

Digitization Projects in Switzerland

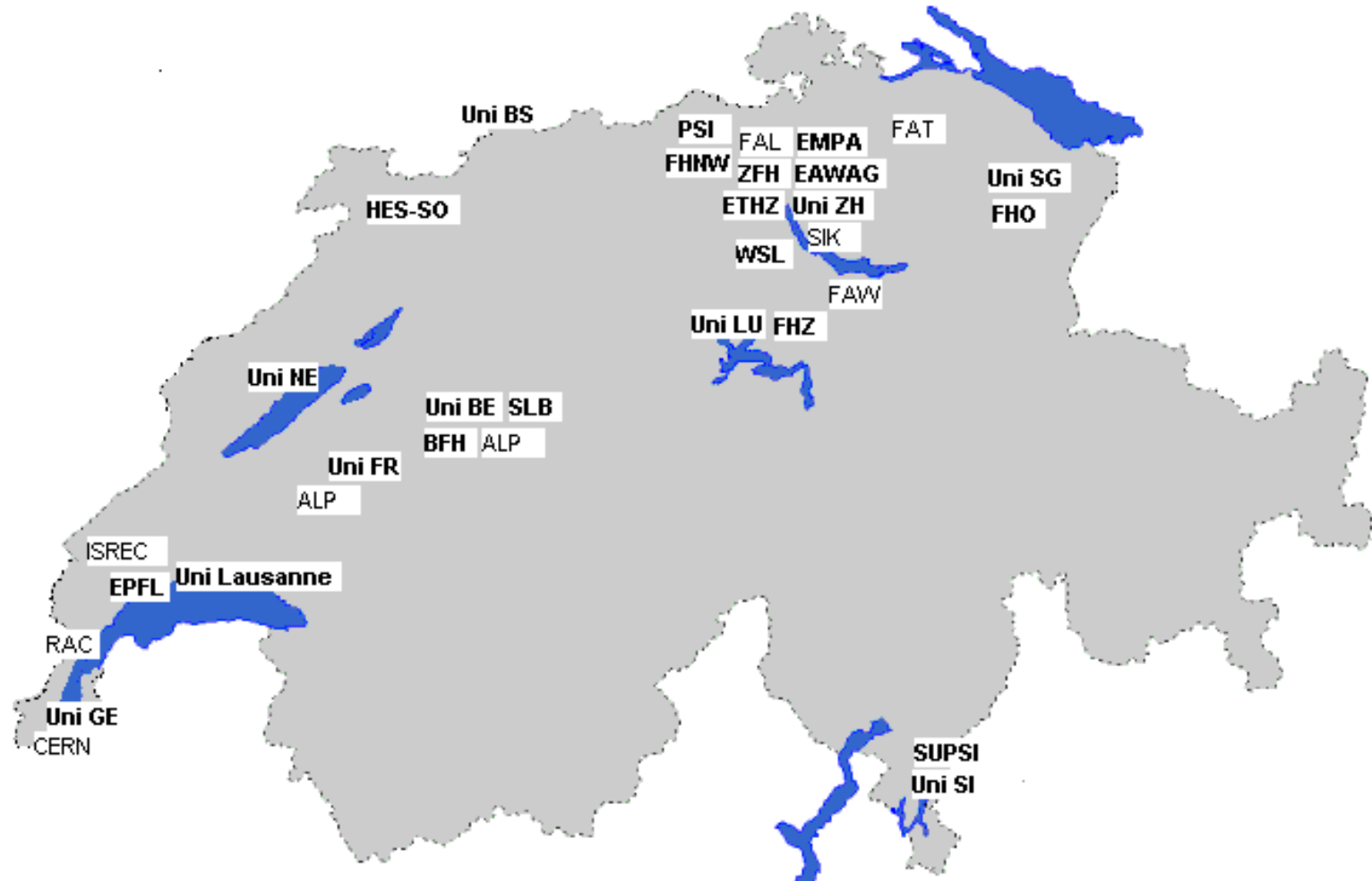
BESS-Seminar – Digitization Options for Libraries: Challenges and Opportunities,
Torino, 08 October 2007

Dr. Matthias Töwe
Consortium of Swiss Academic Libraries

Outline

- The Consortium and its project on E-Archiving
- Dimensions of digitization projects
- Overview of projects in Switzerland
- Practical experience from the Consortium's project
- Conclusion

