EVA

The Acquisition and Archiving of Electronic Network Publications In Finland

Kirsti Lounamaa
CSC - Center for Scientific Computing
FUNET network services
Kirsti.Lounamaa@csc.fi

Inkeri Salonharju
Helsinki University Library - The National Library of Finland
Inkeri.Salonharju@helsinki.fi

Abstract

This paper presents the current state of the effort to capture and preserve electronic documents published in the Finnish Internet. We describe the policy used to select the material to be included in the collection, the capturing process, the storage technology and accessing method. Some advanced tools to extend the access to the documents as well as the results and some statistics from the first collection are also presented. Electronic documents are not yet as a part of the Legal Deposit in Finland, but the revision of the Act is going on and the aim is to preserve also electronic publications for the coming generations.

Background

Helsinki University Library - The National Library - is co-ordinating the project EVA, which is a joint activity by libraries, publishers and expert organisations, being a part of the Information Society Strategy Program by the Finnish Ministry of Education. The main aim of the project is to test methods of capturing, registration, preserving and providing access to the on-line documents published by the established publishers or freely in the Finnish Internet. In this presentation we mainly concentrate on the latter type.

EVA is a cross road of several other development projects on the Nordic basis and serves as a test bed for new tools. The Dublin Core metadata template and converter, the URN generator and the harvesting and indexing application NWA (Nordic Web Archive) are tested in EVA. The design and specifications of NWA are done in co-operation with the Royal Library of Sweden and the CSC/FUNET - Center for Scientific Computing/Finnish University Network - is responsible for the implementation of NWA in Finland.

The advanced tools tested in EVA

Metadata template. The Dublin Core metadata template was originally created by Lund university library NetLab unit with the Nordic Metadata project. The template builds all the required HTML syntax automatically and allows the user to concentrate on creation of the content. The Perl script that creates the template is available for free and has become very frequently used world wide. There is already a significant number of documents in the Web which contain their own description in the Dublin Core. We realised early on, that we needed to offer an easy to use and fast template to encourage writers to produce metadata themselves. The template is an electronic form which, after filled in, immediately returns data parsed as a Dublin Core record.
(http://linnea.helsinki.fi/cgi-bin/dc.pl)

The URN generator which can build Uniform Resource Names based on National Bibliography Numbers, NBNs. The syntax of the produced URNs is authorised by the IETF URN WG and the numbering schemes are assigned by the Finnish and Swedish National Libraries. URN identifiers are persistent and unique: the URN given to a document will

never change, if the intellectual content remains the same. The browsers are at the moment not capable to find the document when a user puts the URN into the Location-box. In the future the Nordic Web Index system's national full-text and metadata databases will offer the full and correct searchability of all URNs used in WWW documents published in Nordic countries.
(http://linnea.helsinki.fi/cgi-bin/urn.pl)

The metadata harvesting and indexing application. This is in practice an enhanced version of the Nordic Web Index which can extract metadata in Dublin Core format and other formats from HTML documents and make this information searchable via metadata databases.
(http://nwi2.funet.fi/)

The Dublin Core to MARC converter, which can extract Dublin Core data from a document and convert it into a MARC record. From libraries point of view it is important to be able to utilise the Dublin Core records in the maintenance of the National Bibliography database and in the library OPACs.
(http://www.bibsys.no/meta/d2m/)

On-line documents with limited access

Not all the electronic documents can be acquired using automatic methods. The library must use other methods for collecting restricted documents from commercial publishers. In the future deposit of on-line documents is defined on national level either in the Act of Legal Deposit or in voluntary agreements between the National Library and publishers. According to the proposal for the new Finnish deposit law publishers will be responsible for sending the on-line deposit documents to the library.

The National Library is experimenting the transfer of the documents with different methods. At present, document delivery using ftp-protocol seems to be the most effective way to accomplish this. Situation may change rapidly when new Internet-tools come to the market. To prevent document spoofing during transmission, digital signatures are used. In addition to this purpose, a digital signature can be used for two other purposes, firstly for authentication of the sender, and secondly for authentication of the content.

