

e-Diss.ch : collecting and archiving online theses at the Swiss National Library

Barbara Signori

*Federal Department of Home Affairs FDHA, Swiss Federal Office of Culture FOC, Swiss National Library NL,
Hallwylstrasse 15, CH - 3003 Bern, Switzerland, E-Mail: barbara.signori@nb.admin.ch*

Abstract

e-Diss.ch is a project undertaken in collaboration with the Swiss university libraries as part of the e-Helvetica project at the Swiss National Library. The mission of the Swiss National Library is to collect and archive all Swiss publications, both printed and electronic. The objective of the e-Helvetica project is to fulfill this mission for electronic publications. The goal of the e-Diss.ch project is to set up a collection of all Swiss online theses and to preserve them in the Digital Archive of the Swiss National Library. In collaboration with several Swiss university libraries we have designed and tested a shared workflow for the collection, cataloguing, archiving and dissemination of online theses. Standards are used to ingest the publications into the Digital Archive of the Swiss National Library (webform, OAI-PMH), to describe the publications in the catalogue (XMetaDiss, MARC) and to identify the publications within the Digital Archive (URN on the basis of NBN). Furthermore a policy is being developed for the assignment of URN-NBN in the entire Swiss academic domain. This paper gives some general information on the e-Helvetica project and how the Digital Archive is being built at the Swiss National Library. The approach chosen for the collection, cataloguing and dissemination of the online theses will be highlighted as well as the creation of a social infrastructure of the Swiss university libraries and other partners.

Introduction

It is the task of national libraries to preserve the intellectual and cultural heritage of their countries. Nowadays much of this heritage is expressed in electronic form and many national libraries have started projects to collect and archive electronic publications. As a result standards and best practices have emerged that new initiatives are able to draw upon.

The Digital Archive of the Swiss National Library^[1] draws on the Open Archival Information System (OAIS)^[2] that describes a technical and organizational infrastructure for preserving data and making it available to a defined user group. The federal governance approach in Switzerland requires the creation also of a social infrastructure to support the archiving.

At the Swiss National Library the various activities and projects that are needed to build these capabilities have been brought together under the umbrella of the e-Helvetica^[3] project. As part of e-Helvetica, e-Diss.ch is a good example of the approach being taken to address the issues that emerge in the preservation of Swiss heritage for the future.

The Swiss National Library and the e-Helvetica project

The Swiss National Library was founded in 1895. It is the main source worldwide for written material on Switzerland and the Swiss. The Swiss National Library has a legal mandate (Federal Act on the Swiss National Library of December 18, 1992) to collect, catalogue, preserve and make available printed information as well as information stored on other information carriers with a bearing on Switzerland (this is called "Helvetica"^[4]). The collection of the Swiss National Library consists of over 3.8 million items with an average growth of 50,000 items per year.

The Swiss National Library is not subject to legal deposit legislation for either printed or electronic publications. However since 1915 there has been an agreement on the gratis supply

of Swiss publications, signed by the Swiss National Library and the Swiss Booksellers and Publishers Association^[5]. This agreement was renewed in 1961. A specific agreement for electronic publications does not yet exist.

The 1992 Federal Act does not specify the type of information carrier merely that the information must be related to all that is Switzerland and the Swiss. With the rise in importance of the Internet since 1993, the Swiss National Library began addressing the question of how in the future electronic publications could be stored on a long-term basis and thus preserved for subsequent consultation. National libraries have the task to work actively against the loss of intellectual and cultural heritage. Electronic media represents more and more an important part of this heritage. As a result, national libraries are facing new, additional and far-reaching tasks.

Since 2001 the e-Helvetica project of the Swiss National Library has taken on these tasks.

The goal of e-Helvetica is to establish the foundations for the collection, cataloguing, archiving and dissemination of electronic Helvetica, both offline (electronic publications on data carriers such as CD-ROMs or floppy disks) and online (electronic publications published on the web without any data carrier such as e-journals and websites).

