eChallenges e-2011 Conference Proceedings
Paul Cunningham and Miriam Cunningham (Eds)
IIMC International Information Management Corporation, 2011
ISBN: 978-1-905824-27-4

Web Archiving Using the Collaborative Archiving Services Testbed

Ingemar ANDERSSON[1], Lena LINDBÄCK[2], Göran LINDQVIST[2],
Jörgen NILSSON[1], Mari RUNARDOTTER[1]
[1]*Luleå University of Technology, 971 87 LULEÅ, Sweden*
Tel: +46 920 49 10 00, Fax: +46 920 49 13 99, Email: firstname.lastname@ltu.se
[2]*LDP Centre, Teknikvägen 3-13, 961 50 Boden, Sweden*
Tel: +46 921-573 00, Email: firstname.lastname@ldb-centrum.se

Abstract: Websites constitute one category of official records and as such should be preserved for the long term in compliance with Swedish legislation. Collaborative Archiving Services Testbed (CAST) supports actors involved in the selective web archiving process, from harvesting to the creation of an information package ready for transfer to a long-term archive at National Archives of Sweden. CAST is developed in compliance with the ISO standards Open Archival Information System (OAIS) and Producer-Archive Interface Methodology Abstract Standard (PAIMAS), and do also consider other well-known and established standards and recommendations in digital preservation area. CAST promotes cooperation, knowledge acquisition and sharing among users in an experimental step-by-step workflow, encouraging a proactive approach resulting in authority websites better adapted to digital preservation recommendations. CAST is developed at LDP Centre, a national competence centre in Sweden, within the digital preservation area.

1. Introduction

Digital objects are vulnerable to digital obsolescence over time because of hardware and software constantly evolving. Currently there is lack of established standards, protocols, proven tools and methods for preserving digital information in general [1]. In most cases National Archives restrict the formats they accept, demanding preservation activities before digital collections are acceptable for long-term preservation. Most of the preserved digital collections transferred to the National Archives come from government authorities and public sector, which is a group where digital preservation strategies often are lacking. These organisations are in need of support to develop methods and tools that provide aid in the process required preserving and transfer preserved collections to the National Archives in an orderly manner, as well as to cover pre-transfer preservation activities [2]. In compliance with Swedish legislation authorities official records are to be preserved for long-term. Initially these records are to be stored and managed in a local Archival Information System (AIS) by the authority that created the record, and with regularity, or at request transferred to the National Archives of Sweden (NA) for long-term preservation. Websites constitute one category of official records and as such continuously be preserved according to Swedish regulation. The web is also an important part of cultural heritage, and web archiving has become an interesting challenge for the e-society.

Website content characteristics make adequate web archiving a challenge with unique difficulties: linked objects, high dynamics, volatility, format variety, rapidly evolving technologies and huge number of content providers [3][4]. Other problems highlighted are that supplied materials do not match the archives' expectations due to unclear definition of what should have been delivered to the archive and delivery commitments that are difficult to fulfil for the information producer (e.g. authority) due to lack of support. This means

long and costly lead times in the harvesting and delivery process and reduced quality in archived collections, because errors in transmission are detected late or not detected before use [5].

Altogether this constitutes the background for a research and development project at the LDP Centre in Sweden [6]. Within the project a web-based e-service, Collaborative Archiving Services Testbed (CAST), was designed and implemented. An e-service with a purpose to provide a method and software support for authorities (information producers) preserving their websites to be stored in local AIS and assist in delivery of harvested web collections for long-term preservation at NA, making it easier to monitor and ensure that technical conditions and regulations are met.

CAST is designed and implemented with the aim of supporting Producer-Archive Interface Methodology Abstract Standard (PAIMAS) [5] defined by Consultative Committee for Space Data Systems (CCSDS). PAIMAS covers the first steps in pre-ingest and ingest processes defined by the Information and Functional Model in Open Archival Information System (OAIS) [7] Reference Model (ISO 14721:2003), also developed by CCSDS. OAIS serves as a framework for development of Archival Information Systems (AIS) covering all aspects of digital preservation objects lifecycle, from standardized procedures for reception, storage and management, and access.

