Ingest Workflow for Electronic Publications at the Swiss National Library

Hansueli Locher; Swiss National Library; Berne; Switzerland

Abstract

The report starts with an introduction to the project e-Helvetica at the Swiss National Library. The goal consists in undertaking organizational measures to collect, catalog, prepare and preserve electronic Helvetica over the long-term, while establishing an archival system for these publications. After that it describes the Ingest system that has been implemented to automate the process of receiving data and metadata, preparing and transforming it for long-term preservation and saving it in archival storage. Special attention is paid to the aspects of harvesting websites.

Introduction

The mission of the Swiss National Library (NL) is formulated in the Federal Act on the Swiss National Library, dated December 18, 1992. Article 2 describes the mission as follows: The National Library has the mission to collect, catalog, maintain and provide printed material **or information stored on other information media** that relates to Switzerland.

The mission includes all types of electronic publications (on data media such as CD-ROMs, floppy-disks, DVDs as well as on the Internet in online publications). There is considerable demand for solutions since electronic publications assume an increasingly important position within our society.

The Project e-Helvetica

The NL instituted the project e-Helvetica (http://www.e-helvetica.admin.ch) to create the prerequisites to meet the statutory mission in the area of electronic publications.

One the one hand, the project e-Helvetica has a goal of undertaking organizational measures to collect, catalog, prepare and preserve electronic Helvetica over the long-term, while establishing an archival system for these publications. The project is budgeted for a period of 10 years (2001-2010). At the end of the period, archiving of electronic publications is planned to become operational.

Pilot Projects

It is important for the NL to search for partnerships with all participants in the process as part of process definition. We are looking for cooperative agreements with the producers of electronic publications from the outset since there is no legal requirement in Switzerland to deposit publications at the National Library.

The NL is gaining initial, concrete experience within four pilot projects.

Pilot project online publications

Modalities for the automated delivery of electronic publications have been established with two publishing houses. The publishing houses deliver the metadata as well as the publications in the form of ZIP files which are uploaded to FTP servers belonging to the NL.

Web-archive Switzerland

Cantonal libraries report websites relevant to their cantons to the NL by means of a web form. The NL then collects and archives these websites on the Internet.

During the pilot phase the collecting of websites concentrates on three areas:

1. Administrative offices for cantons and municipalities,
2. One specific topic per canton as chosen by the participating cantonal library (e.g. the Matterhorn for the Canton of Wallis),
3. An area of interest common to all cantons (environmental protection).

e-Diss.ch

University libraries have two means to transfer dissertations to the NL during the pilot phase. Two smaller institutions participating in the project report their dissertations and habilitations on web-forms similar to the one used for web-archive Switzerland. The NL can also retrieve the metadata via OAI-PMH (http://www.openarchives.org/OAI/) at two large university libraries.

Dissertations and habilitations are available as PDF files and are retrieved in a follow-up step during the workflow at the internet-link provided in the metadata.

Electronic official publications

The NL is reviewing as part of the latest pilot project how it can retrieve publications from official offices, above all, at the federal level and process them for archiving. Concrete results are not yet available. One office formulated an additional requirement to archive the digital signature used that ensures the legal validity of its publications.

Ingest

The NL is relying on the OAIS [1] reference model in its efforts to establish a system for long-term archiving. As part of system implementation, significant weight is placed on automating the processes as much as possible for workflows required as part of long-term archiving. The NL is attempting to use available tools for individual process steps as well as existing standards.

Development of an archiving system is driven analogously to the modules of the OAIS model for the various phases.

Archival Storage

In an initial phase, the NL, together with the Swiss Federal Archives, procured a storage system for preserving electronic publications. Two ADIC tape libraries together with a Hierarchical Storage Management (HSM) operated by StorNext won the bid as part of a WTO public procurement process. The tape libraries will be replaced this year by a redundant Network Attached Storage Systems (NAS).

The Ingest Process

The procurement of the Ingest system also took place as a WTO public procurement. The offer from Elca won the bid by creating a system that integrates existing tools into the process and demonstrates the flexibility required for further development. Of course, data management is also being established together with Ingest. The Ingest system at the NL was commissioned on February 1, 2007. The system processes form an essential prerequisite for the long-term archiving of electronic publications at the NL.

