IAEA Nuclear Energy Series

No. NG-T-6.6

Web Harvesting for Nuclear Knowledge Preservation

IAEA
International Atomic Energy Agency

IAEA NUCLEAR ENERGY SERIES PUBLICATIONS

STRUCTURE OF THE IAEA NUCLEAR ENERGY SERIES

Under the terms of Article III.A. and VIII.C. of its Statute, the IAEA is authorized to foster the exchange of scientific and technical information on the peaceful uses of atomic energy. The publications in the **IAEA Nuclear Energy Series** provide information in the areas of nuclear power, nuclear fuel cycle, radioactive waste management and decommissioning, and on general issues that are relevant to all of the above mentioned areas. The structure of the IAEA Nuclear Energy Series comprises three levels: **1 — Basic Principles and Objectives; 2 — Guides; and 3 — Reports.**

The **Nuclear Energy Basic Principles** publication describes the rationale and vision for the peaceful uses of nuclear energy.

Nuclear Energy Series Objectives publications explain the expectations to be met in various areas at different stages of implementation.

Nuclear Energy Series Guides provide high level guidance on how to achieve the objectives related to the various topics and areas involving the peaceful uses of nuclear energy.

Nuclear Energy Series Technical Reports provide additional, more detailed, information on activities related to the various areas dealt with in the IAEA Nuclear Energy Series.

The IAEA Nuclear Energy Series publications are coded as follows: **NG** — general; **NP** — nuclear power; **NF** — nuclear fuel; **NW** — radioactive waste management and decommissioning. In addition, the publications are available in English on the IAEA's Internet site:

http://www.iaea.org/Publications/index.html

For further information, please contact the IAEA at P.O. Box 100, Wagramer Strasse 5, 1400 Vienna, Austria.

All users of the IAEA Nuclear Energy Series publications are invited to inform the IAEA of experience in their use for the purpose of ensuring that they continue to meet user needs. Information may be provided via the IAEA Internet site, by post, at the address given above, or by email to Official.Mail@iaea.org.

WEB HARVESTING FOR NUCLEAR KNOWLEDGE PRESERVATION

The following States are Members of the International Atomic Energy Agency:

AFGHANISTAN
ALBANIA
ALGERIA
ANGOLA
ARGENTINA
ARMENIA
AUSTRALIA
AUSTRIA
AZERBAIJAN
BANGLADESH
BELARUS
BELGIUM
BELIZE
BENIN
BOLIVIA
BOSNIA AND HERZEGOVINA
BOTSWANA
BRAZIL
BULGARIA
BURKINA FASO
CAMEROON
CANADA
CENTRAL AFRICAN REPUBLIC
CHAD
CHILE
CHINA
COLOMBIA
COSTA RICA
CÔTE D'IVOIRE
CROATIA
CUBA
CYPRUS
CZECH REPUBLIC
DEMOCRATIC REPUBLIC OF THE CONGO
DENMARK
DOMINICAN REPUBLIC
ECUADOR
EGYPT
EL SALVADOR
ERITREA
ESTONIA
ETHIOPIA
FINLAND
FRANCE
GABON
GEORGIA
GERMANY
GHANA
GREECE
GUATEMALA
HAITI
HOLY SEE
HONDURAS
HUNGARY
ICELAND
INDIA
INDONESIA
IRAN, ISLAMIC REPUBLIC OF
IRAQ
IRELAND
ISRAEL
ITALY
JAMAICA
JAPAN
JORDAN
KAZAKHSTAN
KENYA
KOREA, REPUBLIC OF
KUWAIT
KYRGYZSTAN
LATVIA
LEBANON
LIBERIA
LIBYAN ARAB JAMAHIRIYA
LIECHTENSTEIN
LITHUANIA
LUXEMBOURG
MADAGASCAR
MALAWI
MALAYSIA
MALI
MALTA
MARSHALL ISLANDS
MAURITANIA
MAURITIUS
MEXICO
MONACO
MONGOLIA
MONTENEGRO
MOROCCO
MOZAMBIQUE
MYANMAR
NAMIBIA
NETHERLANDS
NEW ZEALAND
NICARAGUA
NIGER
NIGERIA
NORWAY
PAKISTAN
PALAU
PANAMA
PARAGUAY
PERU
PHILIPPINES
POLAND
PORTUGAL
QATAR
REPUBLIC OF MOLDOVA
ROMANIA
RUSSIAN FEDERATION
SAUDI ARABIA
SENEGAL
SERBIA
SEYCHELLES
SIERRA LEONE
SINGAPORE
SLOVAKIA
SLOVENIA
SOUTH AFRICA
SPAIN
SRI LANKA
SUDAN
SWEDEN
SWITZERLAND
SYRIAN ARAB REPUBLIC
TAJIKISTAN
THAILAND
THE FORMER YUGOSLAV REPUBLIC OF MACEDONIA
TUNISIA
TURKEY
UGANDA
UKRAINE
UNITED ARAB EMIRATES
UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND
UNITED REPUBLIC OF TANZANIA
UNITED STATES OF AMERICA
URUGUAY
UZBEKISTAN
VENEZUELA
VIETNAM
YEMEN
ZAMBIA
ZIMBABWE

The Agency's Statute was approved on 23 October 1956 by the Conference on the Statute of the IAEA held at United Nations Headquarters, New York; it entered into force on 29 July 1957. The Headquarters of the Agency are situated in Vienna. Its principal objective is "to accelerate and enlarge the contribution of atomic energy to peace, health and prosperity throughout the world".

NUCLEAR ENERGY SERIES No. NG-T-6.6

WEB HARVESTING FOR NUCLEAR KNOWLEDGE PRESERVATION

INTERNATIONAL ATOMIC ENERGY AGENCY
VIENNA 2008

COPYRIGHT NOTICE

Sales and Promotion Unit, Publishing Section
International Atomic Energy Agency
Wagramer Strasse 5
P.O. Box 100
A–1400 Vienna
Austria
Fax: +43 1 2600 29302
Tel: +43 1 2600 22417
http://www.iaea.org/books

Printed by the IAEA in Austria
January 2008
STI/PUB/1314

ISBN 978–92–0–111207–1
ISSN 1995–7807

The originating Section of this publication in the IAEA was:

Nuclear Knowledge Management Unit
International Atomic Energy Agency
Wagramer Strasse 5
P.O. Box 100
A-1400 Vienna, Austria

FOREWORD

The IAEA has taken on the obligation to organize the continued availability of literature in the field of nuclear science and technology for peaceful applications. In the International Nuclear Information System (INIS), millions of scientific citations and the full texts of hundreds of thousands of pieces of non-conventional literature (NCL) have been collected worldwide and have been assembled into the INIS database of citations and the associated collection of NCL full texts.

The next step in the IAEA's endeavour to secure the continued access to scientific and technical literature in the nuclear field which is now available on the Internet to its staff and to Member States. The IAEA is currently conducting pilot projects under the heading NuArch that could eventually become the seed of a comprehensive archive of electronic documents in the nuclear field.

A pilot project was started in the IAEA for the period 2004–2005 and continues for the period 2006–2007. This publication provides information and examples based upon experience in a variety of Member States. It provides general information that present technical aspects of web harvesting in the context of knowledge preservation in the nuclear field, contemporary activities in the domain of web harvesting, document archiving and Internet access technology in order to obtain a contemporary technology overview. Several aspects of possible web harvesting methodologies are presented in some detail in this document which can also serve as a basis to establish future co-operation.

Appreciation is expressed to all the participants who contributed to this publication. Particular thanks are due to A. Mueller Pathle (Switzerland), W. Mandl (Germany) and A. Badulescu (Romania) for their assistance in the compilation of this report. The IAEA officer responsible for this report was Y. Yanev of the Department of Nuclear Energy.

EDITORIAL NOTE

The use of particular designations of countries or territories does not imply any judgement by the publisher, the IAEA, as to the legal status of such countries or territories, of their authorities and institutions or of the delimitation of their boundaries.

The mention of names of specific companies or products (whether or not indicated as registered) does not imply any intention to infringe proprietary rights, nor should it be construed as an endorsement or recommendation on the part of the IAEA.

CONTENTS

1. INTRODUCTION AND BACKGROUND 1

1.1. The role of the Internet for scientific and technical publishing 1
1.2. The volatility of Internet based information resources 1
1.3. The role of traditional libraries in knowledge preservation 2
1.4. The need for Internet archives 3
1.5. The role of the IAEA in electronic document preservation 3

2. METHODS FOR DISTRIBUTED ACCESS FACILITATION 4

2.1. Search engines 4
2.1.1. Components of web search engines 4
2.1.2. Potential use of the document cache for archiving 6
2.2. Directories of web resources 6
2.2.1. Internet directories of nuclear relevance 6
2.2.2. Web links as part of bibliographic references 7
2.3. Permanent web addresses 7
2.3.1. DOI and open URLs 8
2.3.2. DOI and persistent URLs (PURLs) 9
2.4. Metadata initiatives 10
2.4.1. Dublin core metadata initiative (DCMI) 10
2.4.2. Open archives initiative 10

3. OVERVIEW OF ONGOING WEB ARCHIVING INITIATIVES 12

3.1. The “Internet archive” 12
3.2. AOLA — Austrian on-line archive 13
3.3. Pandora — Australia’s web archive 14
3.4. Kulturarw3 — National Library of Sweden 15
3.5. Project NEDLIB 15
3.6. LOCKSS permanent publishing on the web 16
3.7. EU MINERVA 16

4. WEB HARVESTING METHODOLOGIES 18

4.1. Identifying contents 18
4.1.1. Exploiting link lists and Internet directories 18
4.1.2. Analyzing document contents 19
4.1.3. Restricting document types 20
4.1.4. Selecting document formats 21
4.1.5. Language considerations 22
4.1.6. Document size as a quality indicator 22
4.1.7. Spin offs from bibliographic reference banks 22
4.1.8. Resource manifests 23
4.2. Data formats for archiving 23
4.2.1. Preserving native formats 23
4.2.2. Harvesting web content into an MHTML archive 24
4.2.3. Printing to PDF 24
4.3. Archive architecture 25
4.3.1. The resource-locator module 26
4.3.2. The harvester 27

4.3.3. The data storage and access module with version control 27
4.3.4. The long term storage module 28
4.3.5. The search and resource navigation tool 28
4.3.6. Platform independent document reader architecture 29

5. SUMMARY AND RECOMMENDATIONS 30

5.1. Project management 30
5.1.1. Prospective users' expectations and needs 30
5.1.2. Sustainable management of a long term archive 30
5.1.3. Progressive complexity phase-in 31
5.1.4. Choice of data formats and software components 32
5.1.5. Software integration and operational support 32
5.2. Technical meeting on web harvesting 34
5.3. Formal requirements gathering 34
5.4. Organizational form 34
5.5. Persistence of the archive 34

APPENDIX: RELATED RESOURCES ON THE INTERNET 35

DEFINITIONS 37

CONTRIBUTORS TO DRAFTING AND REVIEW 41

1. INTRODUCTION AND BACKGROUND

1.1. The role of the Internet for scientific and technical publishing

The global presence and availability of the Internet has, in the past 10–15 years, an exciting and profound impact on how scientific and technical information is exchanged between peers. Publishing a report on the web is technically simple and, compared to a traditional publication in a scientific journal, cost-effective and fast. Laboratory report series and annual reports that document outstanding achievements and the continuous progress that has taken place in research institutes, are today made available electronically in most cases. Electronic publications reach such a large number and diversity of audience that the circulation of printed literature is dwarfed by comparison. Many commercial publishers of scientific literature feel nowadays obliged to go along with this trend, by publishing peer reviewed journals on the web, or by making traditional, printed documents, either entirely or in some restricted form, available via the Internet.

Information seekers have long adapted to this changed situation, and more often than not, the first initiative to find information on a given subject will be searching the Internet, rather than consulting a library. Full-text search capabilities, geographic independence and instant access are compelling arguments in favour of Internet-based document distribution.

Last, but by no means least, the affordable availability of sufficient Internet bandwidth in developing countries is giving researchers and engineers in less affluent countries, for the first time, a chance to be fully aware of contemporary developments and to participate actively and constructively in mankind's global pool of scientific and technical knowledge.