The project is divided into two separate initiatives. The Organization initiative deals with the librarianship aspects of the e-Helvetica collection – that is, determining the collection's contents, their cataloguing and their dissemination. Archiving covers the IT aspects of the e-Helvetica collection, namely the setting up of the technological foundations and the IT applications for preparing and storing the collection.

The project is being developed by specialists in library science as well as IT experts. There are nine project staff equivalent to a total of 3.8 FTE's.

The e-Helvetica project is on schedule to be completed and operational by the end of 2008.

Establishing the electronic Helvetica collection

The Swiss National Library has decided to build a selective rather than a comprehensive collection of electronic publications. The Swiss National Library does not have the resources nor the legal mandate to collect the entire .ch domain. Accordingly, the decision was made to begin with selective coverage but to guarantee long-term preservation for this data. Offline publications such as CD-ROMs and floppy disks will be collected to the fullest extent possible. Online publications such as e-journals and websites will be collected selectively but in as broadly representative a range as possible.

Four pilot projects have been chosen for the implementation phase of the e-Helvetica project. Together these cover a broad range of content types from various producers of electronic publications.

Online theses

The "e-Diss.ch" project focuses on collecting online theses and "Habilitationsschriften"^[6] from Swiss universities. The collection is being developed through the coordinated efforts of the Swiss National Library and the Swiss university libraries, with the support of the Conference of University Libraries in Switzerland^[7].

Online commercial publications

Project “POP” collects online commercial publications belonging to the Basel-based Karger-Verlag^[8] and the Berne-based Stämpfli-Verlag^[9] publishing houses. The collection comprises mainly online journals, monographs and book series. It is being built in collaboration with the publishers.

Web resources

“Web Archive Switzerland” deals with the collection of relevant regional and cultural web resources (grey online literature). The collection is being developed together with the Swiss cantonal libraries, who are responsible for the selection of the publications to be collected.

Online governmental publications

“Electronic governmental publications” (working title) covers the collection of official online publications of the Federal Administration. Partners are the Federal Offices. This project involves the Swiss Federal Archive^[10].

The processing and storage of the electronic Helvetica collection

The Digital Archive of the Swiss National Library is being built according to the directives of the Open Archival Information System (OAIS)^[2]. Adopted as ISO 14721, this reference model (see Figure 1) describes an archive as an organization where people and systems work together on the problem of preserving information data and making it available to a defined user group. The model gives a detailed description of how producer-issued electronic information should be integrated into an archiving system, what preparatory steps must be undertaken for long-term archiving, and how information stored in the archive can be accessed.

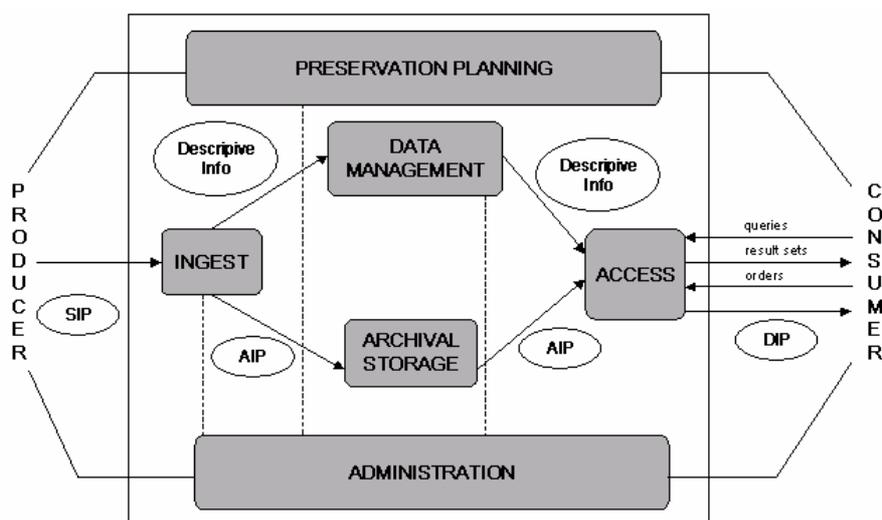


Figure 1: OAIS reference model

In e-Helvetica each OAIS process is tackled as a separate IT project.