In the future the National Library as a Legal Deposit will maintain a list of approved file formats. The organisation responsible for depositing the document (generally the publisher) should also be responsible for converting the document to an acceptable format. If it is not possible to convert the document to a suitable format with reasonable effort, it is not necessary to deposit it.

It is important for National Libraries to retain the original look and feel of the documents. Therefore the list of approved formats should be quite exhaustive and updated frequently. Long-time preservation is in these cases secured by in-house conversion of the document to "better" format, or by emulation of the original usage environment.

Selection criteria for harvesting freely available documents

The general design is that we don't try to make any specific selection but the collection will include all the freely available, published, static HTML-documents with their inline material like pictures, video and audio clips, applets etc. In our approach, 'published and freely' accessible means that the document is accessible by standard HTTP protocol, it is referenced in some other document and there is no fee or password required. Only static documents are captured leaving out programs, database searches and so on. Since we are interested only in documents published in Finland, we have limited the collection, so that we have included only Web sites with network address ending in '.fi'. We are well aware that there are a large number of Web servers which are located in Finland, but whose address ends in '.com', '.net' or something else. These are becoming more and more popular and we have to find a method to recognise them. In the Kulturarw3 project in Sweden this has been solved by asking InterNIC for the list of DNS entries whose owners have given an address in Sweden. This is probably a satisfactory solution for us too.

The Web consists of documents which are linked together with hyperlinks (addresses) embedded in the documents. The documents are identified by their location, so called URL (Uniform Resource Locator). The problem is that the only thing we know about a document is its URL. There is no relation between the URL and the contents of the document, which means that one document may have several locations over time and also that in one location there may be different documents at different times. We see that this problem can't be solved with the current technology before there is a way to identify the content of the document, not the document itself. A URN (Uniform Resource Name) is a unique identifier given by some authorities and it can be embedded in the document, so that it follows the document all its lifetime. Before the usage of URN is more popular, we have to capture the documents by URLs. The limitation is

that there is no concept of document version: a document in a location in one snapshot has possibly nothing to do with a document in the same location in the next snapshot.

Capturing

Regardless of the limitations, we try to create a series of snapshots of the contents of the Finnish Web that exists over time.

Technically the system is quite simple. We have written a harvester robot program whose task is to fetch Web pages. Once a page is captured, our software analyses its content for inline material and cross-references, i.e. hyperlinks. In our model all the inline material, for example pictures and video and audio clips, is considered as essential part of the document's content and has to be fetched regardless of its location. This means that this material is captured also outside of Finland. Where as cross-references are not part of the current document and they are only fetched if they reside on the Finnish Web.

The system consists of four parts: capturing, analysing, indexing and archiving. Every day a program starts which analyses all the documents that were captured the day before. As an output a list of URLs is generated. This list is sorted according to the capturing policy. The policy ensures that Web sites are checked often enough for new and modified documents. At the same it must ensure that the Web server's normal activity is not interfered and the network traffic is kept in minimum. The capturing process then starts reading and processing the list.

The capturing process, i.e. harvester robot, fetches the document and stores it in the local disk to an ordinary unix file system. To maintain the integrity of the snapshot, only the documents with successful HTTP return code (i.e. 200) are stored. The documents are stored together with HTTP headers *as is*, no modifications to their contents are made.

The documents are given a document-id. At the moment we are using the md5-checksum of the document as an identifier because it's guaranteed to be unique. In the implementation the document-id is also the name of a file in the local file system. This is also a way to prevent storing duplicates: if the document already exists, it is not stored again.

Storage

The captured documents are packed and compressed daily using TAR- and ZIP-utilities. This compressed TAR-file is then sent to the archive server using FTP. The file contains 10 000 - 50 000 documents and its size is usually 0.5 - 1 gigabytes. The reason why we have to pack the documents in big files is because at the moment the long-term, low-cost storage is based on tape technology that can't handle small files efficiently.

A concept of hierarchical storage is proven useful in several applications which store large amounts of data for long period of time. In hierarchical storage, there are several layers of storage, usually fast, expensive, small-capacity disk storage at the top and slow, inexpensive, large-capacity tape-drives at the bottom. There may be several layers in between. A software, HSM - hierarchical storage manager - migrates unused data downwards. When the data is accessed the software migrates the data upwards to the fastest device. The system is transparent to the user, the software decides the correct placement of the resource on basis of its usage.