Consortium of Swiss Academic Libraries



Consortium of Swiss Academic Libraries

Headquarters of the Consortium

<http://lib.consortium.ch>

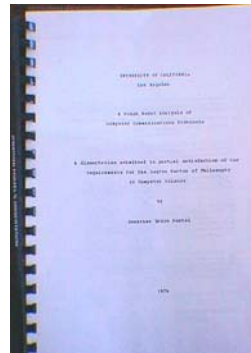
Module Licensing

- Licensing and related services
- 3 FTE
- Federal support 2000-2005
- Since 1/2006: Run by the Conference of University Libraries (KUB) and financed by contributions from all participating institutions

Module E-Archiving

- Permanent accessibility of electronic information
- 2.4 FTE
- Federal support from 2002 as a project of the Swiss University Conference(SUK)

Subprojects / fields of activity



Ein Trippelkinder — bekanntlich in diese
Welt in dem Wappel der großen Vater-
schafft ausgeht. Kann eine der schwie-
geren unter ihm, Minderer anwilligen
— harrt, nur er aus dem Streben nach Ver-
vollkommenheit, sperrt sich vor dem ge-
wöhnlichen Geschick sein Leben fern, er-
scheint, daß er, solange er im gleichen Charakter-
schritte, Tag und Nacht auf dem Trapes steht.
Aber selbst, solange der geistige Charakter-
wunde durch einander abwechselnde Diener an-
sprechen, welche seine wachen und schlaf, was
oben besetzt wurde, in einem konstanten
Gefahren laßt und bewachen. Besonders
Schwierigkeiten für die Umwelt ergaben sich aus
dieser Lebensweise nicht, nur während der un-
rigen Programmierung war es ein wenig
schwer, daß er, wie sich unter anderen, daß
oben gelassen war und daß, nachdem er sich
in solchen Tönen seine ruhig verhielt, bis und
da ein Blick aus dem Publikum zu ihm abtrat.
Doch verhielt sich die Charaktere, weil
er ein selbständiger, unerschütterlicher Künstler
war. Auch sah man natürlich ein, daß er nicht
aus Mitleiden so sehr und eigentlich nur so
sich in diesem Oben schaute, nur so seine
Küsse in diese Vollkommenheit bewahren konnte.
Doch war es oben auch sehr gesund, und
wenn in der wärmsten Jahreszeit in der ganzen
Wunde der Wunde die Schweißkühle aufsteigt
werden und mit der höchsten Luft die Sonne
schlingt in der dämmenden Wärme übersteigt,
dann war es dort sogar schön. Freilich, sein
menschenlicher Verstand war einseitig, und
manchmal Mente auf der Seite der ein-
seitigen zu ihm hinüber, aber er verhielt sich
dem Träger, ist ihm treu und hält an dem
Hauptknoten und phantasie, aber er verhielt sich
Bewerber des Dicht und wachenden einige Worte
mit ihm durch ein selbstes Fremde, aber er über-
prüfte die Feuerwachen die Nachbarschaft
auf der oberen Gasse und der ein wenig
Respektvoll, aber wenig Verstandliche an. Seine
bleib er um ihm still, nachdenklich sah er

E-Journals :
Priority is on
access.
Consideration
of *print
editions*
necessary.

Publications from Swiss universities

Carry forward
traditional *Hochschul-
schriften* (theses etc.)
in electronic form:
*Institutional
Repositories*

Strong interest in
new fields of activity
related to *Open
Access*

Not yet
digitized
Swiss
journals:
*Potential
and need
are there!*

Long-term preservation

Dimensions of digitization - non-exhaustive (I)

- Document types and their characteristics
 - Text, images, manuscripts, rare books, AV-material...
 - Paper and print quality, latin vs. gothic letters...
 - Unique vs. widely available; brittle vs. robust or duplicate...
→ implications for: use/audience, copyright, processing, character recognition, relation to the library's identity...
- Purpose: access, preservation, virtual integration of collections
→ implications for:
 - Required quality (resolution, colour, grey scale, black and white)
 - Requirements for long-term preservation
- Costs
 - One-time project costs of scanning, processing, hard- and software
 - Recurring costs of operation (staff, hardware, storage, licenses...)
 - Costs of long-term preservation

Dimensions of digitization - non-exhaustive (II)

- Collection type:
 - Physical collection or virtual collection of objects from various sources?
- Processing and workflow:
 - Processing in-house or by a service provider?
 - Material from more than one source?
 - Rare or duplicate materials?
 - Scanning from original, scanning from microfilm or exposure of digital images on microfilm?
- Partners
 - Internal partners, editors, publishers, libraries, service providers, other
- Integration:
 - Stand-alone database and/or integration into other retrieval tools?