Archival Storage

The archival storage process has been designed to create synergies within the Federal Administration in collaboration with the Swiss Federal Archive. The storage system was put out to WTO tender; Tape Libraries from ADIC and Hierarchical Storage Management (HSM)

from StorNext were chosen. The data is stored on LTO-2 and AIT-3 tapes running in Neuchâtel and Berne with redundancy across the two systems. In its current configuration the storage system can store up to 30 TB data and is scalable to 300 TB. This configuration will be refreshed next year. The AIT-3 tapes are due to be replaced and the long-term storage system needs to be better integrated into the existing IT architecture of the Federal Administration.

Ingest

The company Elca^[11] was chosen in another WTO tender to set up an Ingest system for the Swiss National Library. The Ingest system manages the automated delivery of the online publications (together with the metadata) to the Library and processes them until the digital objects are finally stored in the Tape Libraries. Ingest integrates the existing tools for the collection of the online publications and is flexible enough to meet future developments. It has been operational since February 1, 2007.

Data Management

The Data Management process was defined alongside Ingest. The Swiss National Library's metadata format is based on the standard developed by the Library of Congress: METS (Metadata Encoding & Transmission Standard)^[12]. MARCxml^[13] is used for the bibliographic metadata. For technical and administrative metadata, the Preservation Metadata format of the National Library of New Zealand (PRESMET)^[14] is used.

Access

The process for ensuring access to the stored online publications is still at the planning phase, scheduled for implementation in 2008. The library catalogue Helveticat^[15] of the Swiss National Library serves today as the entrance point for all media collected at the Library. It is therefore likely that access to the Digital Archive will be through Helveticat. A special interface and a cache system will operate between the library catalogue and the Archive.

e-Diss.ch

The goal of the e-Diss.ch project is to establish an archive for all Swiss online theses at the Swiss National Library with automated processes for handling them. The ultimate aim is to give users permanent access to online theses through any portal.

The technical and organizational infrastructure needed for e-Diss.ch is being developed within e-Helveticat. Accordingly the technical platform to process online theses is built on the basis of the OAIS reference model^[2] and uses standards as much as possible (eg METS^[12], XMetaDiss^[16], MARCxml^[13], URN^[17]).

The building of relations with the Swiss university libraries and other stakeholders has been critical to the success of this project. It has enabled resources and knowledge to be shared among the libraries and institutions in Switzerland and abroad (there is collaboration with the national libraries of Germany^[18] and Austria^[19] regarding standards such as XMetaDiss and URN). Most important however is the creation of a social infrastructure that ensures that standards and processes are followed and that the archiving becomes an integral part of the day-to-day university library workflow.

The key to building the social infrastructure is the integration of the stakeholders into the project from a very early stage - making them a part of the project albeit from a distance. A

number of techniques are used to help build stakeholder commitment. A network of contact people at the Swiss university libraries was established and they are kept updated on progress regularly through letters, e-mails and conferences. These people also helped to test the webform and other ingestion processes. Partnerships were set up with other projects, institutions and networks in Switzerland. No single technique has proven successful by itself; rather it is the combination that is effective. It is worth noting that considerable time and effort is required to establish and maintain the social infrastructure. For e-Helvetica so far this has been 20-30% of the total project effort.

The Swiss Academic Domain

Switzerland has thirteen universities offering postgraduate studies situated in the German, French and Italian speaking parts of the country. According to Swiss National Library statistics between three and four thousand new theses are created per year. As of January 2007, there were 10,625 theses available online in Swiss university repositories. The trend is very much towards online publication with some Swiss university libraries changing their rules to accept the online version as the official version of a thesis.