In this paper we describe the characteristics of Producer-Archive interaction based on the PAIMAS framework and its impact on design and implementation of CAST, an e-service supportive in the process of preservation of websites to be submitted to local AIS at the authority and at request, transferred and approved for long-term preservation at NA. CAST contributes to research on selective web archiving, and the understanding of how to design and implement a support system for selection, harvesting, description, quality review, package creation, and submitting a website collection to a digital archive for long-term preservation. All technical implementation details are not included in this paper.

2. Objectives

The main objective of CAST is to provide technical assistance and method support for users participating in the process of long-term preservation of websites with the main sub-processes: harvesting, content validation (part of quality review), description (adding technical and context metadata), packaging, and submitting the package for preservation in an AIS. CAST encourages cooperation and promotes knowledge acquisition and sharing among users in an experimental step-by-step workflow. Another purpose is to support a proactive approach resulting in authority websites better adapted to digital preservation recommendations.

3. Methodology Used

The development process has taken place in collaboration with a group of eight experts consisting of personnel working with digital preservation at NA and National Library of Sweden (NL). Moreover, archivists at governmental authorities have participated, through an established Archival Network consisting of 30 archivists from universities. Hence, their reactions, thoughts and ideas have influenced the CAST design. The methods for user involvement have been workshops, interviews and expert reviews.

4. CAST - a Collaborative Web Archiving e-Service

The project started with an initial investigation about the current situation at governmental authorities, in order to find out how and what challenges are significant for people working with preservation of websites. The knowledge of how to do in practice is lacking at authorities today. There is need for support in every part of the web archiving process,

from specifying archiving purposes and boundaries, configuration of acquisition tools, checking behaviour and appearance of collected site compared to the original site, content analysis (file formats, crawler response and virus checks), creating metadata (technical and context) to storage and access. Also identified as important among our informants is the need for support in the process of verifying whether intended content actually is included in the collection, suggesting checklist and documentation support. When the website is harvested and validated according to regulations adapted to long-term preservation specified by receiving archive (e.g. local AIS at authority or at NA), preservation description information has to be added. This includes information such as provenance (source and processing history), technical and context metadata, references (identifiers) and fixity (checksums), which then is packaged together with the content and submitted according to receiving AIS requirements. On the receiving side, there is an expressed wish for support to make it easier to monitor and ensure that the basic technical conditions and regulations are met before transfer to the ingest procedure in receiving AIS.
CAST is addressing these challenges and also supporting users in defining a submission agreement and assists users on both sides in the Producer-Archive interaction to fulfil this agreement. The activities in the described process are multiple and varied and often require involvement of a group of professionals with expertise in IT skills, domain knowledge and preservation issues. CAST is designed to create a Submission Information Package (SIP) to be received by an AIS, developed in accordance with the OAIS [7] reference model.

One model defined in OAIS is the Information Model. An information model concept described in OAIS is the Information Package which could be defined as a container holding two types of information, Content Information and Preservation Description Information. Content Information is defined to be that information that is the original target of preservation, in this use case the output from the website crawling process. To create meaning about what is being preserved Preservation Description Information (provenance, context, reference, fixity) has to be added to the Information Package. The Preservation Description Information is information i.e. metadata describing the Content Information. Together these objects form an Information Package and with packaging information, this result in a Submission Information Package (SIP) to be transferred to an AIS. Other information packages defined in OAIS are the Archival Information Package (AIP) an information package used for preservation holding a complete set of Preservation Description Information for the Content Information, and the Dissemination Information Package (DIP) that includes part or all of one or more Archival Information Packages sent to a consumer by the AIS illustrated in Figure 1.

Another model described in OAIS is the Functional Model. Figure 1 show an overview of the Functional Model and CAST interaction with an AIS based on this model.