Examples of projects: Physical collections

- Codices electronici sangallenses (CESG)
 - <http://www.cesg.unifr.ch/>
 - Medieval codices of the Abbey Library of St. Gall
 - High resolution, facsimile quality
 - University of Fribourg and Abbey Library St. Gall
- DigiBern
 - <http://www.digibern.ch/>
 - Digital texts on history and culture of the city and canton of Berne (books, newspapers, maps)
 - Full texts
 - Central Library of the University of Berne

Examples of projects: Google as partner

- Google Book Search
 - Bibliothèque cantonale et universitaire de Lausanne
 - Scanning of 100'000 volumes within two years (c. 5% of the collection)
 - 16th to 19th century (up to 1867)
 - Logistics and scanning provided by Google („highly professional“)
 - In-house selection and preparation through the library
 - No very precious books, no journals
 - The library receives a copy of the digital images and of the full text from the optical character recognition (OCR) for its own use
 - Data will be hosted with Google and within the libraries repository SERVAL
 - Improved access is the primary aim
 - For more information: http://www.bbs.ch/documents/Referat_hvillard.pdf

Examples of projects: Virtual collections

- e-codices – Virtual Manuscript Library of Switzerland
 - <http://www.e-codices.ch/>
 - Extension of CESH to other libraries
- Swiss Poster Collection
 - <http://posters.nb.admin.ch/>
 - Virtual integration of several distributed poster collections
 - Leading house: Swiss National Library
- Digitized Swiss Journals
 - <http://retro.seals.ch>
 - At the moment 450'000 pages in full text, 1 Mio. planned for 2008
 - Consortium of Swiss Academic Libraries, ETH-Bibliothek, scholarly societies, publishers

Examples: Enrichment

- Use of selected digitized images for the enrichment of bibliographic and other databases:
 - Online catalogue „Griechischer Geist aus Basler Pressen“ of 15th to 17th century printed Greek texts:
<http://www.ub.unibas.ch/kadmos/gg/>
 - Online catalogue „Opera poetica Basiliensia“:
<http://www.ub.unibas.ch/spez/poeba/index.htm>
 - Enrichment of library catalogues with abstracts and indices for monographies, e.g. ETH-Bibliothek and ZB Zurich in NEBIS-catalogue (www.nebis.ch)
 - Enrichment of database of rare books:
<http://ad.e-pics.ethz.ch/>
 - Earlier: card catalogues

Examples: special collections

- Maps
 - The Ryhiner Map Collection (ZB/UB Berne):
<http://www.zb.unibe.ch/stub/ryhiner/>
 - ZB Zurich
 - Others
- Collections for special purposes or occasions:
 - Einstein Online (Albert Einstein's scripts) at ETH-Bibliothek:
http://www.ethbib.ethz.ch/eth-archiv/einstein/index_e.html
 - Minutes of the ETH's School Board meetings online (1854-1955)
<http://www.sr.ethbib.ethz.ch/digbib/home>

Examples of other projects

- Newspapers (sometimes from existing microfilm)
 - Bibliothèque cantonale et universitaire de Fribourg (en <http://doc.rero.ch>)
 - Bibliothèque cantonale et universitaire de Lausanne
 - Médiathèque Valais (en <http://doc.rero.ch>)
 - Swiss National Library
 - Others
- Doctoral theses
 - Bibliothèque centrale de l'EPF Lausanne (complete from 1920, not all publicly accessible: <http://library.epfl.ch/theses/>)
 - ETH-Bibliothek (complete run in preparation)
 - Université de Neuchâtel
 - Others
- Collections of images, audiovisuals
 - Bibliothèque cantonale et universitaire de Lausanne
 - ETH-Bibliothek (http://ba.e-pics.ethz.ch/ETH_Bibliothek/Standard/)
 - Phonoteca svizzera (Lugano)

State of affairs

Many small to medium initiatives, few medium to large projects.