The libraries of the Swiss universities are part of the Conference of University Libraries in Switzerland as is the Swiss National Library even though it is not a university library. Several Swiss university libraries have their own local projects dealing with online publication of theses, helping their graduate students create them and storing them in their own institutional repositories. A map^[20] with some brief comments has been published on the web and provides a quick overview of the institutional repositories that have been built at Swiss universities.

There are two major library networks for university libraries in Switzerland: “Informationsverbund Deutschschweiz”^[21] for the German speaking part of Switzerland and “Réseau des bibliothèques de Suisse occidentale”^[22] for the French and Italian speaking part of Switzerland.

The Consortium of Swiss Academic Libraries^[23] coordinates the supply of electronic information for all Swiss universities. The project E-Archiving^[24] of the Consortium deals with the improvement of the accessibility of electronic contents in universities as well as with the long-term preservation of digital objects in accessible and usable form.

Approach

The first phase of the project is limited to online theses and “Habilitationsschriften”^[6] at Swiss universities. If needed in future the collection will be enlarged with theses from Master and Bachelor programmes and with grey literature from Swiss university repositories. If publication has been in online and in print, both versions are collected by the Swiss National Library during the project. When fully operational the Library will limit the collection to the “official form” in which the theses have been published, either print or online.

All online theses are catalogued using the Anglo-American Cataloguing Rules (AACR2)^[25] and entered into the Helveticat library catalogue^[15]. Theses are assigned a persistent identifier and stored in the Digital Archive of the Swiss National Library.

A document^[26] has been written for the university libraries showing the delivery process of online theses to the Swiss National Library as well as giving technical recommendations for the transmission, the format and the metadata.

Cooperation

At the heart of the e-Diss.ch system are the agreements between the stakeholders that form the social infrastructure for this archiving solution. There is an umbrella agreement with the Conference of University Libraries in Switzerland acting on behalf of its members. This agreement covers the archiving and access of the online theses in broad terms. In future the archive and access rights will be covered by formal contracts that specify the full details drawn up separately with each university library. The universities for their part include the rights needed for archiving at the Swiss National Library in the author consent forms.

The agreement with the Conference of University Libraries was established in 2003 and revised in January 2007^[27]. It has served as the basis for the work to date and it covers the following major points:

- The agreement to use the metadata format XMetaDiss
- The commitment to act as data provider and allow harvesting of metadata with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)^[28]
- The agreement to use platform-independent, non-proprietary, structured, stable and widely accepted data formats, such as Adobe PDF, PDF/A, XML etc.
- The commitment to deliver to the Swiss National Library one online copy for each thesis for the purposes of long-term archiving and access
- The recommendation to amend publications rules for postgraduates to formalise the online version

The metadata standard XMetaDiss (based on Dublin Core Metadata Element Set and Electronic Theses and Dissertations Metadata Set in format XML) was first developed and widely used in Germany. Thanks to the close relationship and cooperation with the German National Library, the Swiss National Library has also adopted this standard. The adoption of existing standards makes the exchange of metadata between the institutions much simpler. For example, the common standard between the national libraries of Germany, Austria and Switzerland will be of great value to the DissOnline Portal project^[29] that aims to make the online theses of the three countries accessible. Using the same metadata greatly facilitates the workflows and the exchange of data.

In addition, common standards tend to be more stable and reliable the more widely they are used. Nevertheless changes to standards in the future cannot be excluded and systems must be built with this in mind. The Swiss National Library follows activities and developments around standards both nationally and internationally.

URN's (Uniform Resource Names) have been used as persistent identifiers for archiving material in different national libraries worldwide, including the German National Library. The Swiss National Library decided to follow this standard for e-Helvetica. There is a project underway to test the feasibility of directly assigning URN's for the Swiss university libraries. Partners involved are the two library networks as well as the Consortium. In theory there is much to be gained through shared resources and by having a coordinated, central assignment of URN's. Several options are under study ranging from a central office run by the Swiss National Library to a decentralised system run by the Swiss university libraries. The preferred option at the moment is a mix of both central and decentral assignment. In other words an institution can choose if it wants to assign URN's itself or if the Swiss National Library will do

it on their behalf. There is binding condition that only those institutions with a digital archive intended for long-term preservation may assign URN's.

