

The T_EXDocC-Project: Three Sources and Three Component Parts

Thomas Fischer, Research & Development, SUB Göttingen

The T_EX Document Center (T_EXDocC) is a co-operative project by the University Duisburg-Essen, Duisburg Campus (the former Gerhard Mercator University) and the State and University Library (SUB) Göttingen. The project is funded by the Deutsche Forschungsgemeinschaft (German Research Foundation) and started the end of 2003. The first phase is planned until the middle of 2005, and continuation beyond that date will be provided by the SUB Göttingen.

Sources

The project is based upon experiences in the context of other projects one or both partners have participated in or collaborated with. There are essentially three sources that lead to the conception of the T_EXDocC project.

ArXiv: Preprint Servers for Physicists and Mathematicians

Since the early 1990s, mathematicians have used *preprint servers* to disseminate new results and stimulate discussion among their peers. After a phase of mushrooming servers, consolidation settled in, and now a small but stable collection of preprint servers provides the community with focused information on their specific interest. One outstanding example is the former Los Alamos (LANL) Server, with started in 1991 with a focus on high energy physics, but holds meanwhile one of the largest mathematics preprint collections as well. This e-Print archive has now moved to Cornell University and is available from <http://arXiv.org/> (not arxiv!).

This system showed a number of important points:

- Electronic processing of mathematical papers is feasible and effective.
- The underlying software (essentially Perl) and document format (mostly T_EX) is stable enough to be used for archiving.
- The mathematical and physics community is willing to share their results with their colleagues, even the T_EX “source code”.

Emani: Archiving Mathematics

With the increase of electronic publishing in general and of e-journals in particular, questions about the availability of these kinds of publications became important. While a library can buy a book, put it on a shelf and keep it available for its customers, access to digital objects is much more volatile. If the server is up and running, if the publisher is keeping up this service at all, is a central question for customers asked to invest significant amounts of money in the acquisition of licenses for electronic publications.

Confronted with questions of this kind, the Springer publishing house teamed up with four scientific libraries world wide to provide solutions for a sustained availability of digital objects (<http://www.emani.org/>). Analysis of the material available showed that most papers were presented as PDF files, but usually produced using T_EX. While the Portable Document Format (PDF) is developed and owned by Adobe Systems Incorporated, T_EX is open source software. So to avoid dependencies upon private corporations, using T_EX as archiving format would be preferable. Furthermore, while the experiences from the ArXiv showed that little adjustment was necessary to keep the files “alive”, changes in the PDF format made some earlier PDF documents unreadable for present day renderers like Acrobat Reader.

MathDiss International: Electronic Theses and Dissertations

Electronic publishing has changed academic habits in another area as well. Many universities started to accept electronic substitutes for the traditional final theses or dissertations required for academic degrees. This yields new questions of authenticity and availability which in the German context have been handled by the DissOnline project. While this provided stable standards for the acceptability and handling of electronic theses, the availability for the scientific community could be improved. Subject specific collections should provide focused access to the relevant theses, using the commonly used classification schemes, integrating with the abstracting and reviewing process used in the community. This was achieved with the MathDiss International project (<http://www.ub.uni-duisburg.de/mathdiss/>), a collection of dissertations in mathematics (or closely related fields), categorized using the standard Mathematics Subject Classification, enhanced with keywords and descriptions.

This project fell somewhat short of the expectations in the quest for T_EX source files of these theses: the process initiated with the DissOnline project tended to yield PDF files, which were regarded easier to handle by the libraries accepting and collecting the theses.

On the other hand, the T_EX documents that were collected showed serious problems: some were incomplete and couldn't be compiled by the T_EX engine, others contained an abundance of files, some unnecessary for the final product.

Conclusions

So while the ArXiv project showed the principal feasibility of archiving T_EX documents and the Emani project the necessity to provide some organized structure for archiving, the MathDiss project revealed some serious problems with archiving T_EX documents. Basically some normalization of the T_EX files was lacking as well as a check of the final product. Furthermore, support was needed to make T_EX documents acceptable to the librarians handling the electronic theses.

Component Parts

As a consequence, the idea of a *competence center* for T_EX documents was born. This center should help people start writing T_EX documents and should lead them towards producing well organized documents that can be safely archived. It should build acceptance for T_EX documents outside the specific T_EX community (primarily mathematicians and physicists), and it should provide for an archiving system that makes it easy for the authors to have their papers archived and to find already available articles.

