LIBER QUARTERLY, ISSN 1435-5205

K.G. Saur, Munich. Printed in Germany

Electronic Theses: Swiss Perspectives

by ULRICH NIEDERER, ULRICH WEIGEL, MARIE-PIERRE GILLIERON-GARBER & KARL BÖHLER

INTRODUCTION

To offer theses electronically to its users is something that has been discussed in Swiss university libraries for some time now. There are two main reasons that it has not yet been realized anywhere in Switzerland: On the one hand, the two large university library networks have had big library system migrations to cope with (first the „Réseau Romand" in the French speaking part moved to VTLS, then, in the last two years, the „Informationsverbund Deutschschweiz, IDS" in the German speaking part switched to Aleph by Ex Libris). These migrations used up all available personnel resources. On the other hand, the federalist organization of all things politic in Switzerland, including the university system, means that there are hardly any national programmes to further and finance projects on a national level (or cantonal, for that matter) - in contrast to most other European countries where such programmes have proved to be a very strong incentive for developing various aspects of the digital library.

The discussions in the universities on electronic theses revealed that the existing regulations on publishing theses are widely different even between departments of the universities. Moreover, a thesis printed and published by a prestigious publisher is still highly valued in some academic fields (like, e.g., literary criticism, philosophy etc.), whereas in others the speed of publication is far more important than its form.

Nevertheless, various universities and their libraries begin to show a strong will to publish and archive theses electronically. The following papers present three projects that deal with electronic theses in different ways. They also show that there is still some hope that one day we will find a way to a common solution.

Ulrich Niederer
Zentral- und Hochschulbibliothek Luzern
Sempacherstrasse 10
CH - 6002 Luzern
niederer@zhbluzern.ch

Electronic Theses at the University of St. Gall - EDIS

The possibilities of electronic publishing have become significantly better throughout the last years. On the one hand, more powerful communication technology allows downloading even of large files in an acceptable amount of time. On the other hand, there are nowadays software instruments with which digital documents can be produced fairly easily.

Against this technological background we decided about a year ago to develop a prototype for the electronic publication of doctoral theses at the University of St. Gall. We wanted to solve the storage space problem as well as give fast and cost-effective access to research results. We looked for a solution that does not cause additional costs for the library, but does neither demand any considerable amount of work from the doctoral candidates.

The library and the computing department cooperated in developing EDIS – Electronic Dissertations. As the University of St. Gall uses „Lotus Notes" as its communications platform, we decided to use it to program a prototype. Thus it was neither necessary to acquire additional hardware nor to get acquainted with new application software.

The central components of EDIS are:

a) a module to publish electronic theses
   - a manual for producing a PDF file
   - a help file for filling in a metadata form
   - a help file for formatting the thesis properly

b) a component for indexing PDF files

c) a search module which has the following options
   - full text search for the whole file
   - search for the identification number
   - browse various categories (author, title etc.)
   - further search options are currently being prepared.

d) the software Acrobat Distiller which is needed in order to produce a PDF file

e) information on the EDIS project (basic data, help, short description etc.)

Work is organized in a way that it can be done by the doctoral candidates themselves. The single steps are independent of location. The first step consists of converting a text file (usually Word or WordPerfect) to a Postscript file that is then changed to a PDF file with Acrobat Distiller. This process does not take more than 15 minutes for a normal text file. In order to guarantee the identity of paper and electronic version you have to make sure that no „cut and paste"-operation has been executed. The doctoral candidates also have to certify the authenticity of the electronic document.

The next step includes the metadata description of the thesis by the candidate. The form has two main categories:

- data specific to the thesis, i.e. the bibliographic data (author, title etc.), subject data (abstract, keyword) and administrative data (thesis supervisor, expert, date);
- formal data (thesis language, format etc)

Filling in the metadata form probably causes more trouble, as students should find appropriate key words from an authority list; they also have to write an abstract that makes sense.