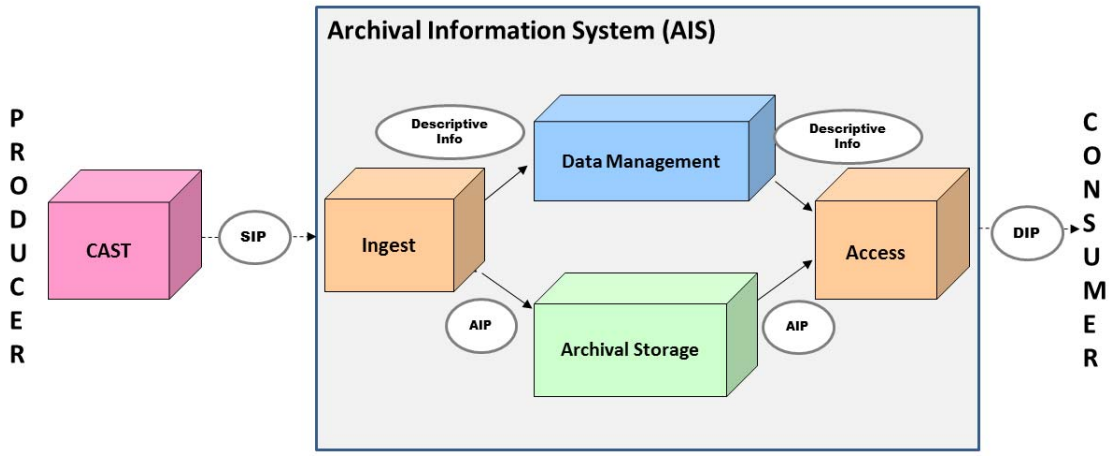

Figure 1: CAST – Interaction with an AIS

The Functional Model consists of different functional entities among them Ingest responsible for receiving and validate a SIP. A SIP forms the basis for all Information Packages handled by the AIS, and as such affects all functional entities in the model. A SIP created and delivered using CAST is aimed to be a good basis for creating the AIP and DIP, that only requiring minor adjustment of delivered SIP.

CAST consists of four sub-processes and starts {1} with harvesting content on the website according to constraints defined in submission agreement. Next CAST performs an analysis of the collected content {2}. In the third step, CAST creates collection arrangement, context, and metadata supplements according to metadata standards defined by receiving AIS, and validates content information and metadata prior to creating a SIP ready for transfer {3}. The fourth and final step is the transfer of SIP to the Archive {4}, which ends with a notification to the Producer whether the transfer was accepted or rejected. Figure 2 show a detailed interaction flowchart of CAST.

