Harvester results in a digital preservation system

Tobias Steinke

Deutsche Nationalbibliothek
Adickessallee 1
60322 Frankfurt
Germany
t.steinke@d-nb.de

Abstract

In the last few years libraries from all around the world have build up OAIS compliant archival systems. The information packages in these systems are often based on METS and the contents are mainly e-journals and scientific publications. On the other hand Web archiving is becoming more and more important for libraries. Most of the member institutions of the International Internet Preservation Consortium (IIPC) use the software Heritrix to harvest selected Web pages or complete domains. The results are stored in the container format ARC or the successor WARC. The files' quantity and the sizes of these archival packages are significantly different than those of the other publications in the existing archiving systems. This challenges the way the archival packages are defined and handled in current OAIS compliant systems.

This paper compares existing approaches to use METS and Web harvesting results in archival systems. It describes the advantages and disadvantages of treating Web harvests in the same way as other digital publications in dedicated preservation systems. Containers based on METS are set side by side with WARC and its possibilities.

Background: Preservation systems and Web archiving

In the last few years cultural heritage institutions like national libraries began to build up dedicated archival systems for digital preservation. Coming from the traditional collection of books and journals the focus was on similar digital entities like e-theses, e-journals and digitized books. These items can be in a variety of file formats and quantities but each single object is clearly defined and contains seldom more than a few hundred files. Nearly all of the archival systems are more or less designed according to the OAIS reference model, which identifies components and tasks of such a system. To fulfill the task of preservation it is necessary to ensure access to the content of the objects even when software and hardware will change completely. In the OAIS model the needed activities are called *Preservation Planning*. Current implementations try to do this basically by the strategies migration and emulation. The basis for both strategies is supporting metadata especially about the technical aspects of each archived object and file.

On the other hand cultural heritage institutions have to face a completely new challenge: The collection and archiving of Web pages. Depending on the institution and existing legal deposits, this could include certain sub domains, pages related to a specific topic or a complete top-level domain like .fr. The common way to collect the pages is to use software called harvester. This automatic program gets an address to start with and then follows every link on each page within given parameters. The result is either saved in separate files according to the original file formats (HTML, JPEG, etc.) or in one aggregated file. One of the most commonly used harvesters is called Heritrix. It saves the results in a aggregated format called WARC. WARC is an ISO draft which contains the files itself and metadata about the harvest activity.

As the process of collecting the Web pages and giving access to them is a challenging process for itself, the actual storage is currently often done without the same requirements for preservation as for other digital objects. Existing archival systems for digital preservation have often not been designed to deal with the complexity of Web pages. Strategies for preservation may be difficult to accomplish on the scale of Web harvester results.

Rebecca Guenther and Leslie Myrick wrote an article in 2006 about the way Web harvester results could be handled as archival packages with the metadata standards METS and MODS [1]. Since then the WARC format became relevant as a more advanced format for Web harvester packages including metadata and on the other hand dedicated archival systems for digital preservation - like the one developed in the German project kopal - became more sophisticated.

Preservation systems and the object model

The ISO standard "Reference Model for an Open Archival Information System (OAIS)" [2] describes an abstract model of an archival system dedicated to long term preservation. This reference model and especially its functional model define the functional entities and terms commonly used in all developments of digital preservation systems. The objects in the OAIS model are called Submission Information Package (SIP) at the moment of ingest, Archival Information Package (AIP) within the archival storage and Dissemination Information Package (DIP) for the access. Each Information Package is a conceptual container of content information and Preservation Description Information (PDI). The OAIS model does not define or restrict what the content information actually is.

On of the first implementations based on the OAIS reference model was the e-Depot of the National Library of the Netherlands [3]. It was conceived for digital publications, which are mostly PDF files. Therefore the object model was suited to handle single files and low complexity objects.

The German project kopal and the Universal Object Model

The German project "kopal: Co-operative Development of a Long-Term Digital Information Archive" (2004 - 2007) [4] used the same core system (DIAS by IBM) as the e-Depot, but enhanced it with a new object model to enable more complex objects and support the preservation strategy of file format migration. Although the object model was conceived to be able to handle all kinds of file formats and objects with hundreds of files, the focus was still on digital publications by commercial publishers, scientific publications and digitized books.