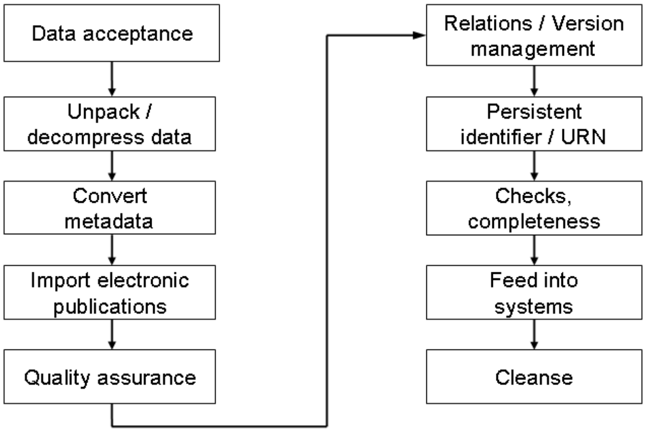


Figure 1. *The Ingest process*

Data acceptance

The Ingest system provides various interfaces for handing over data from producer systems such as FTP, web forms, and OAI-PMH harvesting. The data format provided by the metadata is agreed on with each data provider. It is important for the NL to search for partnerships with all participants in the process as part of the workflow. We are looking for cooperative agreements with the producers of electronic publications from the outset since there is no legal requirement in Switzerland to deposit publications at the National Library. The Ingest process periodically reviews the interfaces to producers and imports the available data or metadata in the process.

Unpack / decompress data

The files are unzipped if supplied as ZIP files.

Convert metadata

The data format provided by the metadata is agreed upon with each data provider. Normally, the metadata is imported in the form used for the producer's pre-print process. As a rule, this is generally XML or SGML files that are converted to the NL internal data format via XSL transformations. The NL internal format essentially consists of a METS container (http://www.loc.gov/standards/mets/), which includes bibliographical metadata in the MARCXML format (http://www.loc.gov/standards/marcxml/). PRESMET (Preservation Metadata) [2] from the National Library of New Zealand is used for technical and administrative metadata.

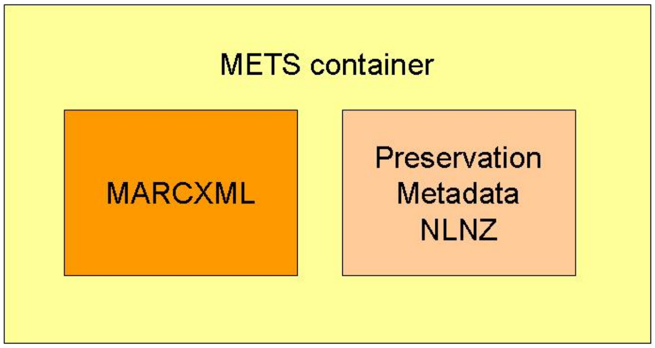


Figure 2. *Metadata format at the NL*

Import electronic publications

Often only metadata is supplied for data import, the initial process step for Ingest. In other words, the Ingest system must subsequently fetch the electronic publications under the link recorded in the metadata. Individual PDF documents as well as entire websites are collected under this process step. The chapter "Harvesting" provides a more detailed description of the procedure.

Quality assurance

Quality assurance steps are undertaken at various stages. Among the items checked are viruses, completeness of the delivered data packages, file formats, compliance of the metadata with the agreed upon metadata schema and the duplicate delivery of data packages. JHOVE (http://hul.harvard.edu/jhove/index.html) is used to check the file formats.

Securing rights is also a part of quality assurance. This includes the right to archive the digital publication in question. It encompasses both the permission to create the number of copies the NL deems is necessary as well as future preservation measures, e.g. executing migrations. Furthermore, the NL agrees to access rights to the archived data with the producer ranging from a prohibition on access over the next few years to being freely available.

Relations / Version management

The hierarchical relations between archive packages are recorded in Ingest analogous to the library catalog. For example, periodicals have an archive package for the publication title and additional packages for each individual issue. Planned, but not yet implemented, is the assignment of individual articles to a serial number. Articles are currently not archived as separate packages, but rather all articles for a number are compiled in an archival information package.

A title entry is also generated for websites collected on a periodic basis. A snapshot of the corresponding website is assigned to this entry.