The coding of the metadata is done by the system according to the internationally renowned Dublin Core Standard, which is also used in numerous other electronic dissertations projects for the description of the files.

The thesis is attached to the metadata file. The whole file is then sent to our theses server. In addition we plan to archive the original versions.

The next step consists of the complete indexing of the PDF document and of the metadata file; it is done by the indexing module. That gives us the prerequisite for searches with various criteria. The results first show the findings from the metadata files so that there is no downloading of large PDF files which turn out not to be what you looked for.

We should perhaps mention that electronic theses do not infringe on the author's copyright. He is and remains the author of the published work. The library should therefore ask for the author's written approval of electronically publishing and archiving his or her thesis.

Electronic theses are still new in the libraries of all German speaking countries, even if Die Deutsche Bibliothek has already received 660 electronic theses. But the degree regulations have yet to be adapted in a great many universities (also in St. Gall!), and the doctoral candidates should be made aware of the new publishing possibilities in a better way. The significant reduction of the cost of publication is not the least important factor in this project, but the advantages of going digital are not yet known well enough. With our prototype we would like to give a new framework for setting up (electronic) theses to the University of St. Gall and its doctoral candidates.

Dr. Ulrich Weigel
Universität St. Gallen
Bibliothek
Dufourstrasse 50
CH – St. Gallen
ulrich.weigel@biblio.unisg.ch

THÈSES ÉLECTRONIQUES À L'UNIVERSITÉ DE GENÈVE

La diffusion des thèses en mode électronique à l'Université de Genève, en même temps que deux autres projets ayant trait aux nouvelles technologies, a été proposée en décembre 1998 par trois bibliothécaires, suite à leur participation à l' „International Summer School on the Digital Library" Tilburg (NL) l'été précédent. La Commission des bibliothèques ayant sélectionné cette option, un premier plan de travail a été préparé durant le premier trimestre 1999.

Ce projet a été conçu pour se développer et se mettre en place sur une année et demie. Une structure de coordination sera chargée de son démarrage et de son suivi et il est envisagé de nommer un groupe de pilotage, nécessaire au niveau politique et structurel. De petits groupes de travail intermittents, selon les options spécifiques à décider, seront organisés en utilisant des personnes compétentes dans les différents domaines, ces groupes ayant une vocation essentiellement pratique et opérationnelle. La mise en oeuvre de ce projet a été planifiée en quatre phases, certains aspects se déroulant tout au long des quatre étapes, d'autres se déployant sur une ou deux phases spécifiques.

Le montage a été soumis au Rectorat en mai 1999, mais ce dernier a souhaité, avant de s'engager plus avant, des précisions quant au droit d'auteur et à l'imprimatur, points qui faisaient partie de la phase d'étude du projet. Pour des raisons de manque de moyens financiers, ce développement n'a pu être réalisé dans la foulée de la décision du Rectorat et n'a été mis en route que depuis mi-janvier 2000.

Les objectifs de ce projet sont :

- de mettre à disposition des usagers les nouvelles thèses de l'Université de Genève sur le Web, afin de valoriser le travail de production scientifique de l'institution,
- d'augmenter l'accessibilité aux travaux scientifiques et d'en offrir la conservation électronique,
- de diffuser de l'information de qualité sur Internet, d'en offrir l'accès à un maximum de personnes et d'offrir aux étudiants la diffusion de leur texte,
- de compléter la formation universitaire des étudiants par le biais de la promotion de la production de documents électroniques, de l'appréhension des problèmes liés à la publication et de la sensibilisation aux problèmes du droit d'auteur.

La structure de coordination prévue sera chargée de la promotion du projet durant son implantation et de l'information auprès des utilisateurs (information au cours des différentes étapes, articles dans les milieux professionnels). A la fin, une évaluation sera probablement faite, tant sur le plan qualitatif que quantitatif, par le biais d'enquêtes et de production de statistiques d'utilisation. Cette structure pourra se dissoudre lorsque l'organisation définitive et la répartition des responsabilités de fonctionnement du système auront été mises en place.