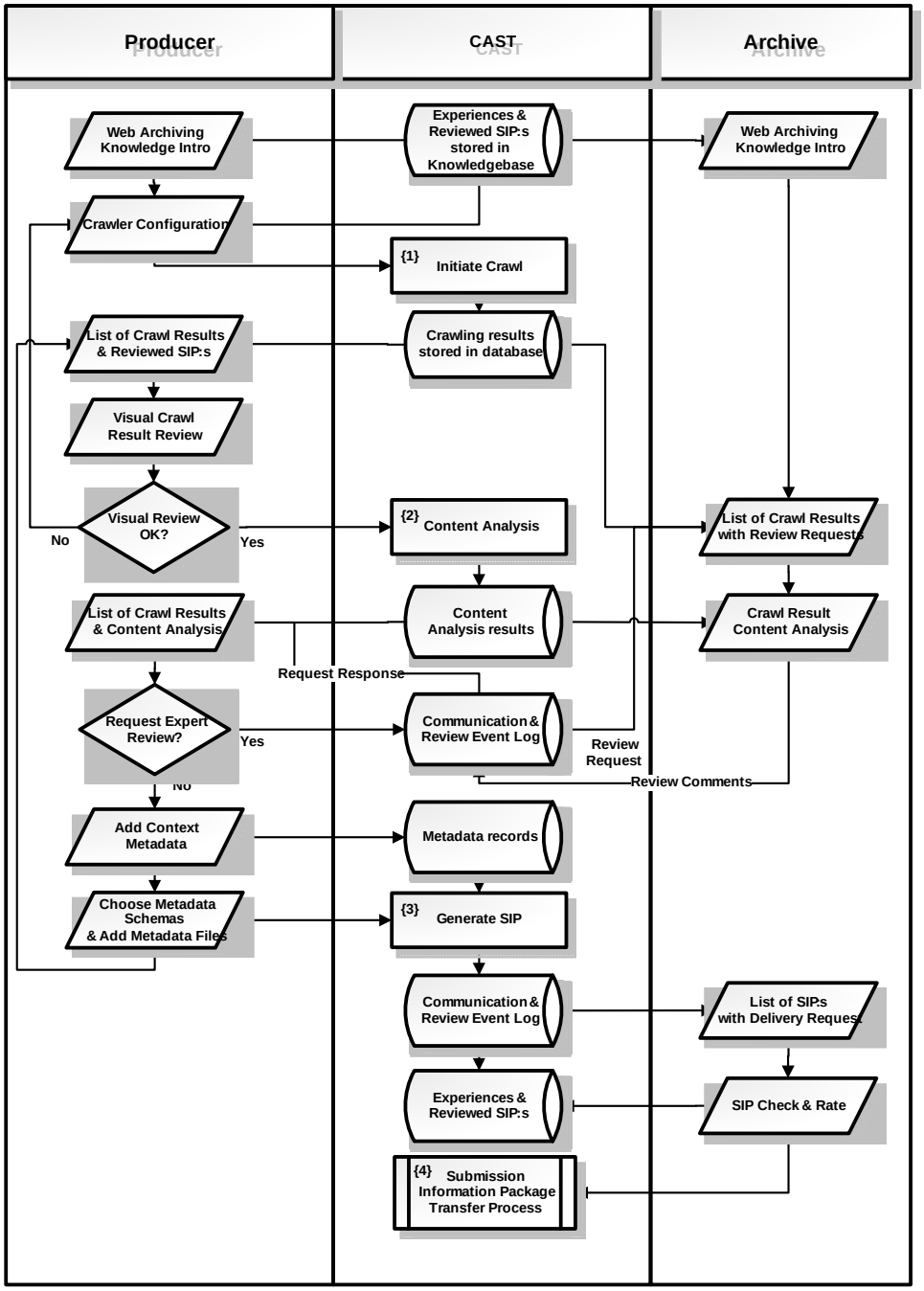

Figure 2: CAST - System Interaction Flowchart

"Web Archiving Knowledge Intro" is the starting point in CAST, introducing users to the web archiving area and the SIP creation process: it provides access to SIP reviews and CAST Knowledgebase. Next step is "Crawler Configuration" where users define URL seeds for the web crawler and select a crawler profile according to submission agreement. The step provides ability to configure crawler to avoid crawler traps, and exclude parts of site that are of no interest to collection or not suited for the harvesting technique in use. "Initiate Crawl" starts crawler (Heritrix[1]) job, generating WARC[2] output files. "List of Crawl Results" lists crawled collections with reviews, request/response status and rated SIP deliveries. "Visual Crawl Result Review" provides visual examination of collected website using Wayback[3] and a browser. "Content Analysis" is a technical metadata extraction procedure for WARC files using DROID[4], extracting data from crawling process logs, and virus-check. "Crawl Result Content Analysis" shows the compilation of detailed file format analysis process linked to PRONOM[5] technical registry. This step presents crawler process status codes, provides possibility to locate and compare specific content in the archived website with content on the live website. This is also a communication surface for content review and support. "Add Context Metadata" is input form for context metadata with information about the creators of the preserved collection and circumstances of record creation and use. "Choose Metadata Schemas and Add Metadata Files" specify metadata schemas to be used in the SIP creation process. Default schemas in use are METS[6] and PREMIS[7] with the ability to upload other metadata files to be added to SIP. "Generate SIP" procedures create an information object: content information with preservation description information, generates fixity checksums and sends a delivery notification request to the Archive. "List of SIPs with Delivery Request" displays a list of SIPs with request for delivery approval. "SIP Check & Rate" a visual control and rate process with accordance to the submission agreement. Sub-sequent activities depend on agreement policy and grading comments and are logged in the Knowledgebase. "Submission Information Package Transfer Process" generates a tar-package to be transferred to receiving AIS. Transmission responses are available to CAST through a web service-interface.

CAST is developed in conformance with the PAIMAS [5] framework in such a way that in the preliminary phase, an iterative and often time consuming process, where the negotiation on a draft of a submission agreement is accomplished, CAST offers, through the knowledgebase entity, documented experiences of all users exposed through process logs and SIP reviews. A list of contact persons with responsibility for specific topics is registered and available in knowledgebase during each step of the process. CAST offers a collection process with configuration and review support performed by experts at receiving archive giving producers advice on what is expected to be part of the digital collection and how to handle deviations from the agreement. CAST supports an automatic and manual inspection and rating process according to submission agreement before transfer to receiving archive: automatic validation according to metadata schemas, review of formats, and visual appearance. A reference to a format registry is available for every identified format version and due to the ability to provide detailed specification of file formats,

