieve its mission, IIPC set up working groups focused on specific subjects. These dedicated committees (framework, researchers requirements, access tools, metrics and test bed, deep web and content management) strive to accomplish the goals of IIPC [11].

The framework working groups focuses on developing standards and models for web archiving procedures and enabling technical interoperability between the IIPC member libraries. This working group defined at the beginning of the consortium the general high level architecture on which the other groups are based.

The metrics and test bed working groups focuses on defining metrics for collecting and preserving internet content and evaluating the performance of web archiving tools. This assessment also takes in consideration the behaviour of the harvesters when dealing with the usual problems of content formats and technical limitations with content access and retrieval within the websites.

The access tools working group focuses on initiatives and tools that allow present and future access to internet archive material. This includes the selection phase access, content management access, digital archive management access and end user access.

The deep web working groups focuses on creating strategies and tools for the archiving of web content not accessible to web harvesters. It is very important to handle correctly the database-driven web content and indexing the maximum possible of deep web and information stored in databases.

The content management working group focuses on developing a model for content stewardship and tools for managing the collection of internet content. It is important that the members share a common vision of collection coverage.

The researchers requirements working group focuses on discussing with experts on this field the characteristics and properties web archives should present. This groups provides advice on how to index and what to include in web archives.

These groups have been working on this area and have presented papers describing the best practices and procedures to be used in web archiving. They have been promoting the development of tools required to work with web archives. The IIPC is not directly responsible for the projects but may fund and support them. The tools are always

developed having in mind the open source philosophy, so as to be maintained and continuously developed by those who whish to improve it.

The most known tool is the Heritrix Harvester developed by the Internet Archive and the Nordic national libraries. Some of the members have been using Heritrix and find it to be a good choice. Other tools have been developed. The National Library of Australia developed a tool for browsing the content of an archived database, stored in XML format, which is called eXplore. The National Library of France developed a tool for manipulation Arc files and a tool to extract data from a database, called Deep Arc. The consortium has been working on a web archive toolset which comprises Heritrix, an archive format manipulation tool called BAT, an access tool called WERA and a search engine called NutchWAX.

The consortium configuration will change in July 2006. New members are welcome to join, help solve existing problems and perform a positive evolution of IIPC.

Challenges

Web archiving poses different challenges. The fast growth that web has experienced, combined with its dynamic characteristics, make it difficult to decide what to preserve and keep it up-to-date. The web is very large and still growing every day.

Its decentralised organization results in different policies and adoption of standards by the organizations responsible for the sites. Each different site owner chooses its own standards and builds its site at its own way.

Pages and websites change, appear and disappear, often with no trace. Its fluid essence results in links that no longer work, whether pointing to the wrong place or no place at all, and often returning a 404 error that web users so well know. This means that at any one time a large amount of links on the internet will be dead or link to the wrong site. Internet domain names sometimes disappear or change ownership. Life of a link may be conditioned by link-depth, file format and site owner, as personal web pages have often different lifetimes than institutional sites.

In the beginning of the WWW pages were essentially static but the evolution of web-based technologies made the internet more dynamic. Web pages created from dynamic databases are difficult to collect and replicate in repositories. Much database information will be invisible to harvesting robots.

When a dynamic page contains forms, JavaScript or other active elements, the archived site will lose its original functionality. The more simple and static the site is, the more easily it is harvested.

Robots exclusion headers present in robots.txt are a way of instructing automated systems not to crawl the respective sites (Standards for Robot Exclusion - SRE). Compliance with robots.txt is voluntary and some sites do not even have this file. However, if respected, the site may lack a substantial constituting part or may not be harvested at all.

Legal issues and copyright are another problem. Questions related to defamation, content liability or mistaken information data may present a problem when archiving is already performed. It may happen that online contents may be changed or removed, even by court order, and the collection is already fulfilled. The lack of legal deposit mechanisms and aspects concerning data protection address another hurdle.

Web archive collections tend to be sizable. The reason is that web collections usually are set to harvest not only text but all the resources in a web page. Non-text contents such as images, archives and multimedia elements tend to enlarge the size of the final collection. Dynamic pages often also provide a multiplicity of views over the web page and, as a result, a small sized resource may become huge if one does not set limits or if the harvesting is not made under control. Large web archive collections may get difficult to manage and manipulate.