Our HSM software is UniTree running on a HP machine and the tape technology is StorageTek Timberline cartridge robot. The archive technology was selected simply because we already had it installed and it's been used for many years. No investments were needed. The tape technology is very conventional and we are quite confident that the data will be preserved for 15 years, what is guaranteed by the manufacturer. This gives us enough time to decide what to do with the archive and to which storage technology it will be migrated next. For safety the UniTree software writes two copies of files to separate tape pools of which the other is to be kept in the vault as a backup copy.

In this first version of the model we decided to store the document and its parts, i.e. inline material separately. This is probably not a good idea. When the document will be presented, it is difficult to maintain the integrity if all its parts are stored in different times in different files. For example, how to guarantee that the pictures are the right ones? A new approach is to change the capturing process so that all the inline material is captured immediately after the document. We also have to select an appropriate file format to store the document and its inline material in an aggregate. The idea is that everything that is needed to show the document in the browser as it was at the moment of capturing, is collected together and stored sequentially. A similar approach can be used as in RFC 2112, which describes how to encapsulate multimedia documents to be transferred in a single e-mail message. This is anyway something we a working on in the project in the near future. The drawback of this design is that it will make the archive much larger than today because

for example all the pictures will be stored in several copies. But since the data is stored in the low-cost, high-capacity tapes it is not a problem.

Statistics

We finished the first snapshot at the end of March 1998. Now we are taking the second snapshot. The first snapshot contains about 1.8 million documents from about 7500 Web sites in domain '.fi' . The majority of documents is text, 86 % of documents are of types html and plain text. About 10 % of documents are images, whose main types are giff, jpeg, tiff and x-bitmap. About 4 % of documents are applications, whose main types are tar, zip, octet-stream and pdf . The rest (less than 1 %) include audio and video and about 180 other types of document of which most are unknown to us. Main multimedia types are realaudio, wav, midi, mpeg, quicktime and ms-video.

About 50 000 documents are collected every day. If there are duplicates of documents already in the archive, they are dismissed. (About 12 % of documents are duplicates.)

Even if there are a lot of different document types found, we will be able to show most of them in the future because the most common file types represent over 90 % of all types. Even if we had to throw away the rest, the snapshot is representative.

Other thing that justifies taking of snapshots is that the documents are not modified too often. The average time between modifications is 200-300 days. Of course this reflects our selection policy, only static documents are collected. But it can also be seen that it is normal practice that documents stay untouched for long period of time after they are published.

The size of documents on the other hand is growing rapidly. On March 1998 we calculated that the average size of a Web document is 15 kilobytes, now it is over 20 kilobytes. This obviously has to do with the increasing usage of multimedia features. This might cause problem in the future. At the moment the estimated growth of the archive is 0.5 terabyte per year. If it grows much faster we might have to change the selection policy towards more strict rules.

Access

Since FUNET runs a Finnish service point of NWI - Nordic Web Index, we have combined the archiving and indexing processes. All the captured documents are indexed using the NWI-profile. The NWI database is accessible by Z39.50 (host: nwi.funet.fi, port:2100) and it contains records of all the online Web-documents. The NWI-profile contains:

- Date the resource was checked
- Date the resource was last modified
- Content type
- Content size
- MD5 checksum
- Availability: URL to the on-line document
- Title
- Headers
- Sample text (about 20 % of the full text)
- Cross-references: linkage and title

We have also started to collect a database FinMeta, which contains all the Dublin Core descriptions that are found in the Finnish Web. There are currently about 1000 records in the database.

The methods for providing access to the archived material will be studied, specified and tested later this year within the NEDLIB project.

References

EVA - Electronic Virtual Archive: http://linnea.helsinki.fi/eva/
The Kulturarw3 Heritage Project: http://kulturarw3.kb.se/html/projectdescription.html
NEDLIB- Networked European Deposit Library: http://www.konbib.nl/nedlib/
NWI - Nordic Web Index: http://nwi.funet.fi/