[1] Heritrix: http://crawler.archive.org/
[2] WARC (Web ARChive file format): http://bibnum.bnf.fr/WARC/index.html
[3] Wayback: http://archive-access.sourceforge.net/projects/wayback/
[4] DROID (Digital Record Object Identification): The National Archives of the United Kingdom http://droid.sourceforge.net/
[5] PRONOM UK National Archives Technical Registry: http://www.nationalarchives.gov.uk/PRONOM/Default.aspx
[6] METS (Metadata Encoding & Transmission Standard): http://www.loc.gov/standards/mets/
[7] PREMIS (PREservation Metadata:Implementation Strategies): http://www.loc.gov/standards/premis/

migration needs can be identified, giving the opportunity to correct the content information object in an early stage. CAST guides the user in adding necessary context metadata, while technical metadata and fixity is generated automatically, transfer responses from receiving AIS is communicated through a web service-interface.

The most time-consuming part of the CAST-process is the configuration of web crawler according to archiving purpose, web site structure, and content requiring the combined expertise of website domain knowledge, IT, and archiving.

The biggest technical challenges are the support of crawler configuration process: analysis of log files from the crawler process; the content analysis: often dealing with a couple of hundred thousand files for each website to be presented in a useful way; and the SIP-creation process: a flexible solution supporting different xml-schemas. These processes include a lot of text- and xml iteration algorithms. CAST also supports a web service interface between receiving AIS and CAST. The technical design follows a multi-tier architecture with the presentation, logical business processing, and data access separated in different layers.

5. Results

With the described implementation of CAST as a selective web archiving for long-term preservation our system, besides supporting the website harvesting process, covers the following aspects:

- Providing knowledge acquisition and sharing between users of the system lowers the threshold for introducing novice users to the area of digital preservation in general and web archiving in specific
- Supporting a comprehensive model of collaboration with well-defined tasks and responsibilities for each user in a quality review process
- Visualisation of a sophisticated content analysis with the ability to identify and locate the position and references to an information object in the collection
- High level of automation and flexibility in the creation of SIP: supporting preservation community standards that facilitate interaction with receiving AIS, through a flexible solution that in present version offers the creation of SIPs with metadata files based on following standards: METS, PREMIS, Dublin Core, and ADDML.
- Supporting the establishment of a submission agreement between producer and archive that contributes to a pro-active approach in daily operations of the website
- Simplified user interface to complex software and high level of automation for selective web archiving.

A short validation of PAIMAS [5] is that it's more concrete than OAIS [7], but it has a strict chronological order with a lot of steps (86) and it needed some restructuring to be useful in the development of CAST. We have used PAIMAS more as a checklist than a step-by-step construction plan.

A similar solution, to the one presented in this paper, is Hanzo Archives [8] which has a well-developed web crawler that offers more than a standard implementation of Heritrix crawler, including support for harvesting social media sites and links embedded in Flash etc. Another solution is the Web Curator Tool (WCT) [9] using Heritrix as web crawler offering a quality-review process where you can delete and add files in the harvested collection. Both of these solutions supports a workflow process with a primarily defined focus on the web harvesting process and have support for adding descriptive metadata, and access tools to browse the archived web sites.

In addition to CAST functionality described in this paper and in comparison with described comparable solutions (Hanzo, WCT), CAST focuses on offering users with

different skills (IT, website knowledge, archive) to interact supporting the Producer-Archive interaction based on the PAIMAS framework. CAST provides support for communication and knowledge sharing between users of the system in a testbed approach that brings guidance to users in a step-by-step workflow ending up with a review of SIP before transfer to an archive. Using CAST holds promises of high SIP-quality, meaning an information package in accordance with the requirements set by receiving archive and in accordance to long-term preservation recommendations. CAST also aims to contribute to increased understanding, within the organisation, of the demands placed on the website content and design, to be persistent in the long term.

We conclude this section by clarifying that CAST: is not a digital repository or archive (an external repository or archive is required for storage and preservation), is not a document or records management system, and is not an access tool (even though you can access your harvested content inside the system).

6. Business Benefits

Since many authorities lack of knowledge in web archiving there is most certainly a need for help in the Producer-Archive interaction and website archiving processes: harvesting, content analysis, packaging and transmission of collection to an archive.