The object model defined for the kopal project is called Universal Object Model (UOF) [5]. The idea of the UOF is to define an information package, which should contain all files of one logical unit (e.g. a book, a thesis, an article) and all necessary metadata to enable preservation actions like migration. Descriptive metadata could be part of the package, but only the preservation metadata is mandatory. The UOF should also be self-sufficient in a way, that is to be suitable to enable exchange of objects between different archival systems. The package itself is a file container (ZIP or a similar format) and a XML metadata file conforming to the Metadata Encoding & Transmission Standard (METS) [6].

METS is a widely used standard to encode different metadata information and structural information about a digital object in a XML file. It is very generic in a way that there is no restriction on the kind of metadata to be included. Therefore the concept of profiling was established to define restrictions for specific use cases. Rebecca Guenther and Leslie Myrick describe in their article [1] the METS profiles of the project MINERVA which includes descriptive metadata in the format MODS and hierarchal structural information. The METS profile for the UOF demands preservation metadata in the format LMER [7] but allows all kinds of descriptive metadata. All files of the package must be listed in the METS file and there should be a record of technical information included for every file. Structural information could be of any complexity, but this should be restricted on files within the package.

Web archiving and harvester

Web pages are part of the cultural output of our society and therefore cultural heritage institutions feel the obligation to collect them like any other digital publications. But the structure of the Web is global and there are no clear national borders in the virtual space. The traditional collection policy of national libraries to collect everything from or related to their own country is difficult to apply to Web pages. This problem is addressed by restricting the collection to pages of a certain top-level domain (e.g. .de, .fr, .uk). As an alternative or in addition there could also be a selective approach to collect topic-related.

Another problem is the dynamic character of the Web. There is never a fixed or final state of a Web page. The content of a Web page could be changed at any point in time. The content could also be dynamic itself, computed at the time of access based on input by the user. As a result, collections of Web pages are always time specific snapshots of certain states. It is not possible to collect "The Web".

The actual collection is done by a harvester (a.k.a. crawler). Starting with a URL these programs follow each link on a page and save every file on their way. The International Internet Preservation Consortium (IIPC) [8] was founded by national libraries and the Internet Archive to collaborate on preserving the Internet content for future generations. Currently it consists of 38 member institutions from all over the world. One of the projects of the IIPC is the (further) development of the open source harvester Heritrix [9]. It uses very sophisticated methods to fetch as much content as possible. The result is stored in ARC files, which are containers for the collected files and the additional information about the harvesting itself.

The WARC format

The WARC format [10] was developed as a successor of the ARC format. It currently exists in a draft status and was submitted as an ISO standard. Every WARC file is a container of records. The records can contain the unchanged binary files of the page (e.g. HTML, JPEG, GIF), general information about the Web crawl, network protocol information, revisitation information (about changes since the last snapshot of the same pages), conversions (migrated file versions) and metadata about each file. The metadata could be WARC specific, Dublin Core or conforming to any other schema. Heritrix will generate one or more WARC files for each crawl depending on a configurable WARC file size.

Approaches to use METS in Web archiving

Most of the institutions which use Heritrix store the resulting ARCs in a file based system and use software like Wayback [11] to give their users access to the stored snapshots. The focus is on managing the harvesting process. Existing preservation systems are separated from these processes.

METS is widely used for SIPs in OAIS compliant archival systems. As the result of a harvester like Heritrix is already a container (ARC or WARC), the containers could be referenced in the METS files or each file in the containers could be referenced individually.

METS in the MINERVA project

The MINERVA project [12] at the Library of Congress (USA) established an archive of event-related collections of Web pages. Although this project was not primary about preservation, Rebecca Guenther and Leslie Myrick [1] described a concept of METS and preservation information for MINERVA. They argue that in order to handle the complexity of the Web material it is necessary to define two METS profiles: One to describe the levels of aggregation and one for every capture. The structural map of the aggregate-level METS files consists of pointers to lower-level METS objects. MODS is used in the METS file to describe the intellectual object on the aggregate-level. The METS files on the capture level includes MODS for page-specific content information, several metadata schemas for technical information on file level and PREMIS for preservation information. The Structural Map and Structural Link section of METS could be used to reflect the links on each HTML page.

