

TEXDocC: A Service Center for the Use of TEX Documents in Academia and Libraries

Thomas Fischer (SUB Göttingen), Sebastian Pokutta und Günter Törner
(University Duisburg-Essen)

Basic Considerations and Background

The TEX Document Center (TEXDocC)¹ is a co-operative project run by the University Duisburg-Essen, Duisburg Campus (the former Gerhard Mercator University) and the State and University Library (SUB) Göttingen. The goal of the project is to provide services to handle TEX documents efficiently, to support authors in the creation of TEX documents and to improve the quality of the documents for presentation and archiving. The project is funded by the Deutsche Forschungsgemeinschaft (German Research Foundation) and was started at the end of 2003. The first phase is scheduled until the middle of 2005, and continuation beyond that date will be provided by the SUB Göttingen.

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 experiences from three sources that lead to the conception of the TEXDocC project: The ArXiv preprint server (recently moved from Los Alamos to Cornell University), the EMANI project and the MathDiss International 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 interests.

One outstanding example is the former Los Alamos National Laboratories (LANL) Server, which started in 1991 with a focus on high energy physics, but by now holds 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/>.

Experiences with this system proved a number of important points:

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

Furthermore, members of the ArXiv project were willing to co-operate and allow glimpses into the inside of the machinery that runs the ArXiv almost automatically with only minimal human intervention. While this software had grown somewhat unwieldy over the last decade,

¹ <http://www.TeXDocC.org/>

it provided insights into the way T_EX documents can be handled and provided a starting point for the development of a modularized program to handle T_EX documents.

Emani: Archiving Mathematics

With the increase of electronic publishing in general and of e-journals in particular, questions about the long-term 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. Is the server up and running all the time, can the publisher guarantee to keep up this service over extended time periods, these are central questions asked by customers before they are willing 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 and founded the *Electronic Mathematics Archiving Network Initiative* (EMANI, <http://www.emani.org/>). Analysis of the on-line material 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. Using T_EX as archiving format would avoid dependencies upon private corporations and so 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 rendering engines like Acrobat Reader.

On the other hand, closer analysis of editions of exemplary journals showed that T_EX documents in a commercial environment need stable, albeit changing settings for compilation. So the Springer Publishing House, for example, uses a system of stacked style sheets to produce the necessary “look and feel” for their journals, with a general Springer Style at the base and specific style sheets for the different journals and their respective volumes on top of that. In addition, some other styles might be used for different reasons. To provide an archiving environment with long term stability, a system had to be devised that allows for the archiving of the articles as well as the needed style files and incorporates a mechanism that joins the article with its particular styles for later compilation and presentation.

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 (<http://www.dissonline.de/>) 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 in the area, integrated 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 (see <http://www.sub.uni-goettingen.de/ssgfi/mathdiss/>).

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 without major revisions, others contained an abundance of files, some unnecessary for the final product. Different T_EX dialects (basic T_EX, L^AT_EX, enriched with personal styles and macros) based on different distributions from different times proved difficult to be reconciled to one system.

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.

As system was needed that would support the creation and handling of T_EX document without human intervention, since only little man power can be invested in this project over an extended period of time. This system has to be stable and reliable (in particular the long term archiving part of it!) to be accepted as a useful service by writers and librarians. It should be well integrated into the standard workflow and resources of the working mathematicians, like established preprint archives (e.g. ArXiv) and search engines (e.g. EMPRESS²) as well as review journals (e.g. Zentralblatt). And on the other hand, experience showed that the acceptance of such a system would be the higher the less additional input it required from the individual authors, so the need for (repeated) input of data and metadata needed to be minimized, with as many metadata as possible created or extracted from the submitted material automatically.

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 lead them towards producing well organized documents that can safely be archived. It should build acceptance for T_EX documents outside the specific T_EX community (primarily mathematicians and physicists), and provide for a system that makes it easy for the authors to have their papers archived and to find already available articles. This system should be reliable enough to be used as an archive for commercial mathematical publications as well, providing long term access for subscribers to electronic journals and – preferably – the general scientific community.