- *Digitization is an important topic for most scientific and cultural heritage institutions, but still some scepticism (expenses, usage)*
- *However, there are differences between university libraries and cultural heritage institutions: other needs and challenges (e.g. access vs. preservation)*
- *So far there is a lack of central coordination, much is done „bottom-up“*

Dimensions and decisions in practice (I)

- Document type: Swiss journals
 - Non-unique, reasonably robust, duplicates available for flat scanning
 - Copyright: Legally, each author would have to be contacted. Practically, scholarly authors want their work to be distributed. No objections so far. Consent of the editing society as representative of authors is secured.
- Purpose: improved visibility and access
 - Character recognition as a „must“ for full text search. Virtually only latin letters. No correction.
 - Manual capture of correct metadata
 - Open access wherever possible – moving wall as concession to publishers.
 - Reasonable quality, but not always facsimile

Dimensions and decisions in practice (II)

- Costs
 - One-time project costs (related to quality): different models, e.g. equal contributions from Consortium, partner library and publisher
 - Recurring costs: contributions from editing society, funding agency, sponsors
 - Costs of long-term preservation: it may be more economically sound to recreate images if necessary. In practice, data that is regularly *used* is much less endangered by obsolescence.
- Partners:
 - No competition with publishers intended.
 - Technical service: Some pass their currently produced files regularly over to the service.

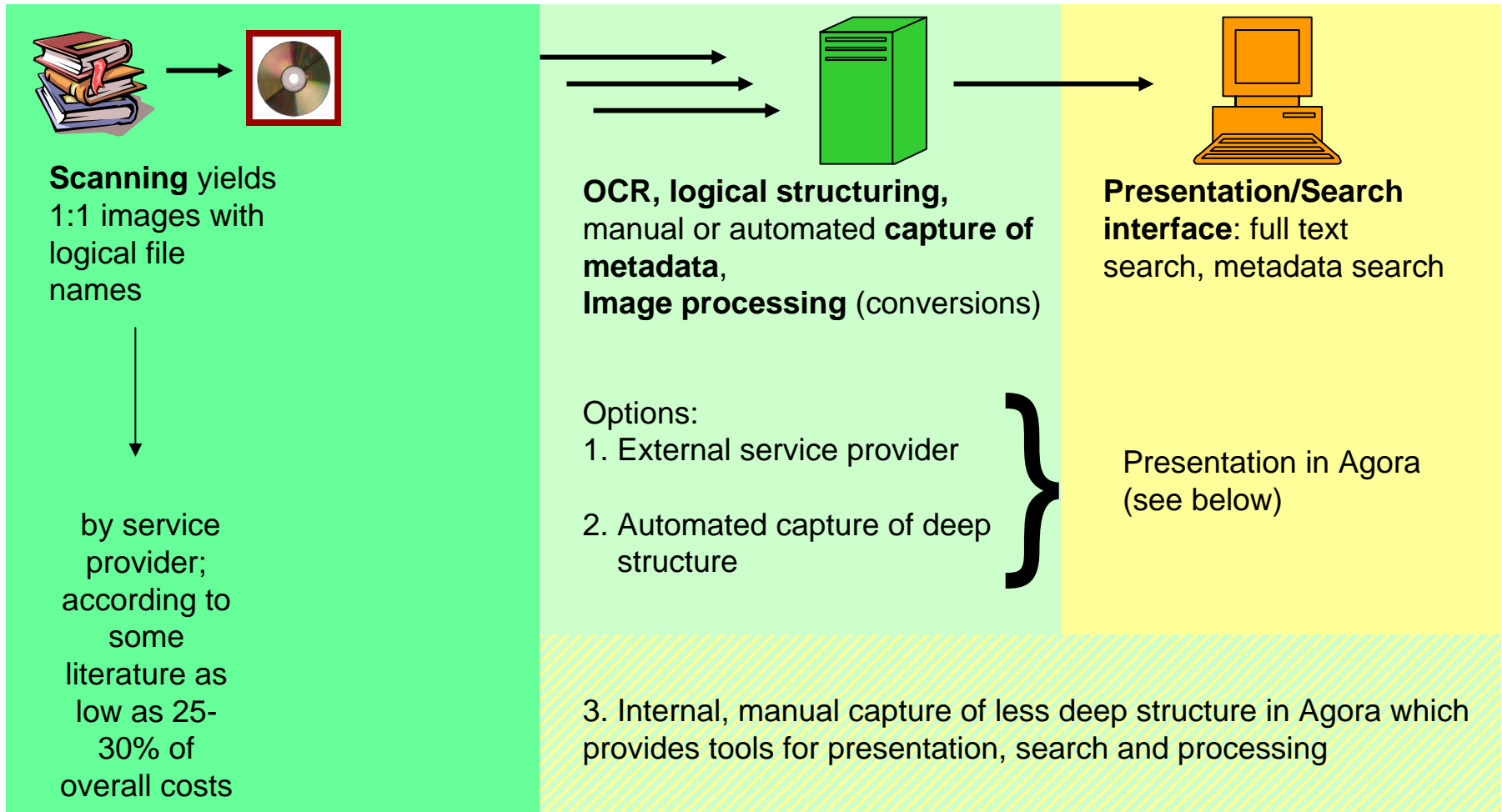
Dimensions and decisions in practice (III)

