Th. Fischer
Staats-und Universitätsbibliothek (SUB) Göttingen

Math*Diss* Service

The goal of the MathDiss International project is to provide a convenient, comprehensive and subject specific access to mathematical dissertations. Towards this end the Math*Diss* Database has been set up at the Goettingen State and University Library (SUB). My presentation will deal with the advances made towards this goals, the obstacles met and the problems to be solved.

The foundation for this project has been laid by the Dissertations Online project, which developed a unified method to collect and distribute dissertations together with their respective metadata. Math*Diss* International investigates subject specific questions which can not be handled by this general approach in the case of mathematics.

The Math*Diss* Database has to handle data that are quite complex for different reasons:
- It has to meet the requirements set by the different national standards for scientific dissertations.
- It has to be linked with an archival function to provide secure access to the full text of the dissertations in different formats.
- It has to deal with the symbols and formats provided by mathematicians in their texts.

While some of the problems have been solved, others are still open. In particular, standardisations prove to be more urgent and more difficult to solve than was expected. This refers to the standard of metadata and their representation in HTML and XML (and/or RDF) as well as to the standardisation of TeX-installations.

The current Metamaker (My Meta Maker for Theses, MMMfT) as a general tool to create and edit metadata for dissertations in science will be presented as well as the Math*Diss* Database as the beginning of the collection of dissertations with browsing and search facilitates and the option to add and edit entries to the database.

Math*Diss* International and Math*Diss* Database

Thomas Fischer

Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB)

(Göttingen State and University Library)

Overview

- ➢ History
- ➢ Sources
- ➢ Goals
- ➢ The Gory Details …
- ➢ Standardisation
- ➢ The State of Affairs
- ➢ The Future: Co-operation

1

History: The basis of Math*Diss* International

- DissOnline (http://www.dissonline.de/) defined a procedure and a metadata set for online dissertations (http://www.ub.uni-duisburg.de/dissonline/metatags.html)

- "Die Deutsche Bibliothek" (DDB, German National Library) has started to collect and archive online dissertations (http://www.ddb.de/professionell/dissonline.htm) and has a detailed metadata set (http://deposit.ddb.de/metadiss.htm)

- More universities allow the electronic publication of dissertations, and more dissertations are published (see the collection of links at Duisburg: http://www.ub.uni-duisburg.de/dissonline/emathdiss.html)

Sources

- Networked Digital Library of Theses and Dissertations (NDLTD) (http://www.ndltd.org/) provides a framework for scientific electronic documents and a metadata format.

- My Meta Maker for Theses produces output conforming to national and international standards (Example) (http://www.mathematik.uni-osnabrueck.de/cgi-bin/MMMTh/V0.5/MMMfT.cgi).

- The SUB Göttingen is the Special Subject Library in Germany for Pure Mathematics and has the facilities to provide archival functions as well as convenient retrieval (see e.g. the MathGuide http://www.MathGuide.de/).

Goals

- While the DissOnline project treated electronic dissertations in general, and the DDB archives all electronic dissertations, scientists want a specific access to their specific data.
  (This refers to Chemists, Physicists, Astronomers ... as well.)
- Math*Diss* International will start to develop a subject specific access to Mathematical theses. This includes:
  - Classification by MSC 2000
  - Exploitation of the T$_E$X Format
  - Access to specific parts of the documents like tables of contents and bibliographies

  These Math*Diss* Services will be freely available internationally.

The gory details 1

- Math*Diss* International wants to provide metadata conforming to national and international standards. This makes the data fairly complex.
- These data shall be delivered in RDF format in XML. But the standards for representing qualified Dublin Core metadata in RDF/XML are still under development.
- The database will be accessed via Internet, using HTML and or XML. The presentation of Mathematics under these conditions is still problematic (MathML, IBM techexplorer).

The gory details 2

Problems in the T$_E$X files include:

- Special formatting: custom style sheets (not two dissertation have the same type, some are T$_E$X, some LaT$_E$X...). Most of them do not compile correctly on a standard installation.
- German dissertations come with Umlauts. We found ä as ¨ a, ¨a, ""a, "" a, "a...
- Some dissertations come with ligatures ff, fi ... which do not print at all with standard fonts.
- Since the database is essentially text based, _ is an additional problem.

The gory details 3

- Additional work is necessary to complete the metadata sets. The sets we received were (and still are) missing important data, like the MSC classification, keywords or personal and institutional data.
- Other data sets contain abstracts of an excessive size.
- The T$_E$X files are sometimes a collection of necessary and unnecessary objects (e.g. backups), while others (e.g. styles) are missing. In general T$_E$X files are hard to handle.

The State of Affairs

- My Meta Maker for Theses is available in a pre-release version and working.
- The Math*Diss* Database is installed and working.
- Math*Diss* International co-operates with the Zentralblatt der Mathematik, the NDLTD and the Open Archives Initiative
- Dissertations start to flow in. An efficient work flow is needed to unite the efforts.
- The formatting problem for mathematical text is not solved yet.