Collecting the web is a time consuming task and harvesting activities must be well coordinated so as to perform a well succeeded web collection. Internet Archive has collected by far the largest web collection by harvesting web documents all over the world. Despite that, it still contains only a part of the content available.

Deep web *vs.* surface web

A significant part of the web is not accessible or is of restricted access (usually a password and a username is required) and is very hard, if not impossible, to collect. Sometimes websites are real islands, isolated, with no link to the outer web. This inaccessible subset of the web is known as "deep web" in contrast with the freely accessible "surface web".

The surface web is easily accessed and harvested by crawlers, as the contents are freely available for every user. It is the case of a normal web page, where all data is attainable with a scanning of the page by a crawler. This does not happen with pages that contain database-driven content, in which one will not be able to guess all of what the source database contains. It also does not happen with pages that are password protected or access restricted.

The deep web is a serious issue to web preservation initiatives, especially the ones based on the harvesting approach.

Approaches

Three main different approaches are taken when considering web archiving:

- deposit, in which web documents or parts of sites are transferred into a repository;
- automatic harvesting, in which crawlers attempt to download parts of the web. The Internet Archive follows this approach;
- selection, negotiation and capture, in which web resources are selected, then their inclusion is negotiated with the site owners and finally it is captured. The National Library of Australia follows this approach.

These approaches do not have to be taken in separate. Combinations of them may be used, in order to achieve a more successful result.

One may consider two different main harvesting methods. The first is to collect in depth in order to have a complete idea of the contents of a website at a given time. The second is to collect broadly in order to have a representative collection of web material. These methods are contemplated in the three main harvesting policies used:

- selective harvesting is used to collect a large percentage of a limited number of chosen sites, using in depth harvesting;

- cross section harvesting is used to take a snapshot of the web domain, using broad harvesting;
- thematic or event based harvesting is issued to collect as much as possible a number of websites related to a theme or an event, using a combination of in depth and broad harvesting.

Selecting what to archive may sometimes be a difficult task. Selection guidelines must be established in order to perform a successful collection. Archiving every single resource present on the internet is, in fact, impractical, if not impossible.

Time dimension may enlarge an archive's value. If a collection is made with successive harvesting sessions through time, one gets a comprehensive record of the resources evolution on the internet. This enables the composition of a relative temporal model of the documents in which dependencies and inconsistencies may be made explicit.

Problems in harvested resources

Harvested resources may present several problems. Sometimes images are missing, resulting in a negative effect on site navigation. Links, sometimes significant, are at times not retrieved resulting in incomplete archives. The absence of multimedia or plug-in-based content related to web pages interferes with site navigation.

Dynamic contents and javascript may represent a threat to functional navigation of harvested resources. This may lead to redirection problems in links. Flash contents also lead to problems in the structure and navigation of harvested resources.

Preservation strategies

Archiving systems are useless if one is unable to successfully access data stored in five, ten or more years. Special challenges, such as the enormous amount of data or the heterogeneity of file formats, arise when concerning web archives.

Digital contents are expensive to maintain over time as they rely on hardware, software, data and standards that are replaced every few years.

Four strategies have been used with respect to long-term preservation: emulation, migration, normalization and metadata [12].

The main sources of data loss in archival repositories are media failure and decay, media or format obsolescence, human and software errors, and external events, such as fires or natural disasters. Every document in the archival repository has a number of manifestations. A manifestation is a way in which a document can be rendered to a user. If there are no more ways in which the document can be rendered to a user, then the document has no more manifestations and is considered to be lost. A manifestation can be presented to a user in different modes of access, for example, playing a sound, rendering a picture or displaying a document.

Web archives face a big challenge when handling file formats. Each site from where the web material is collected has its own schemes of which formats to use. In order to be complete, web archives must accept material in every format. To correctly render the objects an appropriate application is needed. The application is generally chosen by the local setup of the terminal used to access the archive, based on a file extension or a

mime-type. This solution will not stand forever, as formats and standards change over time. In this way, not only the contents need to be archived but also the formats and applications [13].

For an archive to be useful a user must be able to access the documents in the archive. When a viewer in not present, documents may still be accessed if a suitable converter exists. The problem with converters is that sometimes they do not always respect the original layout or structure of the documents and, in result, the quality may be reduced and the look-and-feel will not be the same.