So the T_EX document center will consist of essentially three component parts.

Support for writing in T_EX

Writing T_EX is essentially very flexible. Everybody can create her or his own macros, making writing (possibly) easier, but making exchange of data cumbersome. Some T_EX “dialects” have evolved that restrict this freedom and provide some “guardrails” for creating documents, most notably L^AT_EX. But still the variation of styles can be very broad, so the first task of the T_EX Document Center is to provide a standard environment for L^AT_EX documents against which the individual documents are measured. This will give the authors a framework document to start writing T_EX and a collection of additional files that are considered to be useful, e.g. for writing T_EX in German. Obviously these would have to be different for Chinese T_EX.

Further help will be given through links to available T_EX distributions and information, to a collection of questions and answers regarding problems handling T_EX. So this part will essentially provide everything needed to create a high quality T_EX document that is easily transferred to some other system and can be archived without any losses in content or presentation.

Validation and Compilation

While the above part will give the means to create high quality documents, this part will ensure and prove that the given documents are actually conformant to the standards set in the first stage.

A *validator* will check submitted T_EX file collections for completeness and redundancy. This will build on the software developed for the ArXiv site and provide appropriate feedback to the author on how to improve the document if validation fails.

If the validation is passed successfully, the document can be compiled, and output in various formats can be produced. While PostScript and DVI are options, the preferred output format most likely will be PDF. Given adherence to the formulated standards, high quality PDF can be produced, including linked tables of contents and indices, interlinked bibliographies etc. This will be highly scalable and fit for reading on-screen as well as for high resolution printing.

We think that this feature will allow “non-T_EXies” like librarians to handle T_EX documents easily. An electronic thesis supplied as a collection of T_EX files will be just transferred to the T_EXDocC server, and the PDF file returned will be used for the printed version and the web presentation of the paper.

Archiving

To go beyond a mere service for producing and handling T_EX documents, the final step will be a facility to archive these documents and make them readily available. Towards this end, an upload for T_EX files will be installed, which will build on the previous stage: no upload without validation. Additional information will be

extracted from the files and fed into a metadata template that will give a basic description and classification of the given document. Nevertheless the author will be required to provide additional personal and bibliographic data, but this should be minimized by the system.

The uploaded document and the provided metadata are the starting point of the full bibliographic description required for serious long term archiving. The T_EXDocC database will provide its own search facilities to make its archive a full-scale server for electronic documents; but a linking to the German reference journal “Zentralblatt der Mathematik” will make the uploaded articles available through the appropriate section of this journal (e.g. preprints) as well.

Outlook

The project is still in its primal stage, gathering available information, hunting for hints and tips, developing the necessary software and starting to build up the server.

For the success of this project, contact and communication with the T_EX community will be of vital importance.

To facilitate this, a Wiki will be installed on the website, creating the opportunity of exchange and discussion. A Wiki (cf. <http://c2.com/cgi/wiki?WikiHistory>) is a system that allows anybody to contribute to a public discussion on the topic given. While in principle this would allow for arbitrary rubbish to be published, experience shows that the system is very stable and can handle deviations from the aims of the website very efficiently. The best known example is the online encyclopedia *Wikipedia* (<http://www.wikipedia.org/>), by now probably the largest encyclopedia available, with generally high quality content. This is an experiment, and we will be watching it closely.

On the other hand, the T_EXDocC does not stand alone in the SUB Göttingen. There is already a plethora of mathematical resources available:

- The SUB holds the central German collection of books in pure mathematics,
- *MathDiss* International Database is the repository of the electronic theses in mathematics,
- *MathGuide* is the subject gateway to quality controlled mathematics websites,
- the *Göttinger Digitalisierungszentrum (GDZ)* holds a vast collection of retro digitized mathematics journals and monographs (over 1 million pages),
- Göttingen provides a mirror to the server of the European Mathematics Union (EMU), is partner in the EU projects Renardus and Euler, starts building a Digitization Registry and is involved in several other activities related to mathematics.

To integrate all the relevant sources, the SUB Göttingen plans to build a Virtual Library for Mathematics. While an independent project in its own right, T_EXDocC will find an appropriate integrating framework in this endeavor.