Persistent Identifier / URN

Each archive package is assigned a unique identification in the form of a Uniform Resource Name as a National Bibliography Number (URN:NBN). It is registered via email with an XML attachment to the resolving server of the German National Library together with the associated URL (Uniform Resource Locator). The Ingest system initiates the next process step only after the registered URN can be queried on the resolving server.

The NL assigns the URN based on a sequential number. An internal identification is used for archive packages, not made available via the Internet that is setup like a URN.

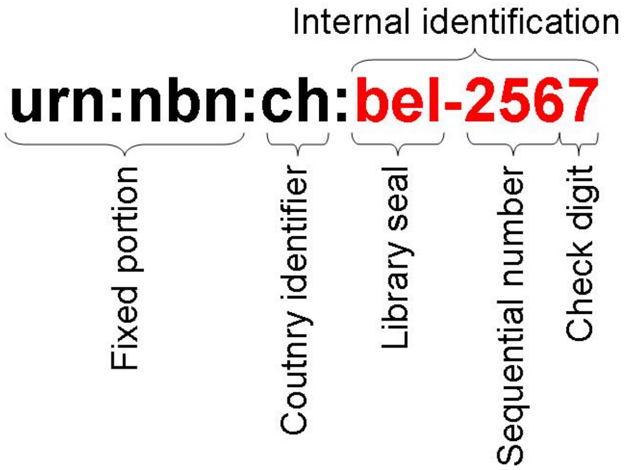


Figure 3. *URN:NBN setup*

For example, a complete URN would be urn:nbn.ch:bel-2567, whereas bel-2567 is used for the internal identification. This makes it easy to assign a URN after the fact, for archive packages that are opened to the public after some time, to simple to add "urn:nbn:ch" to the internal identification and registering it at the resolving server for the German National Library.

Checks, completeness

As part of a final check, the system automatically checks whether all the information acquired as part of processing is recorded in the metadata and ensures completeness of the data.

Feed into systems

Systems closely related to the Ingest are supplied with information as part of this process step.

All the metadata in XML format is stored in data management.

The Archival Information Package, containing both the actual digital publications as well as the metadata is fed to long-term storage in the form of a tarball.

The bibliographic metadata is converted from MARCXML to MARC21 and transferred in this form to the Helveticat library catalog, which for its part automatically adds it to the catalog.

Cleanse

Finally, under cleansing, superfluous processing data derived as part of processing to an Archival Information Package is deleted from data management.

System Environment

The Ingest system must fit into a previously existing system environment. Numerous interfaces to other systems must be considered. Figure 4 provides an overview of this system environment.

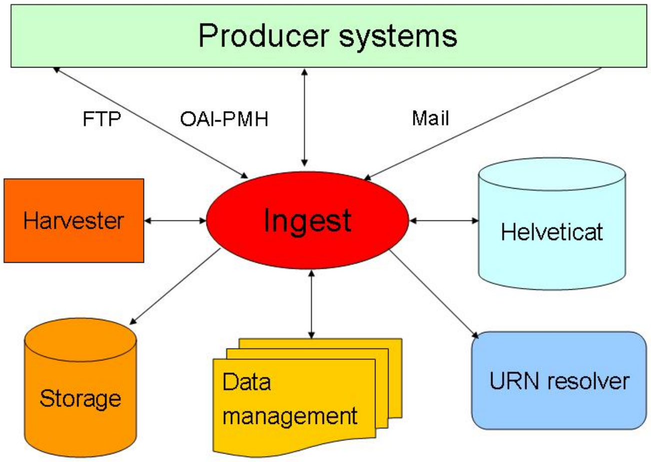


Figure 4. *System environment*

Storage

Long-term storage is located in a network area secured by an additional firewall. Archival Information Packages are transmitted via scp to the storage system.

Data management

Metadata arising as part of the Ingest process and all metadata previously supplied by the producer are stored in data management. This consists of an Oracle database accessed via JDBC.

URN resolver at the German National Library

The URN of long-term archived publications is transmitted to the URN resolver at the German National Library in Frankfurt together with the associated URL (http://www.persistent-identifier.de/?lang=en). It is responsible for the correct issuance of the URN.