1.2. The volatility of Internet based information resources

The Internet, as a system, has no memory; specific archiving services have to be created to provide it. Printed literature is usually distributed in many hard copies to libraries and to private collections. The continued availability of the knowledge content of commercially produced and distributed literature is therefore ensured, even in the case when the commercial producer goes out of business, or if an individual library is closed down. Information on the Internet, on the other hand, is usually distributed in soft copy only. A research paper that is available almost instantly worldwide from a single web server today will be purged from the world's intellectual heritage in the moment when it becomes unavailable from that single source. The reasons for the discontinued availability of an information resource can be diverse, ranging from budget restraints over politically motivated restriction of information to the loss of interest in a particular knowledge domain at the host of the information resource.

Whatever the reasons for the discontinued availability of Internet based resources are, the Internet is a volatile medium and the eventual loss of access to online resources is the rule rather than the exception. A recent example of this information decay process that has reached some public attention is the discontinuation of the "Library without Walls" at Los Alamos National Laboratory[1]. A large amount of no longer classified materials had been made openly available on the laboratory's web site in an attempt to demonstrate openness and transparency.

When Los Alamos National Laboratory decided in 2002 to discontinue the open access to the materials on its "Library without Walls" web site, this pool of information would have been

[1] The US Department of Energy web site of the "Library without Walls" still exists at http://library.lanl.gov/lww/, but it no longer gives public access to the library's contents.

lost to the public, if it had not been for the initiative of two individuals[2] who had harvested the contents of the library in time and who continue to give access to most of the reports on their own web site[3] (Fig. 1).

Fig. 1. Screenshot of the Federation of American Scientists (FAS).

1.3. The role of traditional libraries in knowledge preservation

Libraries have played a historical role in mankind's quest for knowledge and intellectual development. Libraries of clay tablets were already established in ancient Mesopotamia, 5 000 years ago. Throughout history, the two main functions of traditional libraries were:

(1) to make rare or expensive materials available and easily accessible to scholars in a single location, and
(2) to archive and preserve printed hard copies of texts over long times.

The first of the two main library tasks, to concentrate materials on a given subject in a single, well accessible location was pivotal for the advancement of science. This is the main function of most university libraries, today. It can, however, also be very dangerous to keep single copies of historic documents in one location only. The most famous library of ancient times

[2] Almost all of the withdrawn reports were acquired and preserved in the public domain by Gregory Walker and Carey Sublette.
[3] At the time of writing, Los Alamos reports and publications are available at http://www.fas.org/sgp/othergov/doe/lanl/index.html

was the Alexandrian Library in Egypt. It housed more than 4 000 000 scrolls of papyrus. When the Library was destroyed, much of ancient history, literature and learning were lost forever.

The systematic archival and preservation of documents has become an outstanding task of libraries since the advent of book printing. Copies of scientific literature are now available in sufficient numbers to supply many libraries and therefore, the demise of a single library, would no longer threaten the survival of historic documents, on a global scale.

1.4. The need for Internet archives

The modern practice of Internet publishing has, oddly enough, had a rather negative impact on medium and long term knowledge preservation. Often single copies of scientific and technical reports are kept on dedicated web servers. This makes modern scientific literature as vulnerable to the risk of total loss, as were the scrolls of papyrus in the Alexandrian Library.

A new aspect that became available with electronic publishing is, however, that the creation of indistinguishable copies of original documents is now technically nearly trivial and inexpensive. In order to preserve the knowledge of our generation to our successors, the role of librarianship needs to be extended to documents that are available on the Internet today. The acquisition of online documents can be automated in most cases by using a so called web harvester. The subsequent steps are then again traditional library work, even if the technology to be employed is somewhat different: secure storage, document classification and access facilitation.

The need for Internet archives has been recognised by a number of public and commercial bodies in recent years, and first initiatives to preserve those volatile information and knowledge resources on the Internet are taking shape. The individual reasons for preserving today's Internet resources are varied and range from commercial interests, over the legal obligations of national libraries to the idealistic views of the open source movement.

1.5. The role of the IAEA in electronic document preservation

In the field of nuclear science and technology for peaceful applications, the IAEA has taken on the obligation to organize the continued availability of literature in this field. Via the International Nuclear Information System, INIS[4], millions of scientific citations and the full texts of hundreds of thousands of pieces of Non-Conventional Literature (NCL), have been collected worldwide and have been assembled into the INIS database of citations and the associated collection of NCL full texts.

The next step in the IAEA's endeavour to secure the continued access to scientific and technical literature in the nuclear field that is now available on the Internet, to its staff and to its Member States, is presently taking shape. The IAEA is presently conducting pilot projects under the heading FAWNI (Freely Accessible Web-based Nuclear Information) that could eventually become the seed of a comprehensive archive of electronic documents in the nuclear field.

The present report describes the technical aspects of web harvesting in the context of knowledge preservation in the nuclear field.

[4] http://www.iaea.org/inis/

2. METHODS FOR DISTRIBUTED ACCESS FACILITATION

In the early years after the creation of the World Wide Web, the pool of resources was small enough, so that manually maintained directories were sufficient to give the initial community of users a good overview of what was available on the web[5]. However, given the rapid growth of the reservoir of resources on the World Wide Web, automated means had to be introduced, rather soon.

Methodologies for facilitating access to the contents of the World Wide Web, without making the attempt to create a web archive, are only of peripheral interest in the context of the present report about web harvesting. The following aspects are therefore touched only briefly in as far as they are relevant for the location of resources – a prerequisite for the subsequent harvesting and storing.

2.1. Search engines

General purpose search engines are coming to our mind first, when means of locating resources on the Internet are discussed.

2.1.1. Components of web search engines

A general purpose Internet search engine consists of three distinct functional components, the web crawler, the indexer and the retrieval engine. The crawler is a computer application that starts by downloading web pages from predetermined web locations, the so-called seed URLs. It will then analyse the retrieved web contents for links that it can in turn use to retrieve the next group of web pages, and so forth.

The first crawlers could only follow explicitly coded links in well formed HTML pages, and for more primitive crawlers that is still the case, even now. Modern crawlers are required to extract links from resources other than HTML pages, like for example from resources that are coded in PDF or in Macromedia Flash format. Other hurdles that are to be overcome are the widespread use of Java-Script and similar "active" components in web pages, as well as the correct handling of cookies in connection with dynamically created web pages.

By following clickable links, crawlers can in principle only reach those parts of the web that are known as the "shallow web". Beyond the so called shallow web there exist vast information resources that are only accessible by filling in online forms or by submitting credentials such as username and password. These hidden resources are called the deep web. Today, only few web crawlers, such as the Convera spider[6], are able to penetrate into the deep web. The deep web is estimated to contain up to 100 times more content than the shallow web.

The indexer extracts from the crawled web contents lists of keywords and other metadata that are needed to enable the functionality of the retrieval engine. The comprehensiveness of the index, that means the fraction of the entire (shallow) web that has been indexed, is one of the main quality criteria of a web search engine. In earlier days, when no individual search engine covered more than a few percent of the entire web, so-called meta search engines were commonly used. The function of meta search engines is to submit a given query to many different search engines simultaneously and to collect and arrange the results from all search engines in combined results lists. Deploying up to 30 different search engines simultaneously

[5] Tim Berners-Lee, Weaving the web, Orion Business Books, UK, (1999).
[6] http://www.convera.com/

was quite common with this technology. Nowadays the indexed fraction of the shallow web is high enough with the leading search engines, that the use of meta search engines has fallen out of fashion.

The retrieval engine, finally, is the component of any search engine that impacts the perceived quality and reputation of that search engine most. It determines which resources, out of thousands or even millions of web pages that match a given query string, are listed at the beginning of the search results list. Retrieval engines used to be rather primitive, using only simple keyword count statistics for their results ranking. When Google[7] introduced its new, revolutionary algorithm for ranking search results, it made the Google search engine the market leader within a short time.

Recent developments to improve the search precision include the analysis of the semantic contents of web documents. Convera's search engine Excalibur is one of the first services to attempt the automatic classification of contents based on a semantic analysis (Fig. 2).

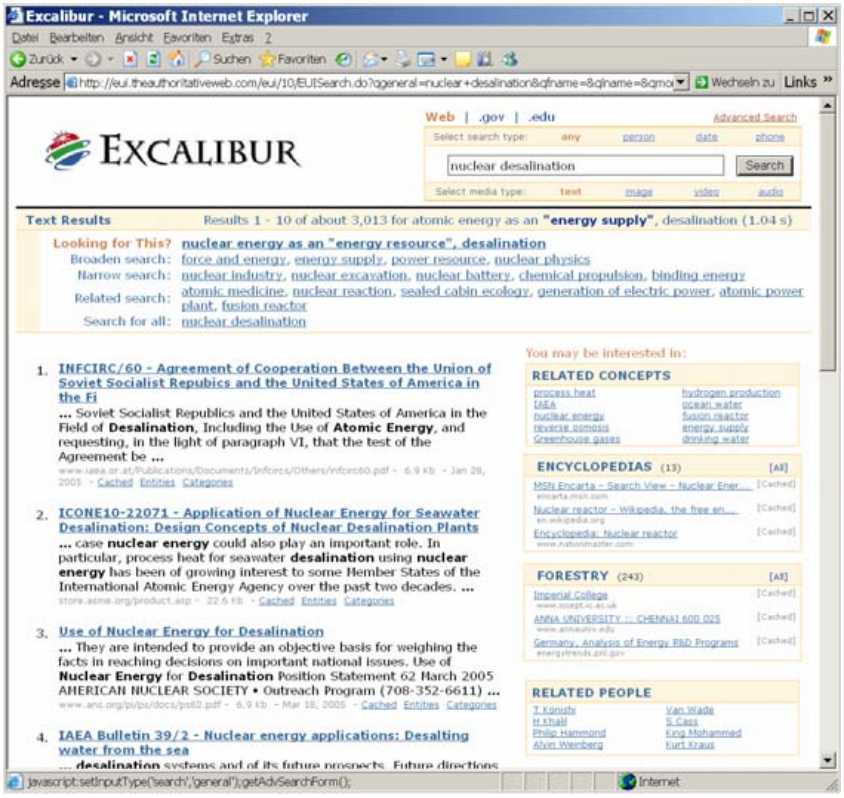


Fig. 2. Excalibur Internet Search Engine.

2.1.1.1. Deep web vs. shallow web

The concept of the deep web is becoming more complex as search engines such as Google have found ways to integrate deep web content into their centralized search function. This includes:

– *The content of databases accessible on the web.* Databases contain information stored in tables created by such programs as Access, Oracle, SQL Server and DB2. Information stored in databases is accessible only by query. This is distinct from static, fixed web

[7] The domain google.com was registered on September 15, 1997, and the company was incorporated as Google Inc. on September 7, 1998.

pages, which are documents that can be accessed directly. A significant amount of valuable information on the web is generated from databases. In fact, it has been estimated that content on the deep web may be 500 times larger than the fixed web.

- *Non-textual files* such as multimedia files, graphical files, software, and documents in formats such as Portable Document Format (PDF).

However, even a search engine as innovative as Google provides access to only a very small part of the deep web.

2.1.2. Potential use of the document cache for archiving

Commercial Internet search engines facilitate the access to the contemporary contents of the web. Although some search engines make the latest cached version of web pages available, too, they lack both the intention as well as the technical means to build a comprehensive archive of the Internet. There is, however, at least the one known case where Alexa Internet[8], the Internet search engine that is owned by Amazon Inc.[9], donates its store of cached documents monthly to a non-profit organization[10] whose declared aim is to build up a comprehensive Internet archive.

In the context of building an Internet archive of the nuclear related resources on the web, one could try to negotiate with operators of existing search engines in order to obtain access to readily available by-products of search engine operations, such as their cached document store. On closer inspection, it is, however, questionable if one should accept dependence on search engine operators who are themselves subjected to commercial necessities and in some cases to political influences[11], too.

2.2. Directories of web resources

A web directory[12] is a directory on the World Wide Web that specializes in linking to other web sites and categorizing those links. Web directories often allow site owners to submit their site for inclusion and human editors review the submissions for their fitness for inclusion.