Without a packaged solution as CAST, an interface and method support for non-technical users for the archiving of websites, may end up with a risk of not preserving websites at authorities.

Looking at the archives perspective a solution as CAST that supports a process with the opportunity to review (monitor and ensure that technical conditions and regulations in SIP are met) prior to transmission, the quality of archived material is increased and significantly reduced post-processing procedures are required.

With use of CAST the entire web archiving process, from initiation to ingest procedure in the archive system is calculated, in terms of time, to be shortened by several months.

The process supported in CAST is not only dedicated to preservation of web. CAST is a tool for knowledge acquisition, sharing and cooperation between users, resulting in significant time savings, higher quality and lower costs for each delivered SIP. A business model is under development as we foresee an e-service with great business potential.

Looking outside the legal requirement of preservation of the web in Sweden, corporations, government agencies and other institutions increasingly rely on web-based publications, using the web as the main channel to distribute information and communicate both externally and internally. The structure and organisation of website content on a disk does rarely reflect the structure and content of a website, relying on backups of databases and servers are for this reason not a viable option. To search for lost documents required for regulatory compliance or legal purposes is normally easier, and thus cheaper, retrieved from an archive than to go through large disk backups.

7. Conclusions

In this paper we have presented CAST, the Collaborative Archiving Services Testbed, supporting ideas in the PAIMAS [5] framework. We have learned that relationships and interactions between an information producer and an archive are critical for the quality of archived information, although rarely simple and easy.

Based on experiences from this project, we believe that the pre-ingest process is critical to a cost effective long-term preservation solution. If this function is utilised effectively, then it acts as a gatekeeper, ensuring that objects have consistency of structure, content and in compliance with long-term preservation recommendations. The normalisation of content

and creation of metadata should occur at the organisation transferring the objects, where the organisational knowledge and expertise are, not at the receiving archive.

We do not have enough basis for drawing conclusions about if the use of CAST results in websites better adapted to preservation recommendations.

An initial evaluation holds promises that CAST will work well as support in the process of achieving a submission agreement between the producer and archive, but that there maybe should be alternative routes in the system for deliveries when an agreement has been established. Another important aspect influencing the adoption of a solution like CAST is the support of a highly automated process with an interface supporting non-technical users interacting with complex software.

During autumn 2011, CAST will be further developed according to supplementary configurations, crawl profiles, metadata schemas etc. and test deliveries of preserved web collections to the AIS at National Archives of Sweden are scheduled.

Acknowledgements

CAST is a result from the R&D project Testplattform at the LDP Centre, Centre for Long-term Digital Preservation, in Sweden. The project was partly funded by the EU. The authors wish to acknowledge staff from the National Archives of Sweden, the National Library of Sweden and Luleå University of Technology, participants of the Swedish University Archival Network and e-Challenge reviewers.

References

[1] S. Ross, M. Hedstrom. Preservation research and sustainable digital libraries. Published online: 13 January 2005, Springer-Verlag 2005.

[2] P. Sinclair, J. Duckworth, L. Jardine, A. Keen, R. Sharpe, C. Billenness, A. Farquhar, J. Humphreys. Are you Ready? Assessing Whether Organisations are Prepared for Digital Preservation. The International Journal of Digital Curation, Issue 1, Volume 6, 2011.

[3] DPE: European Quarterly Preservation Digest. LiWA – Living Web Archives. Published online: 31st March 2008, Digital Preservation Europe.

[4] R. Sharp. Active Preservation of web sites. International Web Archiving Workshop, IWAW 2010.

[5] Consultative Committee for Space Data Systems. Recommendation for space data systems practices, Producer-Archive Interface Methodology Abstract Standard (PAIMAS). CCSDS 651.0-M-1, Magenta Book, May 2004, adopted as ISO 20652:2006 http://public.ccsds.org/publications/archive/651x0m1.pdf

[6] The LDP Centre, Sweden: http://www.ltu.se/org/srt/Centrumbildningar/Centrum-for-langsiktigt-digitalt-bevarande-LDB?l=en

[7] Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1, Blue Book, Issue 1, January, adopted as ISO 14721:2003 http://public.ccsds.org/publications/archive/650x0b1.pdf

[8] Hanzo Archives: http://www.hanzoarchives.com/solutions/web_archiving_overview

[9] Web Curator Tool: http://webcurator.sourceforge.net/