Workflow

The workflow for the automated processing of online theses in the Swiss National Library is made up of the steps shown below.

Receiving metadata

There are two ways for metadata to be delivered to the Swiss National Library. Either a webform is completed by the university libraries and sent by e-mail as an XML attachment or the metadata is harvested directly from the repositories of the university libraries using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). In both cases XML files containing the metadata are directly processed into the Ingest system of e-Helvetica. The webform is used mainly by the smaller university libraries; the larger ones prefer OAI-PMH.

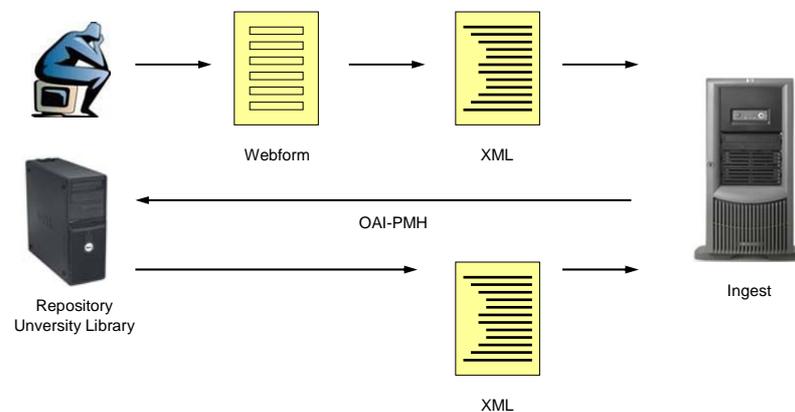


Figure 2: Receiving metadata

Standardising metadata

The metadata formats of the university libraries (either XMetaDiss or MARCxml) are transformed via XSL into the internal data format of the Swiss National Library.

Checking rights

The rights for archiving and accessing online theses are covered by the agreement with the Conference of University Libraries in Switzerland and in the author consent forms. In future these rights will be further covered by a contract with each Swiss university library that specifies the details. In most cases the Swiss National Library obtains the full rights to archive the online theses and to undertake further procedures for long-term preservation such as migration or emulation. A separate set of access rights define how the online theses may be accessed at the National Library, whether freely available or restricted in some way.

Transmitting online theses

Once the rights have been checked the system is ready to acquire the online theses. Ingest uses wget^[30] to pick up the theses as digital objects from the repositories of the university libraries using the links read from the metadata. The digital objects are returned to Ingest

usually as PDF files. This is the common format for online theses in Switzerland today but other formats will likely be supported in future.

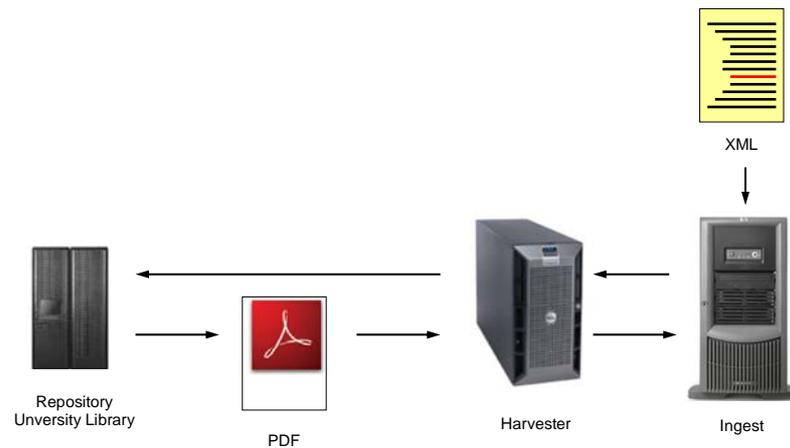


Figure 3: Transmitting online theses

Checking quality

Quality checks are carried out at different stages of the Ingest process. The digital objects are checked for viruses and for completeness. The data formats are checked using JHOVE^[31] and there is a control mechanism to guard against multiple delivery.