Web directories are, like search engines, means to facilitate access to web based resources, without archiving them. Apart from the large, multi-disciplinary directories, like the open directory project, a number of smaller directories that focus on particular subjects are available on the web. Such subject oriented directories range from simple, unstructured "link-lists" that are part of many web sites to large, structured and well maintained resources.

2.2.1. Internet directories of nuclear relevance

In the nuclear field, a number of dedicated web directories are available. One of the most used resources is the IAEA's Internet directory of nuclear resources[13]. IAEA's Internet directory lists thousands of nuclear relevant web sites by subject, by acronym and by country. It is therefore an excellent starting point in the search for web sites that should be included in a web archive of nuclear information resources on the Internet.

---

[8] http://www.alexa.com/

[9] http://www.amazon.com/

[10] http://www.archive.org/

[11] BBC News reported on January 26, 2006:"The company is setting up a new site-Google.cn-which it will censor itself to satisfy the authorities in Beijing." (http://news.bbc.co.uk/1/hi/technology/4645596.stm).

[12] Examples of web directories are Yahoo! Directory (http://dir.yahoo.com/), LookSmart (http://search.looksmart.com/), and the Open Directory Project (http://dmoz.org/).

[13] http://www.iaea.org/inis/ws/index.html

In particular, the IAEA's Internet directory maintains a list of other gateways or lists of nuclear related links on the WWW[14].

2.2.2. Web links as part of bibliographic references

With the increased, and still growing, importance of the Internet as an immediate source of full texts, web links have been introduced into bibliographic databases like the INIS Database[15] or the ETDE. While this practice offers immediate conveniences to users who can now access documents "at the click of a button", it raises some concerns:

(a) In the cases of INIS and ETDE, the inclusion of a web link is sometimes seen as a substitute for collecting the full text of non-conventional literature for inclusion into the collection of NCL full texts. This undermines the databases' intention to make NCL permanently available to its members, because links are only delivering the documents as long as the hosting organizations are interested (or able) to provide that access.
(b) The structure of the Internet is subject to constant change. Web servers are replaced or reorganized and web addresses are changed accordingly. To track changed links and to constantly update the corresponding entries in the bibliographic databases is often beyond the interests and the means of database operators.

Attempts to overcome these problems include the use of allegedly permanent web addresses, the so called PURLs.

2.3. Permanent web addresses

A PURL[16] is a persistent uniform resource locator. Functionally, a PURL is a URL. However, instead of pointing directly to the location of an Internet resource, a PURL points to an intermediate resolution service. The PURL resolution service associates the PURL with the actual URL and returns that URL to the client. The client can then complete the URL transaction in the normal fashion. In web parlance, this is a standard HTTP redirect.

PURLs are intended to increase the probability of correct resolution and thereby to reduce the burden and expense of catalogue maintenance. To what extend this goal can be reached is, however, unclear at the time of writing. Two requisites to the functioning of PURLs are:

- The administrators of Internet resources (webmasters) must inform the intermediate resolution service about any changes to URLs in their domain of responsibility.
- The intermediate resolution service must be maintained at all times.

The general experience with Internet based services teaches us that neither of the above mentioned conditions is likely to be upheld over long periods of time. Factually, another failure prone step has been introduced into the mechanism of web based resource access. The official PURL homepage, for example, displays a number of error messages at the time of writing (Fig. 3).

[14]http://www.iaea.org/inis/ws/subjects/resources.html
[15]http://en.wikipedia.org/wiki/INIS
[16]http://www.purl.org/

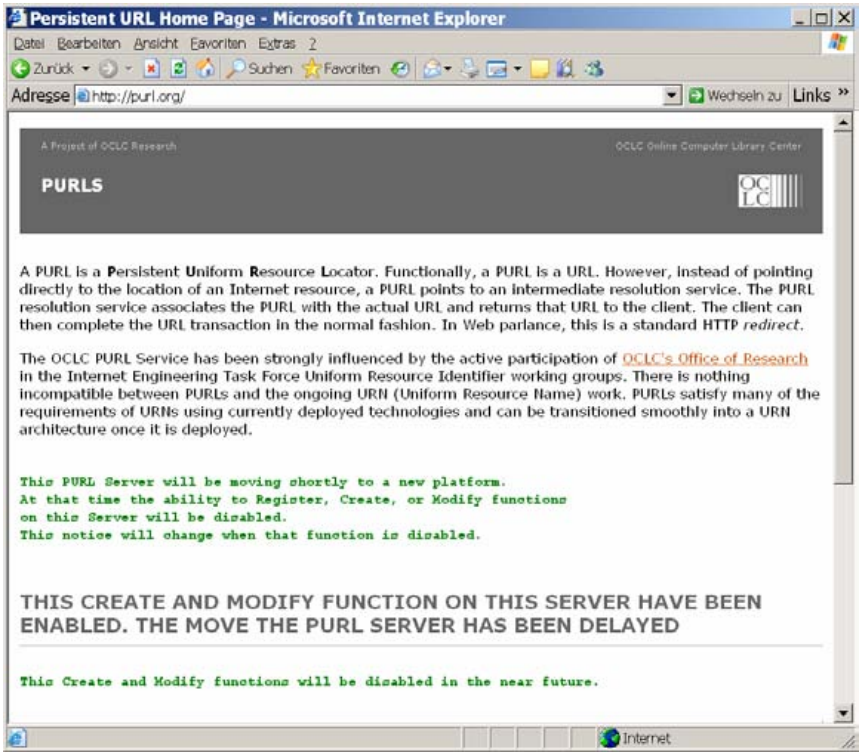


Fig. 3. The Persistent Uniform Resource Locator (PURL) homepage.

Technically, the dependence on a third party for a redirect service is quite unnecessary and can easily be avoided. Each webmaster can ensure the continued availability to resources at changed URLs by simply maintaining a URL redirect list on the local web server.

2.3.1. DOI and open URLs

"When you don't have decent metadata, it's hard to provide decent services. That's why I am an enormous fan of unique identifiers for objects, and systems that allow you to obtain well-structured metadata by using those identifiers. For me the big deal of the DOI/cross ref framework is not necessarily the links they provide, because that might be done in other ways. The crucial importance of that work is in the mere fact that objects are being identified, and that identifiers can lead to metadata about objects. That changes the whole game"[17].

Open URL is a NISO standard syntax for transporting information (metadata and identifiers) about one or multiple resources within URLs. Open URL provides a syntax for encoding metadata (but not a source of it), restricted to the world of URLs (unlike DOI's wider application). This interface can be used to tie together otherwise disparate services such as centralized resolution systems and local knowledge of available resources.

The digital object identifier (DOI) is a system for resolution of identifiers to global services. Open URL is syntax allowing the contextualization of requests to those services to local requirements. Open URL can be used together with DOIs to provide a richer user experience

[17]Herbert Van de Sompel, Creator of OpenURL/SFX, in an interview with Dennis Brunning, The Charleston Advisor, Volume 4, Number 4, April 2003.

that incorporates both the global and the local requirements of the user. A key issue in the open URL world is the transformation of a generic link, say to a publisher's online copy of a journal, into an open URL pointing to the right server for the given user, which must also carry the id and metadata needed to create the contextually appropriate extended service links as described above. In the current deployment this is only being done by the resource pointed at by the URL that the user initially encounters. So in the example of a link to the publishers' copy of a journal, the publisher must: 1) agree to redirect that http request to the user's local open URL-aware server, when appropriate, 2) must add information to the link as needed for the local server to do its job, and 3) must know the location of the local server.

The logically centralized resolution service maintained by the content producers for DOIs has no way to resolve a DOI to a locally held copy of the identified entity. So the synergy between DOI and open URL is clear: open URL needs a source of identifiers and authoritative metadata; DOI provides a single point in the network for the creation and subsequent redirection of open URLs, which is more manageable than asking every content provider to enable this facility. Solving the appropriate copy problem is a significant accomplishment in and of itself, but there are many opportunities for productive collaboration beyond that.

DOIs can be used within the open URL syntax to query local services about availability of resources at a local level, e.g. the following could be used to see if a local copy of a resource were available[18]: The local service could have a list of DOIs that it has a local service for and offer that alongside the global information services obtained by resolving the DOI through the global handle system. Open URL also allows more complex constructs that those illustrated above.

In order to allow for the delivery of context-sensitive services information, recipients of an open URL must implement a technique to determine the difference between a user who has access to a service component that can deliver context-sensitive services and a user that does not. The mechanism used to determine a user's membership of a particular group could be cookies, digital certificates, part of a user's stored profile in an information service, an IP address range, or something else. This user recognition is not a part of the open URL syntax and is separate to open URL. Several library service vendors provide such functionality. If the user is a bona fide member of a group, the local resolution service will be available to that user.

2.3.2. DOI and persistent URLs (PURLs)

PURLs are all HTTP and inherit both the strength and weakness of that approach. PURLs provide one level of indirection, just like a single value DOI handle, but all contained within a single server and that single server is permanently attached to a specific domain name. PURL servers don't know about each other. The redirection is functionally equivalent to the way DOI uses a handle proxy, dx.doi.org, which re-interprets DOI handle queries into HTTP. PURL is equivalent to a local DOI which never goes beyond the proxy server approach and never makes use of the multiple resolutions and data types, metadata approach, and enforced common policy. The DOI system also provides a centrally managed redirection service rather than local purl server management.

[18] http://resolver.local.org/resolutionservice?id=doi:10.1045/1.

DOI sits on top of a system explicitly designed to name digital objects on networks. This system, the handle system, can provide the web-centric functions of a PURL through the use of a proxy server that returns a PURL-like single redirection. But underneath that is a much more extensive set of functionality that can be used as needed now or in the future.

2.4. Metadata initiatives

Most Internet search engines rely on keywords that are contained in the documents themselves for identifying the subject scopes of online documents. In contrast to the detailed data records that are maintained in most bibliographic databases, the Internet is an unstructured data and document reservoir.

In order to improve this situation, most of the document formats that are presently used on the Internet (HTML, PDF, MS-Word etc.) offer possibilities to store additional information about the documents in so called metadata fields. This native meta-information is, however, propriety for each document type, and therefore it is not suitable for providing a metadata standard for the Internet.

Other, document independent initiatives try to define a set of metadata that can be used in a resource independent manner.

2.4.1. Dublin core metadata initiative (DCMI)

"The Dublin core metadata initiative is an open forum engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models."[19]

Unfortunately, the requirement to support a broad range of purposes necessitates that the individual items of metadata are either ambiguous (and therefore applicable in many different ways) or that many different types of metadata are available. In practice, both conditions apply. The element "creator" of the Dublin core metadata element set, for example, applies equally to persons, organizations or a service. Additionally, every implementer of a Dublin core metadata scheme is at liberty to extend the element set with suitable extra elements. This causes globally the availability of very many different elements with partially overlapping or even conflicting definitions. De facto, Dublin core metadata are easily understood by a human observer, but they are suitable for automated processing only if the metadata of a particular resource and the corresponding automated process are exactly coordinated. This makes Dublin core metadata in their practical implementation very similar to any set of propriety metadata.

2.4.2. Open archives initiative

"The open archives initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. The open archives initiative has its roots in an effort to enhance access to e-print archives as a means of increasing the availability of scholarly communication."[20]

The basic aim of the Initiative is not to create new archives but to provide a standard description of existing archives. Following the proposal of the Initiative, archives have to provide a standard mechanism by which metadata about the archive can be accessed (harvested). Whether or not any of the contents of the archive are made available as well, und under which conditions that would be, is outside the scope of the Initiative.

---

[19] http://dublincore.org/
[20] http://www.openarchives.org/

2.4.2.1. Using the open archives metadata manifest for full text harvesting

In the context of harvesting web contents for an archive of nuclear related materials, it should be investigated in how far the open archive interface of participating organizations could be used for harvesting locations and metadata about documents that are readily available for downloading. This might be providing an interesting and rewarding alternative to crawling web links, in some cases. The metadata, in particular, are not only potentially useful for the nuclear archive, they could also be used to filter which of the contents should be considered for downloading.

3. OVERVIEW OF ONGOING WEB ARCHIVING INITIATIVES

The basic idea of downloading parts of the World Wide Web to a local archive, in order to preserve it, is not new. Probably the best known initiative of this kind is the "Internet archive"[21].