So the T_EX document center will consist of essentially three component parts: Support for writing in T_EX, machinery for the automated validation and compilation of T_EX documents and an archiving component that solves the particular problems related to large scale preservation of T_EX documents over time.

² <http://mathnet.preprints.org/>

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.

This information section will also offer advice to handle T_EX documents for people who do not want to write T_EX or look into these files at all; in particular, it should allow librarians at university libraries confronted with a thesis written in T_EX to “do the right thing”, which usually shouldn’t be much more than sending the whole package to the T_EX Document Center and let the system handle it.

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. This will be an interactive process, where the author will keep control of the article from start to end, but the mechanism will guarantee a final product that can be compiled and archived safely.

If the validation is successful, the submitter can choose to have the document 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.

Given the availability of the T_EX source documents, other conversions might prove to be useful. In particular, the MathML standard is increasingly supported by modern browsers and provides convenient presentations for mathematics. So the emergence of conversion tools from T_EX to MathML will be closely monitored to be able to provide this additional format as soon as appropriate tools become available (but probably only in a later project).

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 supported by the system to minimize the effort necessary.

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 link 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.

Metadata

The idea of T_EXDocC is to create an archive of T_EX documents that is easily accessible *and* a safe long-term repository at the same time. While this requires an advanced technical infrastructure together with concepts for the secure preservation of the data (back-up, redundancy), the major problems lie elsewhere.

On the one hand, the programmes to compile the documents have to be available over time. In this respect, T_EXDocC relies on the T_EX community to keep the programmes running in the same way they did over the last two decades. As long as T_EX is the major format for writing mathematics, this will be guaranteed by the general usage. And since the whole T_EX project is built on open source software, this process cannot stumble over copyright and patent issues that arise when commercial companies are involved. Even if the machinery will not be readily available at some point in the distant future, the T_EX files still can be read using a simple text editor, and can be interpreted by anybody knowing the basic rules of the language. More likely will be a transition to some other format, probably some mark up like MathML, for the archive before T_EX falls out of use.

On the other hand, the administration of the data over time is a real challenge, the rules of which are still under development. The widely accepted model today is the “Open Archival Information System (OAIS)³”, describing in an abstract fashion the storage and retrieval of digital objects. Up to now, there are only very few concrete realizations of this model, most notably the e-depot by the Koninklijke Bibliotheek, The Hague, built using IBM’s DIAS system (see <http://www.kb.nl/e-depot/>). Göttingen University is partner in a project that plans to set up such a system together with the *Deutsche Bibliothek* and the *Gesellschaft für wissenschaftliche Datenverarbeitung, Göttingen* (GWDG). In any case, this will require a rich set of preservation metadata in addition to the necessary metadata for administration and retrieval.

³ The central document is “Reference Model for an Open Archival Information System (OAIS)”, for now available at <http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf>. The main web site for the OAIS documents (<http://www.rlg.org/longterm/oais.html>) will no longer, be maintained, information should be available at the NASA web site (see <http://ssdoo.gsfc.nasa.gov/nost/isoas/>).

Since the definition of the data model is still emerging, T_EXDocC for now will start with a flexible and extensible metadata set that builds on the Dublin Core Metadata Set (see <http://dublincore.org/documents/dces/>) and the experience in the *MathDiss* project. The richness of the dataset may depend on the type of the given document, a single article requiring only a simple basic set of metadata, while for official documents (e.g. dissertations), a more complex set is needed. Additional technical metadata will keep track of the dates and requirements of the document.