Generating Uniform Resource Names (URN)

For each online thesis a URN is generated automatically on the basis of National Bibliography Numbers (urn:nbn) for the unique identification in the Digital Archive of e-Helvetica. The URN resolving service of the German National Library is used for this purpose. It records each URN with its corresponding URL's and is responsible for the correct resolution of the URN to these URL's.

Checking for metadata completeness

A final control step automatically checks that all information gathered during the ingestion process has been recorded in the metadata.

Output from Ingest

The completed metadata package is stored in Data Management (see Figure 1). The archival information package that contains the online thesis as well as its metadata is stored in the Digital Archive in the form of a tarball. The bibliographic metadata is transformed from MARCxml into MARC21^[32] and sent to the library catalogue Helveticat automatically creating a catalogue record.

Access

The Access process within e-Helvetica is not yet implemented although it is envisioned that access to the online theses will be through the Swiss National Library catalogue Helveticat. The URN will resolve first to the URL of the original online thesis at the university library repositories. In the event that an online thesis is not or no longer available on the library repository, the URN will resolve to the URL of the copy stored at the Swiss National Library. Since the Digital Archive itself will never be directly accessible to the user, a thesis will be

delivered from a cache (see Figure 4). Copies of highly requested items will be stored on the Cache system and will be delivered immediately from there. Other items will be requested from the Digital Archive as needed and made available to the user via the Cache.

Access to the online theses will also be possible through the university libraries' local catalogues, the library networks' catalogues and through other national and international portals such as DissOnline Portal.

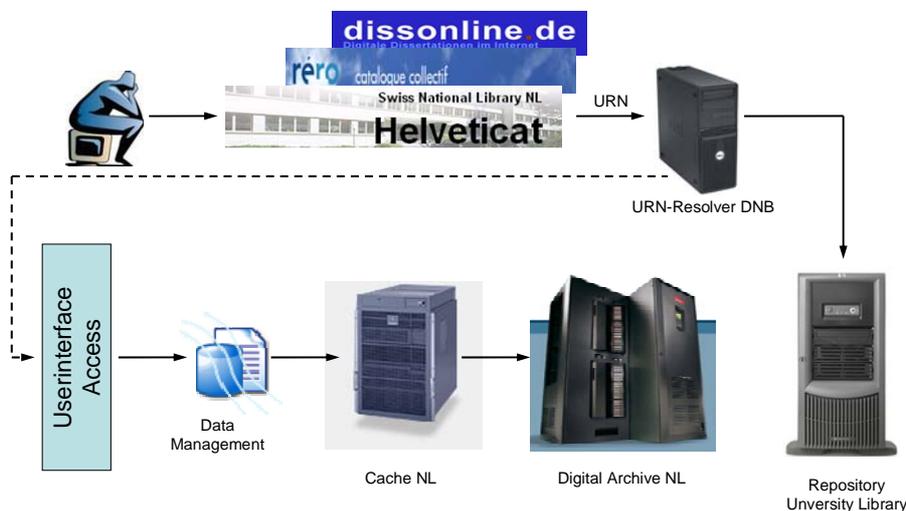


Figure 4: Access

Conclusions

National libraries have the task to preserve the intellectual and cultural heritage of their countries. Increasingly, electronic media represents an important part of this heritage. This is no truer than in the academic world where online electronic media dominates in the exchange of information. This is of course a very logical development since experience has shown that electronic dissemination of information is highly effective. It follows that academic institutions are keen to publish as much online as possible with the result that the size and number of institutional repositories is growing.