Other web archiving initiatives are often related to the activities and the mandate of national libraries.

3.1. The "Internet archive"

The Internet archive (Fig. 4) is building a digital library of Internet sites and other cultural artefacts in digital form. Like a paper library, they provide free access to researchers, historians, scholars, and the general public.

Founded in 1996 and located in the Presidio of San Francisco, the archive has been receiving data donations from Alexa Internet and others. In late 1999, the organization started to grow to include more well-rounded collections. Now the Internet archive includes texts, audio, moving images, and software as well as archived web pages.

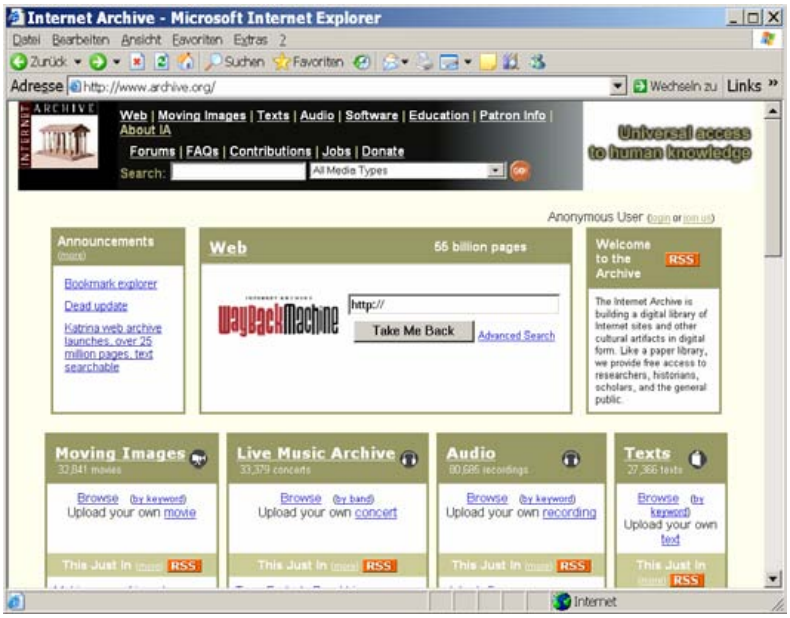


Fig. 4. Home page of the Internet Archive.

The Internet archive develops and operates the open source spider "Heritrix"[22] in order to harvest some of the contents that it archives, but mostly it relies on Alexa Internet for the bi-monthly donation of the search engine's document cache. Each web crawl (Fig. 5) consists of about 100 terabytes of web content spanning 4 billion pages and 8 million sites. Presently, the archive contains over 55 billion web pages.

Besides web pages, the Internet archive archives also other digital artefacts like films, music, scanned texts and even software. An example of a web site that has been retrieved using the "way-back-machine" of the Internet archive can be found in Fig. 6.

[21] http://www.archive.org/
[22] http://crawler.archive.org/

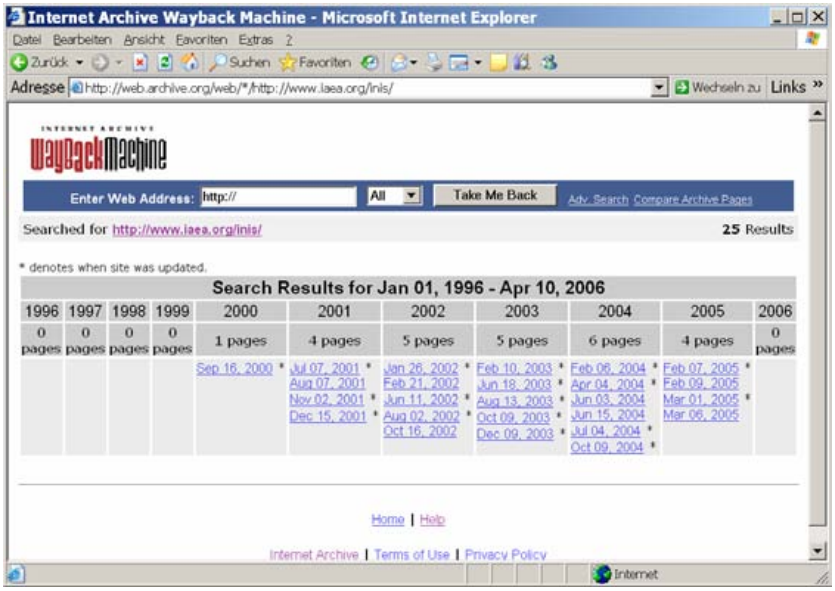


Fig.5. Way-Back-Machine – the user interface to the web archive.

Fig. 6. "Way-Back-Machine" of the Internet Archive.

3.2. AOLA — Austrian on-line archive

The Austrian National Library (OeNB) together with the Department of Software Technology (IFS) at the Technical University of Vienna initiated the AOLA project (Austrian On-Line Archive)[23]. The goal of this project was to build an archive of the Austrian web space.

[23]http://www.ifs.tuwien.ac.at/~aola/

Documents were harvested at certain time intervals to produce snapshots of the Austrian web space.

Initially, the project started with the Nedlib harvester[24]. Several modifications and expansions had to be made in order to make it fit the needs. Between May 7th and 16th 2001 a first attempt was made to take a snapshot of the Austrian web space. Approximately 1 gigabyte of data was collected per day from the .at domain, as well as several manually selected sites from other domains, amongst them .com, .cc, .org and others. In that time about 666 000 unique URLs were harvested from 1 210 different sites. All in all 8.3 gigabyte of data were stored. During this first pilot run numerous problems with the Nedlib harvester were discovered.

For a second run, the combine harvester[25] of the Swedish initiative was used. As the combine harvester was initially developed for indexing purposes, rather than for web archiving, several adaptations had to be performed. In doing so, the project followed a close cooperation with the Swedish Kulturaw3 project, benefiting from their experience and efforts already put into modifying the original harvester. Even though, some of the functionality that is desirable for an archiving system could not be included.

The AOLA project has successfully completed the second pilot phase, using the adapted combine harvester. The harvester was collecting data at a rate of about 7GB per day, having created an archive of 150 GB, including more than 2.7 million pages from about 21 000 sites by June 21st, 2001.

3.3. Pandora — Australia's web archive

PANDORA (*P*reserving and *A*ccessing *N*etworked *D*ocumentary *R*esources of *A*ustralia)[26], Australia's web archive, is a growing collection of Australian online publications, established initially by the National Library of Australia in 1996, and now built in collaboration with nine other Australian libraries and cultural collecting organizations.

The purpose of the PANDORA Archive is to collect and provide long-term access to selected online publications and web sites that are about Australia, are by an Australian author on a subject of social, political, cultural, religious, scientific or economic significance and relevance to Australia, or are by an Australian author of recognised authority and make a contribution to international knowledge.

As a national library, the National Library of Australia has a responsibility to develop collections of library materials, regardless of format, which document the history and culture of Australia and the Australian people. When, in the mid-1990s, unique Australian information began to be published on the World Wide Web, the Library realised that it would be necessary to extend its collecting activity to this new domain. It also realised that, because of the volume of material being published and the complexity of the task of collecting it, it would be impossible to build an archive of sufficient depth and breadth alone. It therefore invited other deposit libraries with similar responsibilities, namely the State libraries, and other cultural collecting agencies (PANDORA participants) to join it.

To support the acquisition and management of increasing volumes of data, as well as to support more efficient archive building for participants at remote work stations, the Library developed the PANDORA digital archiving system (PANDAS[27]), the first release of which

[24]http://nedlib.kb.nl/
[25]http://combine.it.lth.se/
[26]http://pandora.nla.gov.au/
[27]http://pandora.nla.gov.au/pandas.html

took place in June 2001, with version 2 being released in August 2002. Further development of the software continues, with the release of version 3 expected in 2006.

An evaluation system for version 3.0 of PANDAS is expected to be made available in the first half of 2006. The library also intends to make the PANDAS software itself freely available to anyone who wants to use it.

Presently the archive holds over 30 million files, amounting to over 1 TB of data.

3.4. Kulturarw3 — National Library of Sweden

The Royal Library, National Library of Sweden, tasked since 1661 with collecting all Swedish printed publications, has inaugurated a project, entitled Kulturarw³, with a view to the long-term preservation of electronic documents. The aim of this project is to test methods of collecting, preserving and providing access to Swedish electronic documents which are accessible online in such a way that they can be regarded as published. Through this project the Royal Library is also laying the foundations of a collection of Swedish electronic publishing for our time and for coming generations. Kulturaw³ uses NWA-Combine, a modified version of the Combine harvester[28], optimized for archiving purposes.

The collected material is stored on magnetic DLT tape in a tape robot and it is available via hierarchical storage management software. It takes approximately one minute to upload a website to disk. From disk one can surf the archive with the aid of ordinary web browsers. Today, there are two main collections: The all-over-Sweden bulk collection, and the daily newspaper collection:

- The bulk collection consists of approx. 306 million URLs and the amount of data is 10 TB.
- The daily newspaper collection comprises 0.5 TB and 6 million URLs.

Kulturarw3 has acquired a license for ArchiveWare[29], which can keep track of the separate parts of 10 million web pages and make them accessible.

The archive is available only within Library's realms, but visitors may consult the collection at terminals in the Library. The archive is not indexed; hence one must know the exact address of any web page that one wants to access.

Similar activities are carried out in the other Nordic countries. In order to coordinate their activities a working group has been formed, Nordic web archive (NWA), and scheduled meetings are held. The mission of the group is to exchange ideas and experiences but also to supply a forum to create and develop compatible software.

3.5. Project NEDLIB

Project NEDLIB[30] - Networked European Deposit Library - was launched on 1st January 1998 with funding from the European Commission's Telematics application programme. It aims at constructing the basic infrastructure upon which a networked European deposit library can be built. The objectives of NEDLIB concur with the mission of national deposit libraries to ensure that electronic publications of the present can be used now and in the future.

The NEDLIB project has ended as of 31st January 2001.

---

[28] http://combine.it.lth.se/
[29] http://www.fuzzygroup.com/
[30] http://nedlib.kb.nl/

As a result of the project, several reports and publications are available from the NEDLIB web site (http://nedlib.kb.nl/). The NEDLIB harvester is a freeware application for harvesting and archiving web resources. It has been optimised for web archiving, and therefore it has many features, which normal harvesters lack, such as the archive module. There are also numerous features which differ from standard harvesters. For instance, harvesting priorities have been optimised for archiving purposes.

Although the application has been designed for processing very large collections such as the national web space, it can also be used for archiving documents from an individual server or its subdirectory.

The application is maintained jointly by Helsinki University Library and Center for Scientific Computing. Documentation and source code are available at: http://www.csc.fi/sovellus/nedlib/.

3.6. LOCKSS permanent publishing on the web

LOCKSS[31] is different from the other Internet archive initiatives that have been discussed so far in that it attempts to archive a component of the deep web.

LOCKSS (Lots of Copies Keep Stuff Safe) is open source software that provides librarians with an easy and inexpensive way to collect, store, preserve, and provide access to their own, local copy of authorized content they purchase. Running on standard desktop hardware and requiring almost no technical administration, LOCKSS converts a personal computer into a digital preservation appliance, creating low-cost, persistent, accessible copies of e-journal content as it is published. Since pages in these appliances are never deleted, the local community's access to that content is safeguarded. Accuracy and completeness of LOCKSS appliances is assured through a robust and secure, peer-to-peer polling and reputation system.

A library collects newly published content from the target e-journals using a web crawler, similar to those used by search engines. It continually compares the content it has collected with other appliances and synchronises the data stores. Participating libraries provide browsers in their community with access to the publisher's content or the preserved content as appropriate.

Before LOCKSS can preserve a journal, the publisher has to give permission for the LOCKSS system to collect and preserve the contents. Publishers do this by adding a page to the journal's web site containing a permission statement, and links to the issues of the journal as they are published.

When a request for a page from a preserved journal arrives, it is first forwarded to the publisher. If the publisher returns content, that is what the browser receives. Otherwise the preserved copy is served.

3.7. EU MINERVA[32]

MINERVA is a network of Member States' Ministries to discuss, correlate and harmonise activities carried out in digitisation of cultural and scientific content for creating an agreed European common platform, recommendations and guidelines about digitisation, metadata, long-term accessibility and preservation.