- Processing and workflow:
 - Scanning mainly by a service providers, exceptions in-house
 - Material from libraries, societies and their members, printers, publishers: rarely complete back runs available from publisher
 - So far only scanning from originals, there may be cases where partners want to obtain microfilms at the same time
 - Manual vs. automated metadata capture and structuring: test of automated process successful, but costly when no deep structure required
- Integration:
 - So far linking from publisher's website, library catalogues, SFX and databases (ZDB, EZB). Integration into *E-lib.ch* (new Swiss Electronic Library) as an issue.

The Consortium's project: procedures (I)

- Prerequisite: some articulation of interest from scholars or other groups
- Contact with known stakeholders: editors, editing scholarly society, publisher, funding agencies (e.g. academies of sciences)
 - Discussion of benefits of the project and of the technical solution
 - Requirements regarding quality
 - Status of copyright
 - Cost division
 - Possibility of the use of duplicate volumes
 - Signing of an agreement by all involved parties
- Careful check of all available volumes for missing items or pages. In parallel marking for scanning in colour/grey scale/black and white. This information is listed in Excel-sheets which also contain a predefined file name for each page's image.

The Consortium's project: procedures (II)



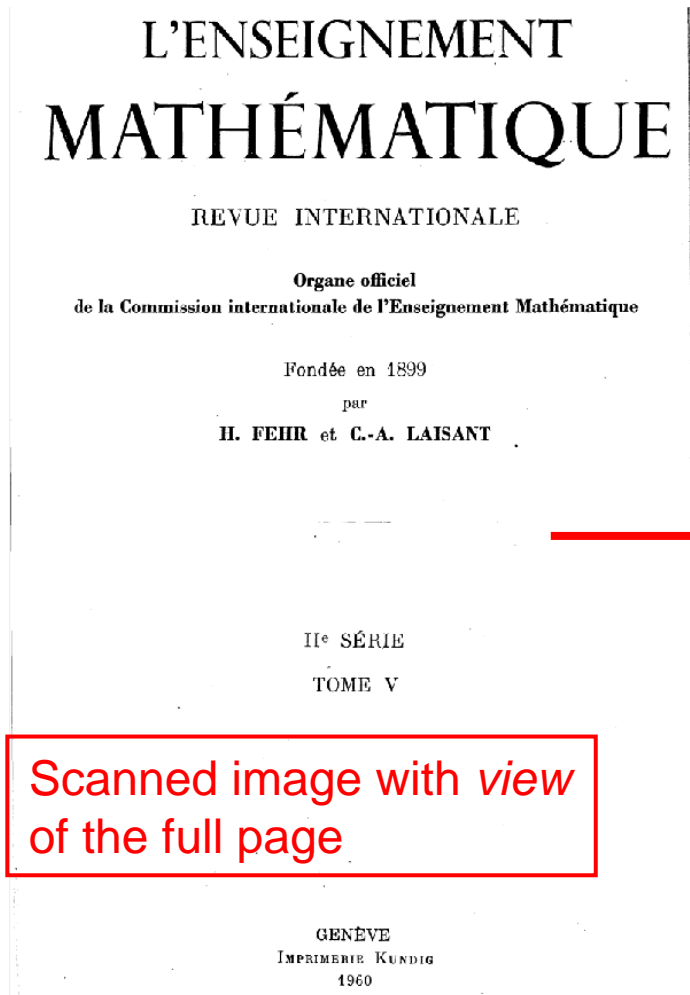
The Consortium's project: procedures (III)

- Routine scanning is performed by commercial service providers who receive the originals and the Excel-sheets
 - Colour and grey scale at 300 dpi
 - Black and white at 300 dpi to 600 dpi
- Special items (large formats etc.) are scanned in house
- Images (TIFF) with predefined file names are delivered on hard-disk or on DVDs
 - Challenge of handling large data volumes (several TBs):
Temporarily high demand for storage space, considerable processing times
- Externally and internally scanned images are merged into proper order

The Consortium's project: procedures (IV)

- Images are converted for a reasonably fast web presentation (mostly to JPEG, sometimes to GIF)
- Optical character recognition (OCR) generates a file containing the full text extracted from each page image including each word's position within the image. This information is used to highlight search hits in the page images.
- Metadata are captured manually with a special XML-editor. This is where each image is looked at and checked. In parallel, the content is structured: which items belong to the same article etc. The result is an XML-structure.
- XML-metadata and structure are fed into a designated database („repository“), search indices are generated for metadata fields and full text.