Doctoral theses, dissertations and “Habilitationsschriften”^[6] are an important part of the research output of an academic institution, often containing much more detail than other forms of publication such as journal articles. The output from all the universities in a country together represents a key source of intellectual heritage for that country. It is therefore of considerable value to Switzerland to archive this information from its thirteen universities and to guarantee its long-term preservation.

The availability of information by itself does not ensure the visibility of that information. The information must be findable and retrievable. The project e-Diss.ch sets standards that guarantee visibility across all academic institutions in Switzerland and that guarantees supply of the information into the future. It is anticipated that the visibility and availability of this information will result in wider dissemination both nationally and internationally with obvious benefits to Switzerland.

Finally, e-Diss.ch is of great value as a project to the e-Helveticat team at the Swiss National Library. It provides a real test of the technical infrastructure of the Digital Archive systems. Just as important however it allows the team to gain valuable experience in establishing the social infrastructure that will ensure the lasting success of digital archiving in Switzerland.

Further information

For further information (in German and French) please check the website of e-Helvetica at <http://www.e-helvetica.admin.ch>. An English version of this website will be available soon.

References

[1] <http://www.nb.admin.ch/slb/index.html?lang=en>

[2] <http://public.ccsds.org/publications/archive/650x0b1.pdf>

[3] <http://www.e-helvetica.admin.ch>

[4] The term Helvetica is used to denote the entire literary production of the country stored on all forms of information media (handwritten manuscripts, printed, digitized, in pictorial form or as sound documents), all works published abroad with a bearing on Switzerland, its population, its past, present and future works (including translations) by Swiss authors.

[5] <http://www.sbv.ch>

[6] A "Habilitationsschrift" is a written scientific work which, as a general rule, is necessary for being able to teach and obtaining tenure at a Swiss (or German) scientific university. It is the highest academic examination in the German-speaking world.

[7] <http://www.kub-cbu.ch/>

[8] <http://www.karger.com>

[9] <http://www.staempfli.com>

[10] <http://www.bar.admin.ch/>

[11] <http://www.elca.ch/live/3/home/>

[12] <http://www.loc.gov/standards/mets/>

[13] <http://www.loc.gov/standards/marcxml/>

[14] <http://www.natlib.govt.nz/downloads/metascema-revised.pdf>

[15] <http://www.nb.admin.ch/helvetica/>

[16] See "XMetaDiss : Schweizer Ausgabe" (in German) on the page:
http://www.nb.admin.ch/slb/slb_professionnel/projektarbeit/00719/00723/00629/index.html?lang=de

[17] <http://www.persistent-identifier.de/?lang=en>

[18] <http://www.d-nb.de/eng/index.htm>

[19] http://www.oenb.at/en/welcome_to_the_oenb.jsp

[20] http://lib.consortium.ch/external_files/Dok32a_V1_2_UebersichtIR_D.pdf (in German)

[21] <http://www.informationsverbund.ch/>

[22] <http://www.rero.ch/>

[23] <http://lib.consortium.ch/index.php?lang=2>

[24] http://lib.consortium.ch/html_wrapper.php?src=project-ea&dir=project&activeElement=2&ea=1

[25] <http://www.aacr2.org/>

[26] See "Empfehlungen für die Anlieferung von elektronischen Dissertationen" (in German) on the page:

http://www.nb.admin.ch/slb/slb_professionnel/projektarbeit/00719/00723/00629/index.html?lang=de

[27] See "Elektronische Dissertationen : Vereinbarung" (in German) on the page:

<http://www.kub-cbu.ch/> > Projekte > Elektronische Dissertationen

[28] <http://www.openarchives.org/OAI/openarchivesprotocol.html>

[29] <http://www.dissonline.de/>

[30] <http://www.gnu.org/software/wget/>

[31] <http://hul.harvard.edu/jhove/index.html>

[32] <http://www.loc.gov/marc/>