---

[31] http://www.lockss.org/
[32] http://www.minervaeurope.org

Due to the high level of commitment assured by the involvement of EU governments, it aims to co-ordinate national programmes, and its approach is strongly based on the principle of embedded ness in national digitisation activities.

It establishes contacts with other European countries, international organizations, associations, networks, international and national projects involved in this sector, with a special focus on actions carried out in the DigiCult action of IST.

4. WEB HARVESTING METHODOLOGIES

Before a web harvesting activity of non-trivial extent can be envisaged, it is necessary to define the framework for the project by answering the following questions:

- Who will benefit from the archive, how will the archive be used?
- What sort of materials shall be considered for downloading?
- How comprehensive shall the archive be?
- How long is the envisaged live-time of the archive?
- How much manpower is available for creating, configuring and maintaining the archive?
- How much Internet bandwidth, storage space and processing power is affordable within the project's budget?

Within the limiting factors and conditions of its framework, the web harvesting project can be optimised to best fit the expectations. Important preparations that need to be conducted in order to define the structure of the web harvesting process and of the resulting web archive are to describe how contents to be harvested are located, in what formats data will be stored and what the fundamental architecture of the archive shall be.

4.1. Identifying contents

The Internet contains a huge and still rapidly growing collection of materials. Most web harvesting projects will therefore not be able to download more than a small fraction of the entire content of the Internet. In order to utilize the limited resources efficiently, a policy for selecting materials must be created. Useful selection criteria within a harvesting policy could be:

- Reputation of the publishing or hosting organization: Download from a set of well known nuclear research institutes, nuclear authorities, etc.
- Relevance of document content: Decide whether or not a document is relevant by analysing its content of specific keywords.
- Document type: Research reports, annual reports, news articles and (i.e. conference registration forms are different document types.
- Document format: Download, for example, only documents in PDF.
- Publication language: Download only documents that are published in one of a given set of languages.
- Minimum document size: Consider documents that are smaller than a certain minimal size as trivial and neglect them.
- Maximum document size: Do not download documents that are larger than a certain threshold in order to preserve resources.
- Exploit links to full-texts in bibliographic databases.
- Exploit resource indexes. Organizations who participate in the Open Archives Initiative, in LOCKSS or in similar initiatives provide indexes of the documents that are available from them.

Some of the practical steps that need to be taken for identifying and selecting contents for a nuclear web archive are discussed in the following paragraphs.

4.1.1. Exploiting link lists and Internet directories

Most Internet sites provide a list of links to Internet resources that are related to their own sphere of interest. These lists are usually assembled manually and can therefore be treated like recommendations by a knowledgeable person. These lists are, unfortunately, often not very

long and therefore they are difficult to exploit, since one would have to screen rather many of them. An attempt, to at least semi automate the selection process, could work as follows:

(a) One starts with a list of known web sites that are of interest in a given context.
(b) A crawler can flag web pages with more than a given number (i.e. 20) of external links as "link lists".
(c) The linked external web sites will be preliminarily harvested.
(d) The content of the harvested documents will be analysed automatically (see next paragraph), and depending on the number of relevant documents found, the web site will be classified as *'of interest in the given context'*.
(e) Any new site that is of interest will again be screened for link lists.

Substantial link lists, usually classified by subject, are provided by Internet directories such as the Open Directory Project[33] or the IAEA's Internet Directory of Nuclear Resources[34]. Such directories can provide thousands of relevant links in a given field and they still have the benefit of having been assembled by a human being.

4.1.2. Analyzing document contents

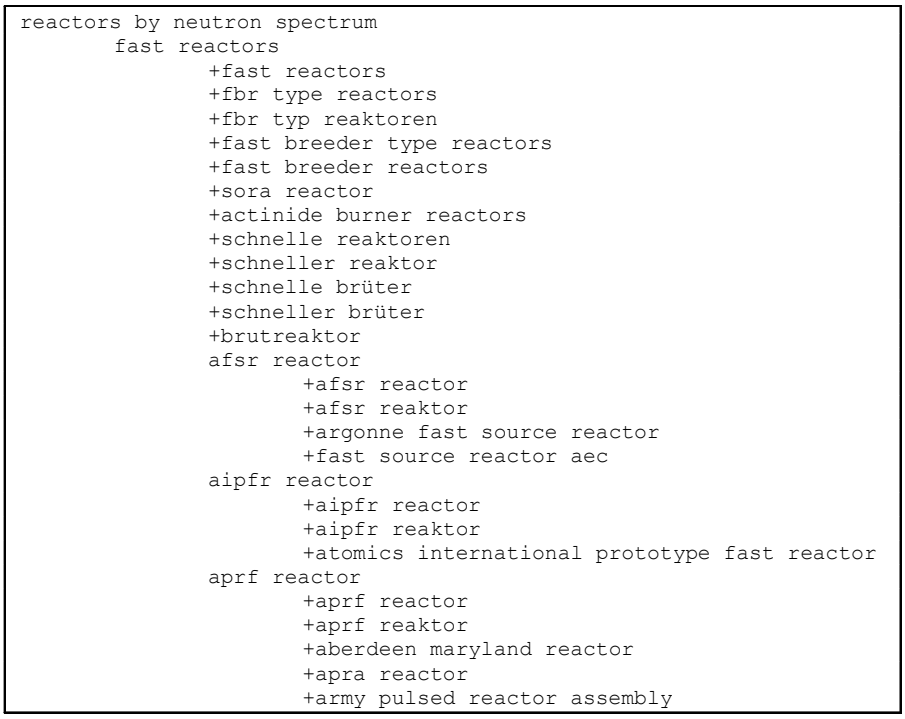
Rather than identifying entire web sites that are relevant in the nuclear domain and then screening them for useful documents, one can use Internet search engines and search directly for individual documents.

The terms used in taxonomies in the nuclear field are a rich source of keywords that can be used to construct query strings for different Internet search engines (Fig. 7.). Since the different ranking algorithms of the various search engines give raise to different result sets, it is worth using several engines in parallel. This is most likely a rewarding application for meta search engines.

Whether or not a given document is relevant in a specific context can easily be recognised by a knowledgeable person, but it is difficult to determine automatically. Simple keyword statistics tend to work rather poorly as a document filter[35]. We can expect somewhat better results, if a search engine that uses semantic concept expansion, like Convera RetrievalWare, is deployed.

Since document filtering is used in any case when the archive is searched by its users, it might be argued that one could simply archive all documents that have been obtained by crawling web sites or by launching searches on the Internet. Document filtering at indexing time has, however, the advantage of saving storage space.

[33] http://www.dmoz.org/
[34] http://www.iaea.org/inis/ws/
[35] The poor performance of keyword statistics as a document filter was the reason why searching the Internet, before Google became available, was a rather frustrating experience. Long lists of keyword-matching, but nevertheless irrelevant "hits" had to be scanned manually in order to find relevant documents.



Under each node, the terms of the corresponding "synset" are prefixed by a "+" sign. These terms can be used to search for relevant documents on the Internet.

Fig. 7. Snippet taken from nuclear reactor taxonomy.

4.1.3. Restricting document types

In this context the type of a document refers to its content, intended use and target audience rather than how the document is coded (PDF, HTML, PPT, etc.). Research reports, annual reports and technical manuals are examples for different document types. Generally, we can distinguish document by a number of criteria:

(1) **content type:** scientific/technical – news item – administrative;

(2) **target audience: experts – students – general public; and**

(3) **temporal validity:** long term – medium term – short term.

Deciding which types of documents should be targeted for harvesting determines, to a large extent, the nature of the resulting archive.

- A scientific archive, for example, will certainly concentrate on documents of a scientific/technical nature that target experts and remain valid for a long time.
- A news archive, on the other hand, will try to capture news items, regardless of their target audience. The temporal validity of news items is usually short term.
- Administrative contents are usually of interest only within their territory of validity (i.e. Procedures required to apply for a research grant at a university is relevant for the students and the prospective students of that university).

In practical terms it is difficult to distinguish the different types of documents in an automated way. The most promising method for selecting documents by type for harvesting is to analyse the structure of major websites and to harvest only from specific locations that are likely to contain documents of the required type.

IAEA's NuArch pilot project can serve as an example where a small number of core web sites in the nuclear field underwent detailed analysis in order to identify specific locations where the targeted document types are available. The harvester for this project was configured to crawl and download only within the identified locations. Practical experience with NuArch shows, however, that the cost for manual site evaluations is very high.

For larger harvesting projects it might be acceptable to harvest all types of documents and to rely on advanced retrieval techniques for filtering only documents of relevant type at search time.

Nevertheless, the temporal validity of documents is of relevance for the frequency with which a web site is crawled and documents are downloaded. While a web site that contains mostly reports of long validity is adequately archived if it is harvested a few times per year, the web site of a scientific e-journal will need to be scanned according to the publication cycle of that e-journal.

4.1.4. Selecting document formats

Different document formats are customarily used for different document types:

- Print formats like **PDF, PS, EPS, TEX, DJVU** are often used for self-contained documents that have reached the maturity to be ready for printing.
- Presentation formats like **PPT, SWF** contain mostly the visual components of presentations. Without the wording of the presentation, they are usually of little value. Similarly, multimedia formats like **MP3, MP4, AVI, WAV**, etc., are often recorded presentations or video clips that serve as illustrations, but are of little value by themselves.
- Word processor formats like **DOC, RTF, SXW** might indicate that the document is still being worked on.
- **XML, XLS** and **MDB** indicate databases or spread sheets.

Native web content is usually coded in **HTML** or in **XHTML**, and file extensions of the type .asp, .aspx, .php, .cfm are often used in order to indicate that the page content is created dynamically. A particular property of HTML coded documents is that they are normally optimised for viewing on a computer screen. Larger documents are therefore in most cases broken down into many individual HTML pages. This presents the automated harvester with the problem of how to decide which pages belong to a document. Indeed, the very concept of well defined documents with sequential pages is at risk in the context of closely interlinked pages of a web site.

Practical considerations make it likely that harvesting is either restricted to document formats, like print formats and word processor formats, that are used for self contained documents or entire web sites are harvested. The former case avoids the complexities that are associated with archiving distributed documents and takes at the same time advantage of the higher average content quality of documents that are ready for printing. The latter case ensures that distributed documents are captured entirely, without having to determine explicitly which files or components belong to a given distributed document.

For the IAEA's NuArch pilot project, harvesting of self-contained documents only was selected. In the process of conducting the project it was discovered that files that are coded in PDF are often fragmented, too. Obviously, webmasters wanted to enable users to download files that are considered large more easily by breaking documents into several PDF documents. In order to capture such fractured documents completely, it is essential that the web crawler is able to follow links inside PDF files.

4.1.5. Language considerations

Whether or not it is useful to harvest and store documents that are presented in languages that are not understood by the project team depends on the intended use of the archive. In the case of the IAEA, the following options has been considered:

- **English only:** English is the official working language of the IAEA. It is understood by every staff member and the majority of nuclear related documents worldwide are published in English.
- **IAEA official languages:** English, French, Spanish, Arabic, Chinese and Russian are the official languages of the IAEA, although few staff members will be able to understand all of them.
- All languages that use the extended ASCII set of characters.
- Any language.

If languages that use characters outside the extended ASCII set of characters are considered, special provisions for storing such characters and for displaying them must be made. This can cause considerable technical difficulties, particularly with languages that are not usually written in from left to right in horizontal lines.

Practical problems might also arise from the fact that the project team might not be able to perform quality control tasks on documents that they cannot read.

Advanced features of the document retrieval software, like the treatment of grammatical variations, verb stemming and in particular semantic analysis will only function properly for a pre-defined set of languages that are supported by the retrieval software.

The envisaged knowledge domain of nuclear related documents and the international character of the IAEA make it favourable to cover documents in any language.

4.1.6. Document-size as a quality indicator

Document-size is a very weak quality indicator. While it is true that short documents of less than one page are likely to contain only abstracts, such abstracts might still carry valuable contents.

Moreover, file sizes depend strongly on methods of data compression and on the graphical content of documents. To suppress the downloading of large documents in order to save storage space and Internet bandwidth is hardly justifiable anymore, since these commodities are nowadays rather inexpensive.