METS in the Web Curator Tool project

The Web Curator Tool (WCT) project [13] is a collaborative effort by the National Library of New Zealand and the British Library, initiated by the IIPC. Its purpose is to manage the selective Web harvesting process. A SIP specification [14] was developed for the use case of submitting the results of a harvesting process to an archival system. The SIP contains all ARC files of a crawl, selected log and report files of Heritrix and a METS file. The ARC files and Heritrix files are referenced within the METS file. The Metadata in the METS file conforms to a specific WCT schema and includes information about the crawl, owner data, agency data, descriptive information and permission data. There is no list of the files within the ARC files or technical information about these files in the METS file. The Structural Map is just a plain list of the ARC files and the Heritrix files.

Preservation strategies and Web archiving

An archival system for digital preservation should be focused on ensuring the access to its content for the unpredictable future. Software and hardware will change and no file format will be supported forever. The two common strategies to face this challenge are migration and emulation. Migration is the conversion of file formats to currently accessible file formats. Emulation is the recreation of another system environment on a currently used system environment. For both strategies it is essential to record as much information as possible about the technical parameters of the archived objects. This is done by generating metadata and storing it together with the content files. METS could be used to build information packages of metadata and content files.

Migration of Web harvester results could be difficult to handle. One crawl can produce thousands of files. A lot of these files are HTML files with links to other files. In case of the migration of one format to another, not only all affected files have to be change but also all HTML files linking to these files. The approaches of the MINERVA project and the UOF enable the recording of technical information for every file and of dependencies between the files in a METS file. In principle this is a good basis for the migration task. But the practical problems of performing all necessary activities (conversions, checks, error corrections) for objects with thousands of files remain. It may also be technically challenging to generate the metadata and the resulting huge METS files on this scale. Migration on the basis of the WCT METS files might be impossible, because there is no information about the technical aspects of the single files within the ARC files. But this approach is helpful for migrations of the ARC files (e.g. to WARC files).

Emulation for Web harvester results could be an easier task than to emulate complete computer systems. Web pages are in principle designed to work on any Web browser of a certain time period. There are dependencies of certain media plug-ins, software specific restrictions and machine related parameters (performance, memory size) but these are harmless compared to the complexity of the emulation of a specific computer configuration. For the emulation approach it is important to know the time period of the crawl and the circumstances of the harvesting process. This is provided in a useful way by the WCT SIP specifications. The ARC files bundle the unchanged content files and the metadata and the reports give the needed information. The MINERVA METS files on the aggregate level would also provide the information. But it could be difficult to hand over all files of one crawl to the emulator. A few ARC files might be easier to handle than thousands of different files. The UOF was not yet used for Web harvester results. If the ARC files were chosen as content files and the technical metadata within the LMER sections described the crawl, the resulting UOF METS files would be similar to the WCT ones.

On the other hand the new WARC format already offers all needed information for the emulation and even a mechanism to store migrated file versions within the container. But WARC files need to be managed in an archival system and therefore a structural wrapper like METS could be helpful. The provided information within the WARC files could be easily extracted to build up METS files which could even support both preservation strategies similar.

Summary

Web archiving is a new challenge for the preservation community. Existing OAIS compliant archival systems use METS and preservation metadata to support preservation strategies like migration and emulation. These concepts could be used for Web archiving as well but a re-design or enhancement of the METS based object models might be necessary. The introduction of the WARC file format offers additional support for the new developments.

References

[1] Guenther, R., and Myrick, L. *Archiving Web Sites for Preservation and Access: MODS, METS and MINERVA*. Journal of Archival Organization 4, no. 1/2 (2006).

[2] ISO 14721:2003, CCSDS recommendation: http://public.ccsds.org/publications/archive/650x0b1.pdf

[3] http://www.kb.nl/dnp/e-depot/dias-en.html

[4] http://kopal.langzeitarchivierung.de/index.php.en

[5] Specifications of the Universal Object Model: http://kopal.langzeitarchivierung.de/downloads/kopal_Universal_Object_Format.pdf

[6] http://www.loc.gov/standards/mets/

[7] Specifications of LMER: http://nbn-resolving.de/?urn:nbn:de:1111-2005051906

[8] http://netpreserve.org/

[9] http://crawler.archive.org/

[10] Draft of the WARC specifications: http://archive-access.sourceforge.net/warc/

[11] http://archive-access.sourceforge.net/projects/wayback/

[12] http://lcweb2.loc.gov/diglib/lcwa/

[13] http://webcurator.sourceforge.net/

[14] http://webcurator.sourceforge.net/docs/development/WCT%20Project%20SIP%20Specification.doc