There is no single best solution to preserving digital material. Different approaches have been presented.

PANIC is a prototype digital preservation system based on a combination of preservation metadata capture, software and format registries, and Semantic Web Services. PANIC provides a semi-automatic, distributed Web services approach aiming at a cost-effective solution to the long-term preservation of large scale collections of complex digital objects. The automatic detection of potentially obsolescent digital objects, and execution of the most appropriate preservation service, may prevent the loss of valuable digital assets.

The Enclose-and-Deposit method follows the encapsulation concept, based on a model of cooperating archival systems that mutually deposit archived resources, improving the reliability of preservation through time [14]. In this method, a resource is enclosed together with its preservation metadata source and this assemblage is finally archived as a resource.

Concluding Remarks

Collecting and archiving the web poses a number of challenges that have to be overtaken in order to perform a well-succeeded web collection. The existing systems take different approaches when harvesting the web. Different countries and organizations have perceived the importance of preserving valuable web contents and started efforts in order to do so. We have seen a few examples of projects in this field. Will they succeed the mission of making current web content available to generations to come?

References

1. José Borbinha. "Depósito e Preservação na Biblioteca Nacional Digital". VIII Congresso da BAD. (2004) [On-line] URL: http://sapp.telepac.pt/apbad/congresso8/comm8.pdf
2. Michael Day. "Collecting and preserving the World Wide Web: A Feasibility Study Undertaken for the JISC and Wellcome Trust" (2003). [On-line] URL: http://library.wellcome.ac.uk/assets/WTL039229.pdf
3. Gail M. Hodge. Best Practices for Digital Archiving: An Information Life Cycle Approach. D-lib (2000). [On-line] URL: http://dlib.org/dlib/january00/01hodge.html
4. PANDORA. [On-line] URL: http://pandora.nla.gov.au/
5. Internet Archive [On-line] URL: http://www.archive.org/
6. Niels H. Christensen. "Preserving the bits of the Danish Internet ", 5th International Web Archiving Workshop (IWAW05) (2005). [On-line] URL: http://www.iwaw.net/ 05/papers/iwaw05-christensen.pdf
7. IWAW [On-line] URL: http://bibnum.bnf.fr/ecdl/
8. Paul Koerbin. "The PANDORA Digital Archiving System (PANDAS) and Managing Web Archiving in Australia: a Case Study", 4th International Web Archiving Workshop (IWAW04) (2004). [On-line] URL: http://www.iwaw.net/04/proceedings.php?f=Koerbin
9. Michael Stack, Full Text Search of Web Archive Collections, 5th International Web Archiving Workshop (IWAW05) (2005). [On-line] URL: http://www.iwaw.net/05/papers/iwaw05-stack.pdf
10. Thorsteinn Halgrimsson. "The International Internet Preservation Consortium (IIPC)", World Library and Information Congress: 71st IFLA General Conference and Council (2005). [On-line] URL: http://consorcio.bn.br/cdnl/2005/HTML/Presentation%20Thorsteinn%20Halgrimsson.htm
11. Catherine Lupovici. Web archives long term access and interoperability: the International Internet Preservation consortium activity. World Library and Information Congress: 71st IFLA General Conference and Council (2005). [On-line] URL: http://www.ifla.org/IV/ifla71/papers/194e-Lupovici.pdf
12. Jane Hunter. "PANIC – An Integarated Approach to the Preservation of Composite Digital Objects using Semantic Web Services", 5th International Web Archiving Workshop (IWAW05) (2005). [On-line] URL: http://www.iwaw.net/05/papers/iwaw05-hunter.pdf
13. Niels H. Christensen. "Towards Format Repositories for Web Archives", 4th International Web Archiving Workshop (IWAW04) (2004). [On-line] URL: http://www. iwaw.net/04/proceedings.php?f=Christensen
14. Koichi Tabata, Takeshi Okada, Mitsuharu Nagamori, Tetsuo Sakaguchi, and Shigeo Sugimoto, A Collaboration Model between Archival Systems to Enhance the Reliability of Preservation by an Enclose-and-Deposit Method, 5th International Web Archiving Workshop (IWAW05) (2005). [On-line] URL: http://www.iwaw.net/05/papers/iwaw05-tabata.pdf