4.1.7. Spin offs from bibliographic reference banks

In recent years, bibliographic databases, like the INIS Bibliographic Database, have started to record the URLs of full-text resources that are available on the web as a component of the bibliographic descriptions of documents. Adding thus referenced documents to the archive has several advantages:

- A full bibliographic description of the documents is readily available.
- The inclusion of documents into a bibliographic database indicates that the document is in a certain subject scope and of a certain minimal quality.
- Harvesting and archiving thus referenced documents adds value to the bibliographic database, which can now refer to the archive as a permanent document source.
- Thus referenced documents are often part of the deep web and would therefore not be harvested if it was not for the links that are stored in bibliographic databases.

In order to be able to draw full mutual advantage from links that are stored in bibliographic reference banks, it is recommended to envisage a close collaboration with the INIS Bibliographic Database, the largest bibliographic database in the nuclear field.

4.1.8. Resource manifests

Resource manifests are data files that enumerate the resources that are available from a given location. Such manifests could, where available, be used in addition or alternatively to spidering web sites in order to locate documents that are available for downloading. The references in manifests have similar value than the above discussed links that are stored in bibliographic databases. They, too, have undergone quality control and may be supplied with a set of suitable metadata.

Examples for widely used resource manifests have been discussed above under the headings "Open Archives Initiative" and "LOCKSS Permanent Publishing".

Given the standing and reputation of the IAEA, it is quite likely that collaborating organizations could be persuaded to provide resource manifests that are purposed for the IAEA's project to build an archive of nuclear materials that are available on the Internet.

4.2. Data formats for archiving

Information on the Internet is published in many different formats. There are not only the different file formats like PDF, DOC, PPT that are used for self contained documents; often a mixture of differently formatted resources (java-script, cascading style sheets, images, etc) are used to represent a given document in HTML. The various data formats can to some extend, but usually not precisely, transformed into certain archival formats. How the different data formats that are today found on the Internet are archived, is an important question since it has profound impacts on the long-term accessibility of the archive.

4.2.1. Preserving native formats

Preserving all resources that have been harvested from the Internet in their native formats seems the most logical and the most straightforward thing to do, at first glance. There are, however, a few drawbacks to this approach:

- Fully qualified URLs will not function in the environment of an archive.
- Some of the browser plug-ins necessary to view today's web contents might not be available in future.
- Server-side components that are needed to create or display web content would need to be emulated by the archive.

In order to address these issues, modifications need to be introduced to data files that are captured in their native format.

All fully qualified URLs that point to resources which have also been captured, need to be replaced by relative URLs. This is trivial to do in the case of statically coded HTML link tags, but it might present considerable problems when links are coded in Java-Script or other scripting languages.

A particular case in this context concerns the redirection of links. Links that are redirected by the server that hosts the documents, or by a dedicated PURL service, can only be turned into relative links if the web crawler records all redirects, and if it makes this information available to a process that modifies the links inside the harvested resources.

Links that point to resources that are not part of the archive should, on the contrary, be turned into absolute URLs, or they should bring up a standard message informing the user that the

requested resource is not part of the archive. If such "external" links are made absolute, then the user is at least able to access the live resource on the Internet for as long as it exists.

Instead of modifying links within the harvested resources, it could also be envisaged that the archive environment provides a mechanism, analogous to server-side redirects, which takes care of serving the appropriate resource.

In order to keep an archive with native formats functional over long periods of time, the software for viewing the files, like contemporary web browsers and their specific plug-ins would have to be preserved as well. This simple sounding task might in reality be quite difficult to achieve, since it is unclear if and to what extent contemporary software will be functional on future hardware platforms. Hence, the maintenance of a web archive with native file formats entails the provision of software emulations of today's software on future computer systems.

One of the biggest hurdles on the way to creating an Internet archive that provides a similar user experience as the "live Internet", is the emulation of essential server-side components. At present no technology that emulates comprehensively server-side components is known. Such emulations would have to be developed and programmed on a case by case basis, which is of course prohibitively time consuming and expensive. Consequently, the preservation of server-side functionality, which includes the access to most of the deep web, must be considered to be outside the scope of present day web archival efforts.

4.2.2. Harvesting web content into an MHTML archive

Some web browsers, such as Internet Explorer and Opera[36], allow the user to save a web page, including resources, as an MHTML file.

MHTML stands for MIME HTML. It is a standard for including resources that in usual HTTP pages are linked externally, such as images and sound files, in the same file as the HTML code. The first part of an MHTML resource is the HTML file, encoded normally. Subsequent parts are additional resources, identified by their original URLs.

Archiving web content in MHTML is quite similar to preserving the native format of web resources, but it simplifies the version control of HTML pages. By including resources like graphics, sounds and style sheets into the MHTML file, it can be ensured that all components of an HTML page are captured at the same time. If the linked resources are stored separately, care has to be taken explicitly that the versions of all resources and of the HTML file are synchronised.

4.2.3. Printing to PDF

Adobe Portable Document Format is ultimately a print format and it is therefore able to capture any screen content that can be printed. According to Adobe's claims[37], PDF should be quite suitable for web content archival:

Open format – De facto standard for more secure, dependable electronic information exchange — recognized by industries and governments around the world (Compliant with industry standards including PDF/A, PDF/X, and PDF/E).

Multiplatform – Viewable and printable on any platform (Macintosh, Microsoft® Windows®, UNIX®, and many mobile platforms).

---

[36] http://www.opera.com/
[37] http://www.adobe.com/products/acrobat/adobepdf.html

Extensible – More than 1,800 vendors worldwide offer PDF-based solutions including creation, plug-in, consulting, training, and support tools.

Trusted and reliable – More than 200 million PDF documents on the web today serve as evidence of the number of organizations that rely on Adobe PDF to capture information.

Maintain information integrity – Adobe PDF files look exactly like original documents and preserve source file information (text, drawings, 3D, full-color graphics, photos, and even business logic) regardless of the application used to create them.

Searchable – Leverage full-text search features to locate words, bookmarks, and data fields in documents.

Accessible – Adobe PDF documents work with assistive technology to help make information accessible to people with disabilities.

The general optical impression of the web page which is preserved (converted to .PDF) comparative with the original one is similar. For example, one difference is that the header part of the page with its search interface and the navigation links is missing in the PDF. The reason for this is most likely that the web page is provided with separate style sheets for displaying in a browser and for printing. Another important difference that cannot be seen on the screenshot is that the links on the page are not functional on the PDF.

These observations indicate that PDF can be used to preserve the optical impression of individual web pages, but the navigation structures that turn a collection of data files into a web of resources, are lost. The suitability of PDF for archiving web content is therefore somewhat limited.

It is however, worth noting that ISO Standard 19005-1:2005 defines the format PDF/A for the long-term archiving of electronic documents. It is based on the PDF reference version 1.4 from Adobe Systems Inc. (implemented in Adobe Acrobat 5) and it is in fact a subset of PDF, leaving out PDF features that are not suited for long-term archiving.

4.3. Archive architecture

The main functionalities of an archive of Internet resources that are relevant to the nuclear field are:

- Locating of relevant Internet resources;
- Harvesting of Internet resources;
- Providing fast data access with version control;
- Secure long-term storage;
- Advanced search facilities and resource navigation tools; and
- Platform independent document readers.

In order to keep the complexity of the archive as low as possible, each one of the listed functionalities is most suitably implemented in a separate component of the archive. These components represent the building blocks of the archive and they should be kept as independent as possible.

Within this modular approach, great care must be taken to define the interfaces between the individual components with sufficient precision, so that they can be developed, updated and replaced independently of each other.

The following sections discuss the individual modules of the archive in more detail.

4.3.1. The resource-locator module

The core component of the resource-locator module (Fig. 8) is the web crawler. It will produce a list of web locations (URLs) of individual resources that are candidates for harvesting. Thousands of web locations that are readily available from the IAEA's Internet directory of nuclear related resources and similar resource listings can be used as seed-URLs for the crawler.

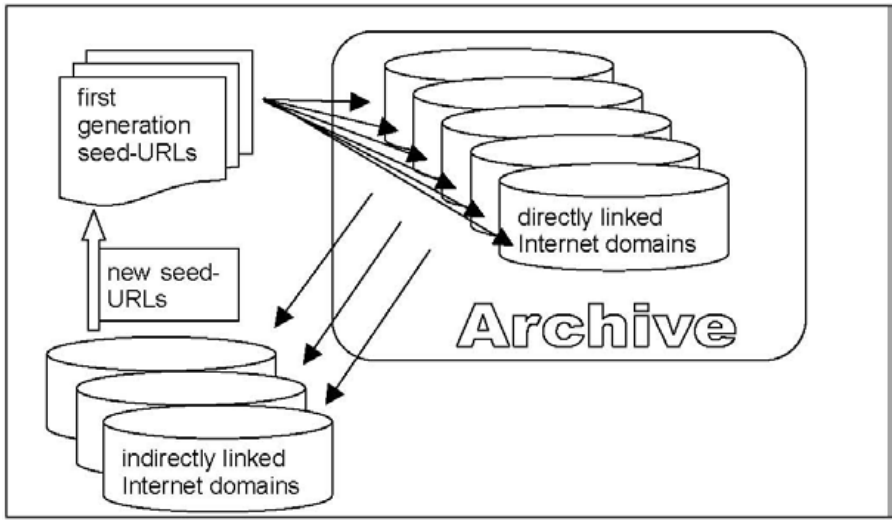


Fig. 8. Internet domains that are directly linked to a set of first generation seed-URLs are by definition in scope, and their contents will be archived. Links from archived web-pages identify a new set of Internet domains that are only indirectly linked to the initial seed-URLs. They need to be screened if they are in scope for the Archive, and the list of seed-URLs needs to be updated accordingly.

Whether or not, links are to be followed, they will be determined by a set of filter functions. A second set of filters will independently determine which files are to be harvested for archival purposes. Regular expressions or somewhat modified rule sets have in the past proven suitable and flexible enough to build powerful filters.

A by-product of crawling the initially referenced web sites will be a list of links to domains that are referenced by web pages of the initial set without belonging to the initial set. Because the URLs on this secondary list of Internet domains have been referenced by nuclear-related web sites, it is quite likely that many of them are relevant, too. Which of these newly discovered domains should be added to the list of Internet domains that are to be spidered, can be determined either manually or automatically, by using the methodology that has been outlined in the above Section "4.1.2 Analyzing document contents".

Another means to extend and to complete the list of seed-URLs is to search the Internet with common Internet search engines for documents that are in scope, as was outlined in Section 4.1.2. Since the operators of high-quality free search engines will in most cases not permit such extensive and automated use free of charge, contractual agreements will have to be negotiated.

Searching the Internet will, apart from potential seed-URLs, deliver the URLs of self contained documents in text formats such as PDF. Such URLs, together with links to full-texts that are contained in bibliographic databases and resource locations that can be obtained from resource manifests (Section 4.1.8) should be used to enrich the list of harvesting requests that is to be assembled and to be passed to the harvester.

In assembling the harvesting list, care should be taken to include all components of self contained text files that have been broken down into several parts, so that they can be downloaded more easily.

4.3.2. The harvester

All modules of the archive must implement measures to reduce the bandwidth that is needed for the maintenance of the archive. This saves project resources, but more importantly, it reduces the burden on the harvested web sites. Files that need to be accessed by the web crawler in order to follow their links should therefore be requested from the data access module, if their header information indicates that they have not changed since the last visit.

In order to be gentle to the resources of the harvested web sites, it is necessary to be able to control both the delay between downloads and the bandwidth that is used for downloading individually for each domain.

Both the crawler as well as the harvester will record a range of metadata that are needed for the operation of the archive. In particular, the harvester will record any link redirects. Another type of metadata that might be essential for the functioning of the harvester are cookies and the hidden parameters of forms.

4.3.3. The data storage and access module with version control

The data storage and access module is the core component of the archive. Its main functions are:

- Exchanging data with the harvester;
- Storing all documents and auxiliary data;
- Maintaining the integrity of compound documents (i.e. HTML with linked images, Java-Script and style-sheets);
- Implementing version control;
- Mapping links to archived documents;
- Emulating server-side redirects;
- Supporting the search interface; and
- Interfacing with the document reader.