Helveticat

All electronic publications processed in the Ingest system are recorded in Helveticat, the library catalog for the NL (http://www.helveticat.ch).

Producer systems

The Ingest process provides various interfaces for the handover of data from producer systems.

- **FTP and sFTP:** There are two different ways of receiving data packages: One feeds the data packages per FTP via a server at the NL into the Ingest process. Producers can also grant the NL access via sFTP to their own servers. The NL then retrieves the new data on a periodic basis.

- **Mail:** Data suppliers can also register the metadata for publications intended for long-term archiving via a web form. The registration arrives as a mail with an XML attachment to certain mail boxes. The Ingest process has access to these mail boxes.
- **OAI-PMH:** Larger university libraries allow for OAI-PMH harvesting of their library catalogs. This is how the NL acquires the metadata for newly recorded dissertations and habilitations.
- **Harvester:** A harvesting farm collects websites and individual documents on the Internet (see below).

Harvesting

The Swiss National Library currently limits the collection of websites to selective harvesting. Harvesting of all the .ch- domains should not be excluded as a possibility; capacity is just not available at this time.

Design and Function

Collection of electronic online publications was originally intended to be implemented directly on the Ingest server. Experience gained as the project progressed suggested, however, that security requirements for the network area, where the server was located, were too high to allow for trouble-free import of the required data from the Internet. As a result, harvesting was outsourced to a Demilitarized Zone (DMZ). An additional five servers were installed there.

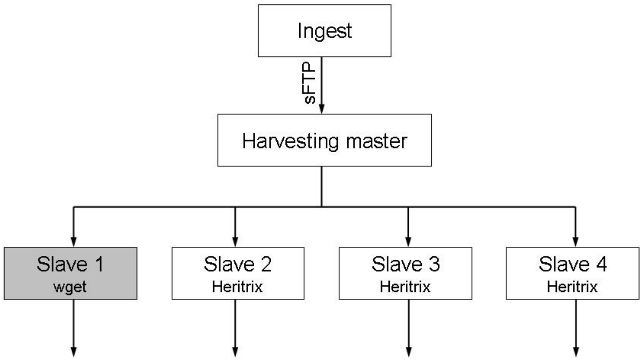


***Figure 5.** Harvesting configuration*

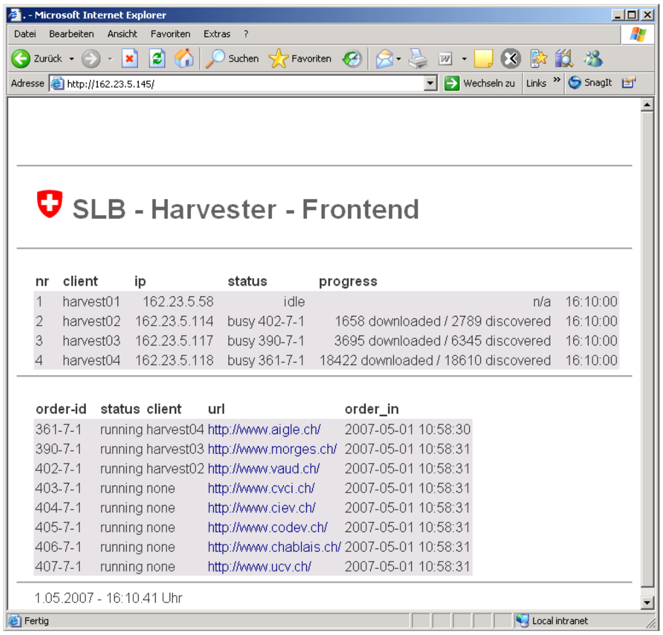
The Ingest server resolves harvesting by transmitting XML files via sFTP to the harvesting master. Periodic harvesting requests are administered on the Ingest server and XML files are once again generated at the appointed time for the harvesting master.

The harvesting master accepts the requests from the Ingest server and distributes them to the four slaves. All five devices participating in harvesting are synchronized using a Network Time Protocol (NTP) so that processes do not hinder each other at startup. The four slaves do not all have the same functions. Three are loaded with Heritrix 1.6 and collect websites.

A specially configured additional slave assumes retrieving individual files, for example, dissertations using Heritrix, since experience indicates that it does not always work without problems. The server collects individual documents with the help of wget (http://www.gnu.org/software/wget/).