Standardisation is necessary!

- Standardisation of T_EX files
- Standards for high quality PDF files
- Standardisation of metadata
- Standardisation of metadata formats
- Standardisation of Workflow and metadata management

Files 1: T$_E$X

> For T$_E$X files a reference installation is necessary. This has to take into account:
> - Different operation systems (Windows, Mac, Linux …)
> - The development of T$_E$X : versions, updates, "flavours" (AMT$_E$X, LaT$_E$X)
> A library of additional files and an archive of outdated files may be necessary to keep the files readable.
> A check method for completeness is needed.
> This means also standards for the inclusion and exclusion of files from the package to be archived.
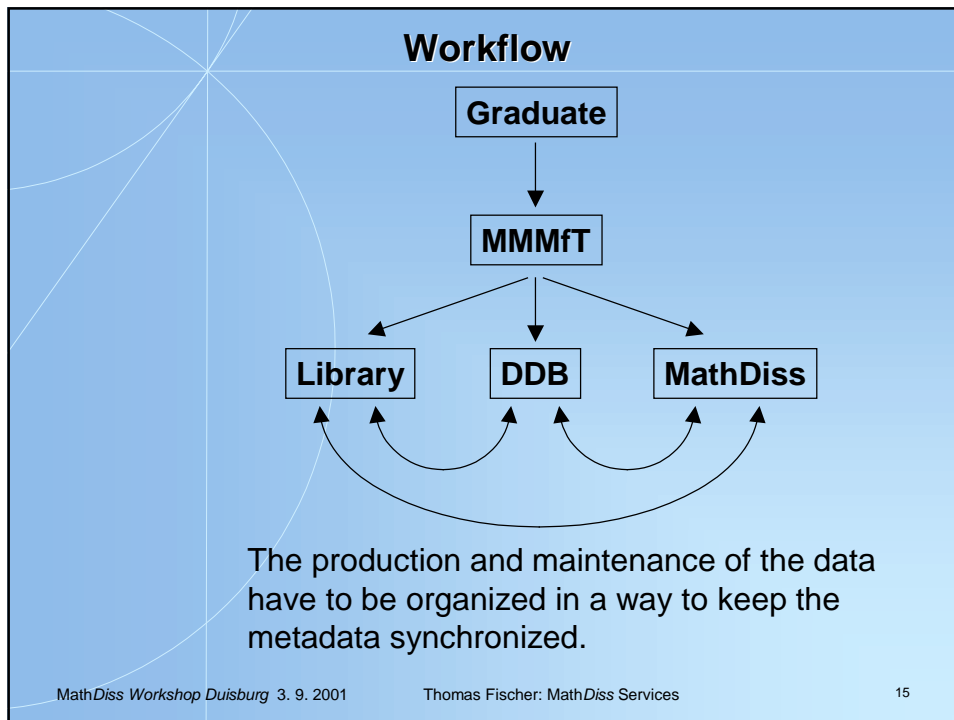
Files 2: PDF

> The standard format for display (and printing?) of the dissertations will be PDF. A high quality of these files is desirable. Standards include:
> - A reasonably high resolution of the files, to be displayed using a browser plug-in as well as for printing, optimised for web access.
> - The internal (and possibly external) linking of the files: using bookmarks, links into the bibliography and within the text.
> - The inclusion of metadata in the files. This should be at least the standard metadata provided by the format (title, author, topic, keywords). If available, Dublin Core metadata will be useful.

Metadata

➢ The set of metadata for the MathDiss files is quite complex. We need rules which data must be provided by the initial entry into the system, and who will provide the other required entries. The MMMfT already includes first checks on the completeness of the data.

  ▪ Basic data (title, author, keywords, abstract) as well as the data of the graduation have to be provided by the author.

  ▪ MSC classification should be provided by the author or else must be awarded by the MathDiss service.

  ▪ Additional classification must be awarded by the requiring library.

Formats

➢ Conform with W3C standards: XML, RDF (+DTD, schemas, namespaces etc.)

➢ Conform with Dublin Core formats for metadata presentation

➢ Conform with NTLTD standards

➢ Provide the data according to the Open Archive Initiative protocol

Workflow

```
                    Graduate
                        |
                        v
                     MMMfT
                   /    |    \
                  v     v     v
            Library    DDB   MathDiss
```

The production and maintenance of the data have to be organized in a way to keep the metadata synchronized.

The Future: Co-operation

- International: NTLTD, OAI,…
- Bilateral: Activities in Austria, Switzerland, France …
- National:
  - DDB, DissOnline,…
  - Conference of Heads of Mathematics Departments, DMV, Math-Net,…
  - Contacts with subject specific activities in other areas (e.g. PhysDis in Physics)
- The goals is to develop a competence centre for metadata for dissertations in general and for mathematical theses and dissertations in particular.