The approach taken by most web archival products is to store the harvested files in a file system. Some products store, additionally, some metadata and auxiliary information in an associated database. This procedure derives its simplicity from the fact that contemporary web browsers are readily able to display and navigate the web pages of a locally stored web site as long as all links are relative. Some products try even to convert all absolute URLs to relative URLs in order to make the user experience as authentic as possible; others accept that following links with absolute URLs will divert the user to the live site.

Some of the weak points of file systems for data storage are:

- Different rules apply for forming file names and for URLs;
- Version control is difficult to implement;
- Server-side redirects are not available;
- File types are usually indicated via file extensions; and
- Converting absolute URLs to relative URLs is not always successful.

In order to avoid the difficulties that arise from the use of file systems as data stores, it is recommended to store all harvested data in a database. The data storage and access module must support efficient interfaces for all the other modules that need to exchange information with the archive's main data storage.

4.3.4. The long-term storage module

Thomas Jefferson is quoted to have said: "The lost cannot be recovered, but let us save what remains: not by vaults and locks which fence them from the public eye and use, in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident"[38].

Storing a backup of the entire system in a secure data vault is of course a good and necessary step towards avoiding the loss of data in case of a local failure. It should be implemented in any case.

If we are concerned with the long term maintenance of knowledge assets, however, we need to recognise that backup systems, however sophisticated they may be, do not prevent the occurrence of the most common reason for loss of knowledge – namely the loss of interest. Whenever the interest in the contents of the vision of the archive wanes, funding will be cut and backup systems will be discontinued. Truly long term storage must therefore account for scenarios that exceed the attention span of the initiating organization.

In order to give the archive the stability that it needs to make the efforts for its creation worthwhile, internationally distributed custodianship and geographically multiplied backup installations are indispensable.

The long term storage "module" must therefore, beyond its technical functionality, be backed up by international arrangements, similar to the presently existing INIS system.

4.3.5. The search and resource navigation tool

Most Internet archive projects permit the users to browse and navigate the archived contents, as long as the user knows precise URLs. Search capability is usually absent or only available in a very rudimentary form.

The technical implementation of an advanced retrieval interface is presently being developed in the context of the IAEA's NuArch pilot project. Important functionalities of the envisaged interface are:

- Search for the latest versions of documents that were once or still are available on the Internet;
- Search only for documents that are still available on the Internet (online when last visited by crawler);
- Search for all documents (including older versions which are not online any more);
- Search for documents which were online during a specific period of time (online-start, online-end);
- The retrieved documents can be downloaded to a local storage medium;
- Link to www for online-version (where available) and access to all stored versions in the archive; and
- Dynamic classification of result sets according to selected knowledge domain.

The last point about dynamic classification deserves particular attention. It requires the development of taxonomies for the knowledge domains of interest and it necessitates a search application that can leverage the semantic structure of these taxonomies.

---

[38] Thomas Jefferson to Ebenezer Hazard, Philadelphia, February 18, 1791. http://www.firstmonday.org/issues/issue2_8/cox/

4.3.6. Platform independent document reader architecture

Since all the documents of the archive are harvested from the Internet, displaying them with a contemporary web browser seems to be a trivial aspect. In the case of a long term archive, however, this is not so!

The difficulties of providing a document reader for a digital archive becomes immediately transparent, if we consider the task of displaying a digitally preserved document that has been created with an early word processor, 20 years ago. We must assume that the originally used word processor has by now sunk into oblivion and the likelihood that any modern word processor might be able to import the propriety format of the old document, is rather slim.

It is clear that, besides preserving the document, the archive must also preserve the software appliances that are needed to visualise the stored documents. But again, we are faced with a similar problem: A 20 year old word processor is not likely to function on a modern computer operating system! The same situation appears if we consider preserving historic operating systems which will of course not function on present or future hardware platforms.

Let us consider three different approaches for providing document readers for the archive:

- *Option 1:* Maintain hardware platforms, together with their operating systems and software.
- *Option 2:* Convert all contents of the archive into an open, standards compliant data format (for example PDF/A) that is suitable for long term preservation.
- *Option 3:* Preserve display applications together with the documents and maintain emulators for suitable operating systems on contemporary computers.

Option 1 is, at least in the beginning, the cheapest possibility to provide access to historic documents. It is however not at all suitable for long term archival, because any hardware will fail beyond repair at some stage.

Option 2 describes a feasible scenario. Its main disadvantage is that any document features that are not supported by the archival format are necessarily lost. If the main intention of the archive is to preserve document contents, rather than the exact look and feel of contemporary Internet resources, option 2 will be cheapest to implement and maintain in the long run. Because it is based on open standards, it offers the biggest chance that document contents can be recovered after truly long storage times, too.

Option 3 preserves the exact look and feel of contemporary Internet resources. It is expensive to maintain, because suitable emulators need to be created, maintained and ported to new computer platforms constantly. However, emulators for most of the first generation popular home computers are readily available today. Although these emulators serve presently mostly to satisfy nostalgic motions, the feasibility of the approach is proven.

In order to maximise durability and user experience of the archive, a dual approach is proposed. All documents should be stored in both their native formats and in an archival format. This ensures the long term preservation of contents, and at the same time it enables contemporary users to experience the full functionality of the archived content. For the medium term future, when present day Internet applications begin to become inaccessible, option 3 remains open and available for implementation.

5. SUMMARY AND RECOMMENDATIONS

Summarizing the above sections, contemporary activities in the domain of web harvesting, document archiving and Internet access technology have been researched and evaluated in order to obtain a contemporary technology overview. In Section 4, several aspects of possible web harvesting methodologies have been discussed in some detail. In this section, ways for managing the project in a sustainable manner and some general recommendations to the pilot project will be outlined.

5.1. Project management

Below are given some recommendations for managing the project using the sum of all insights that have been gained in similar projects and during the preparation of this report.

5.1.1. Prospective users' expectations and needs

The technical evaluation of the web harvesting and archival processes has shown many different possibilities for implementing a web archive. The choices between the different options can and must be used to optimise the entire project in a way that best meets the prospective users' expectations and needs. Gaining the support of the prospective users is a weighty factor among the parameters that determine the success or failure of the web harvesting project.

The most important recommendation is to determine, as precisely as possible, who the future users of the system will be and what those users need or expect.

Examples for distinct user groups and their core domain of interest might be:

Scientist or engineer: research reports, review papers, scientific publications, safety standards, regulations and technical standards, compilations of data

Journalist: news items, white papers, annual reports, opinions, legal texts

Student: lecture notes, scientific publications, text books

Nuclear Safeguard: specific news items, research reports, scientific publications, annual reports, country specific information

Member of the public: news items, opinions

As soon as particular user groups have been identified, a few concrete use cases should be developed and documented. Interested user groups within the IAEA can be very important supporters during the project preparation and implementation.

5.1.2. Sustainable management of a long term archive

The biggest danger for an archive that is supposed to bridge long periods of time is that humans, including the management of international organizations like the IAEA, have only a limited attention span. The importance of a project that was paramount a few years ago tends to wane with time. The only way to secure long time survival and funding, consists in integrating the project firmly into the structure of the IAEA and to give it a status that interlinks directly with the IAEA's Member States.

The International Nuclear Information System, INIS, has proven to be sustainable for more than 35 years[39]. In order to improve its chances for long term sustainability, the web

[39] INIS is quite similar to the web harvesting project, except that it concentrates on printed documents only.

harvesting project could be integrated into INIS' range of activities or, if that is not desirable, a similar organizational setup should be established.

It is also worth noting that the nature of the web harvesting project is only partially IT oriented, its more decisive components are firmly rooted in formal knowledge management. Also, a solid background and overview in the nuclear domain are indispensable for the project management. This distribution of required skills and the placement of stakes in nuclear science and engineering must be reflected in the distribution of responsibilities in the project management team.

5.1.3. Progressive complexity phase-in

The more developed a project is, the more expensive it is to implement modifications in the project specification. Given the overall extent and complexity of a large scale web harvesting project, it is prudent to start by developing basic core components first. More complexity can be added progressively with time, based on the experience gained during previous project stages.

Progressive complexity phase-in should not be confused with "growing the project organically". So called organically-grown projects or products start simple, with no clear view of the future, and additional components or features are added as the need becomes apparent. This works often fine for a short while, but the lack of vision entails that later additions of complexity become more and more difficult as the project progresses. Experience shows that poorly planned projects have to be scraped and started all over again in regular intervals.

A large scale web harvesting project and its associated archive should be planned fully, before implementation commences. During the step-wise increase of the product complexity, only small corrections to the overall project should become necessary. In order to be able to gain the experience that is needed to optimise the final product along its path of development and implementation, it is important that a functional product is available after each complexity increment.

A progressive project implementation for the web harvesting project could look as follows:

Phase 1: Pilot project that encompasses spidering of only a few web sites. Solely self contained documents (PDF) in English are considered. → First experiences with a concept oriented search interface and automatic classification are possible.

Phase 2: Increase spidering to many more web sites. Start including documents that are referenced in bibliographic databases and in resource manifests. → This allows to test the spider on many different web publishing technologies and to gather experience with the scaling of storage, processing and bandwidth.

Phase 3: Include documents, like HTML pages, that are not self contained. → A significantly more complex version of the archive can be tested.

Phase 4: Extend the archive to non-English documents. This entails, among other measures, the development of multilingual taxonomies. → Inspect and evaluate the multilingual features of the retrieval engine and of the automatic classifier. Experiment with a more complex user interface.

Phase 5: Extend the concept oriented search capability and the automatic classification feature to further subject areas. This entails developing new taxonomies and extending existing taxonomies. → Work with users and with the contents of the archive in order to improve the user experience.

The phases 2-5 can be implemented in any order.

5.1.4. Choice of data formats and software components

Specific software products are based on a market study with detailed technical evaluations of the available products, rather than on research into web harvesting methodology. It is, however, clear that certain properties of software products are important for the success of the web harvesting project.

The parameters that determine the choice of software components are:

- Product features;
- Suitability for long term archival;
- Costs for customization and maintenance; and
- License fees.

The first two points touch the fundamental suitability of products to fulfil the required functions and to be sustainable over long periods of time, while the two remaining topics concern the optimal use of project resources.

For a project that aims at preserving information over long periods of time, it is imperative that data formats and interfaces are standards compliant. Proprietary formats are acceptable for testing or for a transitional period only. Cases where standards fall out of use or if they are superseded by new, more appropriate standards, must be covered by an appropriate data migration policy.

Given the standardization of data and data interfaces, the software components are interchangeable. In recent years, an intense debate comparing commercially marketed products, and so called open-source products, has taken place. The main arguments, that open-source products are more sustainable and more standards compliant, which are brought forward by the open-source community, are not necessarily relevant in our case. Standards compliance can be demanded from commercial products, too, and to sustain a major open source product that has lost the support of its initial developers, can be prohibitively expensive (although, it is at least possible). The biggest argument that tilts the odds often in favour of open-source products is, in many cases, in particular in the resources starved public sector, that there are no licence fees applicable. Whether licence fees are decisive or not, in a given case, must be evaluated in the context of the overall cost for the customisation and maintenance of a software product.

Generally, it can be said that widely used, general purpose software, such as web spiders, database servers, web servers, and operating systems for example, are good candidates for the open source market. More specific components, like a retrieval engine or an automatic classifier that are based on semantic context analysis, on the other hand, are probably uniquely available from specialised commercial software producers.

5.1.5. Software integration and operational support

In order to build a web harvesting and archival product, a number of individual components need to be customised, extended and integrated. The IAEA has the choice between charging its own IT personnel with these tasks or to outsource the support. As the IAEA's web harvesting project is an undertaking that is envisaged to require IT support and maintenance over a long period of time, it is important to consider the placement of software integration and operational support carefully.

Table I compares some of the consequences that would arise from the internal or external conduction of the IT tasks.