Information on on-going harvesting processes and error messages are also provided by the harvesting master as an XML file in a folder. The files are retrieved from there by the Ingest system via sFTP.

You can further monitor the harvesting process directly via a web-based user interface.



SLB - Harvester - Frontend

nr	client	ip	status	progress	
1	harvest01	162.23.5.58	idle	n/a	16:10:00
2	harvest02	162.23.5.114	busy 402-7-1	1658 downloaded / 2789 discovered	16:10:00
3	harvest03	162.23.5.117	busy 390-7-1	3695 downloaded / 6345 discovered	16:10:00
4	harvest04	162.23.5.118	busy 361-7-1	18422 downloaded / 18610 discovered	16:10:00

order-id	status	client	url	order_in
361-7-1	running	harvest04	http://www.aigle.ch/	2007-05-01 10:58:30
390-7-1	running	harvest03	http://www.morges.ch/	2007-05-01 10:58:31
402-7-1	running	harvest02	http://www.vaud.ch/	2007-05-01 10:58:31
403-7-1	running	none	http://www.cvci.ch/	2007-05-01 10:58:31
404-7-1	running	none	http://www.ciev.ch/	2007-05-01 10:58:31
405-7-1	running	none	http://www.codev.ch/	2007-05-01 10:58:31
406-7-1	running	none	http://www.chablais.ch/	2007-05-01 10:58:31
407-7-1	running	none	http://www.ucv.ch/	2007-05-01 10:58:31

1.05.2007 - 16:10.41 Uhr

***Figure 6.** Harvester user interface*

It is planned to replace Heritrix 1.6 by Heritrix 1.12. This new version delivers a better data format for the harvested files (WARC). Furthermore it is possible to run different harvesting instances at the same server.

Process Tracking

Websites registered by the Ingest system are currently not analyzed prior to harvesting, but rather are transferred directly to the harvesting process. Heritrix is configured to collect only files located within the registered domain name. If a directory is entered in addition to the domain name, Heritrix considers it the root directory and retrieves only data that is located in the corresponding directory tree.

Employees at the NL only check a website when problems occur during harvesting and Heritrix is specially configured for this case. To date, problems have primarily involved websites with news tickers, local or city maps with scroll and zoom features or sites that provide a calendar of unlimited duration.

Indications are that considering the instructions in robots.txt helps to avoid problems during harvesting. In individual cases, consultations with the webmaster and changes to this file may simplify the request. Experience to date suggests that no significant data loss is expected by considering robots.txt.

The systematic, manual review of collected websites is still not possible at this time at the NL. It is, however, planned for the future.

entarc

The NL had a small program developed, which unpacks the ARC files generated by Heritrix 1.6 and writes the files therein to the corresponding directory structures. The reason for this development was the concern that ARC, as a non-standard file format, may cause problems during decryption in a few years time. Another consideration was to store all processed data in the form of tarballs for reasons of consistency. With the standardization of WARC, the NL will, however, plan this new format for long-term archiving of websites.

Outlook

The NL is continuously striving to integrate additional producers of electronic publications into the Ingest process. The users should receive access within the framework of applicable copyright laws to the archived electronic publications by the end of next year. Finally, suitable migration or emulation strategies as well should guarantee the readability of future existing data.

Regarding the archiving of websites, the NL has high hopes from its membership in the International Internet Preservation Consortium (IIPC).

References

[1] Reference Model for an Open Archival Information System (OAIS), Consultative Committee for Space Data Systems, CCSDS 650.0-B-1, January 2007 (http://public.ccsds.org/publications/archive/650x0b1.pdf)

[2] Metadata Standards Framework – Preservation Metadata (Revised), National Library of New Zealand, June 2003 (http://www.natlib.govt.nz/downloads/metaschema-revised.pdf)

Author's Biography

After being a teacher for several years Hansueli Locher decided to turn his hobby - computer science - into his profession. He worked at the Swiss Federal Statistical Office where he was responsible for database supported evaluations of statistical data. He developed also a library system and supervised information projects with strong IT-links. Since 2000 he is working at the Swiss National Library. As Project Manager "Archiving", he is responsible for the technical aspects of long-term preservation of digital objects. He is also the Head of Information Technologies at the Swiss National Library.