For the individual document, metadata are needed for

- description and discovery of the archived items
- identification of the necessary T_EX environment
- handling of the archived item over time (archival and conservation metadata)

Interlinked with these publication data will be personal and institutional data referring to the author of the document and his or her institutional background. In particular, this will allow authors to enter their personal data only once and reuse them for all further papers. In addition, support for the creation of metadata will be given by extraction of metadata from delivered documents whenever possible and at the same time, technical metadata (for T_EX and archiving purposes) will be created automatically with the delivery of the document. This will possibly be supported by appropriate templates available from T_EXDocC that incorporate metadata in T_EX documents.

T_EXDocC as Mediator

As a general goal, T_EXDocC tries to mediate between different communities that have to deal with the creation and handling of T_EX documents. T_EX documents are *created* usually by active mathematicians or physicists, but the excellent results achieved when using T_EX as a typesetting engine are employed by members of other subject areas as well. For example, T_EX provides a very effective environment for writing music. T_EXDocC will try to cater for their needs, independent of their specific subject area; so although it is conceived in the mathematical context, it will try not to be limited to this particular community and provide templates for any kind of article that is desired.

T_EX documents are *handled* by publishers, librarians and archivists with different purposes. Publishers produce printed output and usually have qualified personnel to transform the document into the appropriate form, if not, T_EXDocC can be used to transform the T_EX document into the desired format, usually some PostScript or PDF. Librarians might want to create an electronic version that can be rendered or downloaded through the internet, as well as printed versions for formal procedures needed e.g. for theses. For the archiving process, the T_EX document has to be preserved with all of its components to keep the document “alive” over time, and enriched with additional preservation metadata. Preferably, this would create an “Archival Information Package” in the sense of the “Open Archiving Initiative System”, to be fed into an appropriate trusted archive.

T_EX itself is a stable, but emerging technology with an active community of developers and users. The ultimate goal of T_EXDocC is to create a forum that allows the exchange of ideas between the people creating or handling T_EX documents and the developers of related tools and the software itself.

Outlook

The project is still in its early 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 is 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. Ideally, the users of the T_EXDoc Center should develop their own online tutorial for the creation and handling of T_EX documents.

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.

The T_EXDocC -Server – features and technical implementation

The T_EXDocC-project is divided into two parts. First the web resources with the user support and second, the T_EXDocC-Server. The T_EXDocC-Server is a core component of the competence center. It provides all the document-processing relevant services such as validation, the extraction of metadata and the preparation for archiving. The ArXiv server inspired this server, but we have slightly different design goals. We want to create a service which is easy to use. This means that an author who is new to the service will be able to publish or archive his or her documents almost instantly. Next, we want a flexible service, which can be easily integrated into a given library infrastructure and hence it is necessary to implement open interfaces. To avoid the use of “yesterdays” technology we decided to implement a completely object oriented framework for the T_EXDocC-Server in Perl. Additionally, we decided to split the T_EXDocC-Server into two components. In fact, this means, that we implement the main functionalities as a “Service” and onto this “Service” we build an “Interface”. The big advantage of this two-component-principle is that almost everybody can build his or her own interface to the T_EXDocC-Server. This ensures that the server can be integrated seamlessly into a given library-system. The “Service” is accessible via some XML formatted requests, all output is XML formatted, too. It provides the validation and archiving services on the one hand,

but manages the metadata on the other hand, too. For every document delivered to the service we store the whole compilation environment to ensure further compilations in the future. The “Interface” encapsulates all the service and integrates them into what is then the T_EXDocC-Server. While developing the “Interface” we decided to use XSLT as “style sheet” language to allow quick and safe modifications of the appearance.

All these abstractions ensure that we will be able to provide a flexible “Service”. The use of XML-techniques allows us to implement some useful features like MSC-based RSS-feeds, easy synchronization between different instances of the T_EXDocC-Server, or eventually a SOAP-Interface. Additionally, we provide an OAI Interface.

What we have done already

The implementation of the backend including the databases is completed. All core services are fully functional. This includes the validation and the storage engine, as well as submitting documents and retrieving them later. Actually we are working on the XML Interfaces and discuss some of the needed protocols.