TABLE I. COMPARISONS BETWEEN INTERNAL AND EXTERNAL TASKS

Item	**IAEA IT staff**	**external contractor**
UN culture	*familiar*	*has to learn*
English working language	*familiar*	*usually not a problem in the IT sector*
international environment	*familiar*	*might be a problem in some countries*
build-up of project expertise	*not possible because of staff rotation policy*	*happens naturally, over time, if contractor remains the same*
availability of manpower	*limited, hiring new staff is difficult*	*flexible, is the contractors responsibility*
placement of new tasks	*depends on the availability of IT staff*	*depends on the availability of funding*
changing specific tasks	*easy*	*subject to contractual restraints*
project management	*difficult because IT staff is under independent IT management*	*easy because supervisor-contractor roles are fixed*
risk of substantial budget overrun	*high in the IT sector*	*the contractor carries this risk*

A similar choice as with the support tasks exists also for the placing of hardware components. Placing hardware components at a contractor's site is of course only beneficial if the same contractor is responsible for the related software maintenance tasks (Table II).

TABLE. II. COMPARISON BETWEEN INTERNAL AND EXTERNAL PLACEMENT OF HARDWARE

Hardware platform for:	**on IAEA site**	**at contractor's site**
web crawler	*possibly bandwidth problems*	*bandwidth is no issue, the contractor must provide what is needed*
indexing	*this CPU-power intensive task, requires dedicated server hardware*	*can re-use the server that was used for software development and testing*
data store	*one copy MUST reside at IAEA for reasons of data security*[40]	*useful for backup purposes*
backup	*secure backup space is limited and expensive*	*secure backup space is cheap*
user interface	*MUST run at IAEA for reasons of data security*	*useful for backup and debugging purposes*

[40] In this context, data security only means that the IAEA wants to secure unrestricted and unobserved access to the data for its staff.

In the short and medium term, it seems likely that a maximally outsourced project bears lower risks and is less expensive. Carefulness should be taken in consideration to avoid dependence on a single contractor. Suitable countermeasures against undue dependence on individual suppliers are:

- To insist on sufficiently detailed project documentation, so that project components can easily be moved to different service suppliers, if the need arises.
- To insist on standards compliant data formats and data exchange interfaces. In the absence of applicable standards, detailed documentation is particularly important.
- To avoid propriety software that is available only from a single source, where possible. If propriety software cannot be avoided, it should be supplied from an established, well reputed software house.

5.2. Technical meeting on web harvesting

A number of web harvesting projects are presently underway or have been conducted recently by several different organizations. In order to be able to build on the experience that was gained in those projects, a technical meeting with the aim of developing a review paper that covers several major web harvesting could be taken in consideration.

5.3. Formal requirements gathering

A major project can only be successful if it is met by the support from its stakeholders. It is therefore important to conduct a careful survey of the user's and stakeholders' requirements and expectations. Important stakeholders should be informed and involved at an early stage, and their consent should be sought.

5.4. Organizational form

In order to obtain sustainability over long periods of time, a decentralised, and collaborative organizational form, similar to how INIS is organized, should be sought.

5.5. Persistence of the archive

In order to protect against loss of data, as many copies of the data store and the access interface should be installed in geographically distributed locations, as possible. Possible locations for such mirror installations could be collaborating nuclear organizations, universities or national libraries.

APPENDIX
RELATED RESOURCES ON THE INTERNET

Open Archives Initiative
http://www.openarchives.org/

LOCKSS permanent publishing on the web
http://www.lockss.org/

3rd ECDL workshop on web archives
http://bibnum.bnf.fr/ECDL/2003/

5th International Web Archiving Workshop and Digital Preservation
http://www.iwaw.net/05/callforpapers.html

Internet Archive, Online Computer Library Center (OCLC)
http://www.oclc.org/

Pandora – Australia's web archive
http://pandora.nla.gov.au/activities.html

International Internet Preservation Consortium
http://netpreserve.org/about/index.php

Uncovering information hidden in web archives (article)
http://www.dlib.org/dlib/december02/rauber/12rauber.html

Harvesting web content into MHTML archive
http://www.codeproject.com/csharp/mhtmllib.asp

MHTML – Article in Wikipedia
http://en.wikipedia.org/wiki/MHTML

Web crawler (Wikipedia article)
http://en.wikipedia.org/wiki/Web_crawler

Internet Archive
http://www.archive.org/

Persistent URL Home Page
http://www.purl.org/

Dublin Core Metadata Initiative (DCMI)
http://dublincore.org/

Web archiving projects (links)
http://www.ifs.tuwien.ac.at/~aola/links.html

DEFINITIONS

Bibliographic database: A bibliographic or library database is a database of bibliographic information, such as article or book citations.

Cache: In computer science, a cache is a collection of data duplicating original values stored elsewhere or computed earlier, where the original data is expensive (usually in terms of access time) to fetch or compute relative to reading the cache. Once the data is stored in the cache, future use can be made by accessing the cached copy.

Cookie: A cookie is a piece of text that a web server can store on a user's hard disk. Cookies allow a web site to store information on a user's machine and later retrieve it. For example, a web site might generate a unique ID number for each visitor and store the ID number on each user's machine using a cookie file.

Deep web: The deep web (or invisible web or hidden web) is the name given to pages on the World Wide Web that are not part of the shallow web that is indexed by common search engines. It consists of pages which are not linked to by other pages like dynamic web pages that are generated in response to a query and contain information stored in databases. The deep web also includes sites that require registration or otherwise limit access to their pages, prohibiting search engines from browsing them and creating cached copies.

Dynamic web pages: Contrary to static web pages that are readily formatted and stored on a storage device such as a hard disk, dynamic web pages are created only in the instant when they are needed. Dynamic web pages often draw their contents from on-line databases and they offer the flexibility to take information into account that is only available at the time when the page is requested. Information that governs or influences the contents of dynamic web pages can, for example, be collected via cookies or via online forms.

Excalibur web search is provided by Convera Corporation, a leader in enterprise search and categorization. The Excalibur web search project is focused on indexing and organizing the World Wide Web into millions of distinct categories. Combining semantic analysis, the largest index of categories in the world and sophisticated authoritative ranking to understand the unique meaning of each web page, Excalibur promises to deliver results with high precision.

ETDE: The Energy Technology Data Exchange (http://www.etde.org/) is an international consortium that collects and exchanges energy research and technology information through the Energy Database. This international energy information exchange agreement was formed in 1987 under the aegis of International Energy Agency (IEA).

Fully qualified URL: A fully qualified URL is a URL that contains the complete Internet address of a resource. It is distinct from a relative URL that only gives the location of a resource relative to its own location.

Hard copy: Traditionally, a hard copy of a document meant its representation in the form of printed text on paper. We extend this definition to any hardware bound distribution of texts or multimedia content, where the hardware is intended to be a dedicated carrier of such information resources. In particular, this includes the distribution on CD-ROM or on DVD. A temporary copy of an information resource onto a digital medium, like a computer hard disk, is not a hard copy.

HTML: Hyper Text Markup Language (HTML) is designed for publishing information online. It enables browsing and addressing online information resources.

Internet: The Internet, or simply the Net, is the publicly accessible worldwide system of interconnected computer networks that transmit data by packet switching using a standardized Internet Protocol (IP). It is made up of thousands of smaller commercial, academic, domestic, and government networks. It carries various information and services, such as electronic mail, online chat, and the interlinked web pages and other documents of the World Wide Web.

Java-Script: JavaScript is the name of Netscape Communications Corporation's implementation of ECMA Script. One major use of web-based JavaScript is to write functions that are embedded in or included from HTML pages to perform tasks that are not possible in HTML alone, but is also used to enable scripting access to objects embedded in other applications.

Macromedia Flash: Since its introduction in 1996, Flash technology has become a popular method for adding animation and interactivity to web pages. The Flash files, traditionally called "flash movies", usually have a .SWF file extension and may appear as an element of a web page or to be "played" in the standalone Flash Player.

Metadata: literally "data about data", are structured, encoded data that describe characteristics of information-bearing entities to aid in the identification, discovery, assessment, and management of the described entities.

Meta search engines: query multiple web search engines simultaneously. Web site results that are returned from all engines are merged into one list of results. Depending on which meta engine has been used, the results may be ordered by relevance and pages from the same site may be clustered together.

NuArch: The aim of the IAEA's NuArch (Nuclear Archive) pilot project is to establish and test the methodology and all relevant procedures for the creation of an Electronic Archive on the example of 20 selected web sites and to produce a fully functioning prototype.

Open directory project: The Open Directory Project (http://dmoz.org/), also known as DMoz (from Directory.Mozilla.org, the original domain name), is a multilingual open content directory of World Wide Web links owned by America Online that is constructed and maintained by a community of volunteer editors.

Open source movement: The open source movement is an offshoot of the free software movement that advocates open-source software as an alternative label for free software, primarily on pragmatic rather than philosophical grounds. The early period of the open-source movement coincided with and partly drove the dot-com boom of 1998-2000, and saw a large growth in the popularity of Linux and the formation of many "open-source-friendly" companies.

PDF: Portable Document Format (PDF) is a proprietary file format developed by Adobe Systems for representing two dimensional documents in a device independent and resolution independent format. Each PDF file encapsulates a complete description of a 2D document (and, with the advent of Acrobat 3D, embedded 3D documents) that includes the text, fonts, images, and 2D vector graphics that compose the document. Importantly, PDF files don't encode information that is specific to the application software, hardware, or operating system used to create or view the document. This feature ensures that a valid PDF will render exactly

the same regardless of its origin or destination. PDF is also an open standard in the sense that anyone may create applications that read and write PDF files without having to pay royalties to Adobe Systems.

Regular expression: A regular expression is a string that describes or matches a set of strings, according to certain syntax rules. Regular expressions are used by many text editors and utilities to search and manipulate bodies of text based on certain patterns. Many programming languages support regular expressions for string manipulation.

Search engine: A search engine is a program designed to help find information stored on a computer system such as the World Wide Web, inside a corporate or proprietary network or a personal computer. The search engine allows users to ask for content meeting specific criteria (typically those containing a given word or phrase) and retrieves a list of references that match those criteria.

Seed-URL: A seed-URL is a starting point for a web crawler. Links found in documents that are located at seed-URLs are processed through pre-determined filters and those URLs that pass the filtering become in turn the seed URLs for the next step of the crawler.

Shallow web: These are ordinary web pages that are served out of database systems such as Cold Fusion or Lotus Domino. They are essentially normal, static pages and belong as part of the surface web. However, search engines are generally fearful of indexing these because its easy for them to accidentally index the same page over and over, because the URL might be slightly different due to different features of the dynamic delivery system.

Soft copy: The user copies entire documents, or parts of them, from an Internet location onto the local computer in order to read them. The copy is in most cases discarded after reading, or at least it is not archived in an organized way.

Web crawler: A web crawler (also known as a web spider or web robot) is a program which browses the World Wide Web in a methodical, automated manner. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. Crawlers can also be used for automating maintenance tasks on a web site, such as checking links.

Web harvesting (also known as web farming, web mining and web scraping) is the process of gathering and organizing unstructured information from pages and data on the World Wide Web.

Web server: A computer that is responsible for accepting HTTP requests from clients, which are known as web browsers, and serving them web pages, which are usually HTML documents and linked objects. Can also refer to a computer program that provides the functionality described in the first sense of the term.

World Wide Web ("WWW" or simply the "'web"): A global information space from which people can read and write-to via a large number of different Internet-connected devices. The term is often mistakenly used as a synonym for the Internet itself, but the web is actually a service that operates over the Internet, just like e-mail. The WWW is the complete set of documents residing on all Internet servers that use the HTTP protocol, accessible to users via a simple point-and-click system.

CONTRIBUTORS TO DRAFTING AND REVIEW

Badulescu, A.	International Atomic Energy Agency
Bachmann, H.	Convera GmbH, Austria
Firbas, P.	International Atomic Energy Agency
Gritsevskyi, A.	International Atomic Energy Agency
Krieger-Levine, C.	International Atomic Energy Agency
Lepingwell, J.W.R.	International Atomic Energy Agency
Mandl, W.	International Atomic Energy Agency
Mueller Pathle, A.	Convera GmbH, Switzerland
Stanculescu, A.	International Atomic Energy Agency
Tolstenkov, A.	International Atomic Energy Agency
Weber, M.	Xinexus Nuclear AG, Switzerland
Yanev, Y.Y.	International Atomic Energy Agency

INTERNATIONAL ATOMIC ENERGY AGENCY
VIENNA
ISBN 978–92–0–111207–1
ISSN 1995–7807