Optical character recognition (OCR)



OCR



L'ENSEIGNEMENT
MATHÉMATIQUE
REVUE INTERNATIONALE
Organe officiel
de la Commission internationale de
l'Enseignement Mathématique
Fondée en 1899
par H. FEHR et C.-A. LAISANT

II^e SÉRIE
TOME V

GENÈVE
IMPRIMERIE KUNDIG
1960

Recognized text,
machine readable

The Consortium's project: current status (I)

- Currently under <http://retro.seals.ch> (ger/fr/en)
 - Ten journals in two collections, more in preparation
 - C. 450'000 pages (c. 1'000'000 pages in 2008)
 - Moving Wall of six months to five years
 - Output as online page image (display in browser, additional viewer where appropriate) and downloadable article-PDFs
 - Content Management System AGORA (Satz-Rechen-Zentrum, Berlin, www.agora.de)
 - Positive echo – is appreciated as a service from libraries
 - Transfer into routine workflows
- Long-term preservation
 - Only TIFF-images and XML-metadata (rest can be re-generated)
 - Digital reproduction not as substitute for the original

The Consortium's project: current status (I)



digitalisierte zeitschriften

part of seals - swiss electronic academic library service

 suchen

- > Erweiterte Suche
 - > Browsen
 - > Sammlungen
 - Mathematik (SwissDML)
 - Baugedächtnis Schweiz Online
 - > Letzte Trefferliste
-
- > Home
 - > Über uns
 - > Aktuell

Sammlungen > Baugedächtnis Schweiz Online

Baugedächtnis Schweiz Online

Um eine Übersicht über die aktuell verfügbaren Bände einer Zeitschrift zu erhalten, klicken Sie bitte auf das jeweilige Bild der Zeitschrift oder deren Titel.

Titel	Bände	Erscheinungszeitraum	Info
> Tec21	106 - 118, 127 - ff	2001 - ff	
> Schweizer Ingenieur und Architekt	97 - 105	1979 - 2001	
> Schweizerische Bauzeitung	1 - 128, 65 - 96	1883 - 1978	
> Die Eisenbahn	1 - 17	1874 - 1882	
> Tracés	127 - ff	2001 - ff	
> Ingénieurs et architectes suisses	105 - 127	1979 - 2001	
> Bulletin technique de la Suisse romande	26 - 104	1900 - 1978	
> Bulletin de la Société vaudoise des ingénieurs et des architectes	1 - 25	1875 - 1899	

Perspective *E-lib.ch*

- New federal project framework from 2008-2011
- Examples for new proposals:
- A joint large scale digitization project of several university libraries
- Extension of Codices Electronici Sangallenses
 - Vision: e-codices as virtual library for manuscripts in Switzerland
- Extension of digitization of Swiss journals within the Consortium
 - More regional content
- Projects for digitization of particular collections, improved indexing by search engines etc.

Conclusions (I)

- Something can already be done on a small scale.
- Know what is your intention when you digitize.
- Money and sufficient resources help to build a „critical mass“ in reasonable time.
- Check quality of data from external service providers carefully.
- You keep learning all the time and you cannot determine every detail in advance.

Conclusions (II)

- Don't underestimate necessary IT/computing-resources. You have to handle large volumes of data (storage, copying, conversions...). Talk to IT-people early.
- Seemingly similar materials can be very heterogeneous. Don't try to press them into same patterns. Keep flexible.
- Concentrate on one document type at a time.
- Keep listening to comments from partners, customers and colleagues, but don't try to satisfy everyone.

Thank you very much!

We thank the Swiss University Conference for its financial support of our project.



Dr. Matthias Töwe
Consortium of Swiss Academic Libraries
c/o ETH-Bibliothek
Rämistrasse 101
CH-8092 Zürich
0041-(0)44 632 60 32
matthias.toewe@library.ethz.ch
<http://lib.consortium.ch>