Les autres tâches, toutes d'égale importance et en interdépendance malgré des fonctions très différentes, sont de quatre ordres: définition d'un cadre juridique, choix des options informatiques, définition du rôle et de l'intervention des bibliothèques et précision des rôles au niveau académique.

Concernant le cadre juridique, il faut très clairement définir la notion de propriétaire de copyrights, établir des conditions et des règles juridiques pour la protection des données, que ce soit au niveau de l'auteur, de la faculté ou de l'université; garantir l'authenticité du document et se prémunir contre le plagiat. Des informations juridiques utiles aux usagers quant à la consultation et à la reproduction des documents devront être élaborées.

Les options informatiques tournent autour du choix ou de la désignation d'un serveur, d'un ou de plusieurs formats numériques et de logiciels bureautiques, en tenant compte de standards compatibles avec des développements futurs. Il faudra régler l'organisation et les détails concernant l'archivage des données dans le but d'en conserver l'intégrité et d'en garantir l'authenticité. Tous les aspects liés à la sécurité des données sont à prendre en considération et des outils de recherche qui permettent la recherche en texte intégral devraient être mis à disposition. L'organisation de la maintenance du serveur est à prévoir, de même que la conception d'une feuille de style pour guider les auteurs des thèses, sans omettre d'organiser une formation à l'utilisation d'outils de travail conviviaux.

Dans le cadre du système décentralisé de l'Université de Genève, le rôle et l'intervention des bibliothèques sont à préciser, en ce qui concerne l'instauration de normes d'édition, la production d'un cadre pour le précatalogage et l'utilisation de metadata afin que les auteurs soient à même de précataloguer et indexer leurs thèses eux-mêmes. Les bibliothèques seraient chargées de l'encouragement auprès des auteurs, éventuellement en parallèle à la formation à l'édition électronique, car il est indispensable de préparer un cadre et des marches à suivre pour la rédaction des thèses électroniques et de garantir aux doctorants la formation aux nouvelles technologies et la mise à disposition des logiciels nécessaires. Ces options nécessitent une réflexion bibliothéconomique préalable, mais une fois mises en place elles ne demandent qu'une réévaluation et une remise à niveau régulière. Ce serait une tâche nouvelle, mais tout à fait liée aux tâches traditionnelles des bibliothèques. Il faut aussi envisager la validation par la bibliothèque du précatalogage et des metadata en apposant un label de qualité au document électronique. De même que les instances académiques valident le contenu de la thèse, les bibliothèques ont leur rôle à jouer dans l'authentification de sa signalisation et de sa conformité aux normes internationales. Les bibliothèques seront aussi en charge de l'intégration des données au catalogue RERO (Réseau des bibliothèques de Suisse occidentale) et de la mise en place d'un lien pour l'accès au texte intégral de la thèse, de l'organisation et de la gestion du dépôt légal et de l'archivage au niveau bibliothéconomique, de la mise en place de la diffusion de l'information et des accès par le biais d'un site Web, afin de permettre une harmonisation de l'accès aux thèses. Pour concrétiser ce changement de pratique et de politique, un service d'aide aux auteurs doit être créé et il faut porter un soin particulier à la communication.

Une analyse du circuit actuel des thèses à l'Université de Genève en vue d'y ajouter les éléments liés à une édition électronique est nécessaire en ce qui concerne peut-être une éventuelle centralisation de ce circuit. Ces points et responsabilités seront arrêtés en collaboration avec les professeurs, doctorants, techniciens (facultés et services administratifs).

Une attention particulière, en tenant compte du contexte suisse, devra être portée aux aspects juridiques, en laissant un maximum de souplesse possible et aux choix informatiques, en privilégiant le long terme et l'établissement d'une chaîne informatique cohérente du document. Il faudra bien évaluer l'intervention et la responsabilité des bibliothèques suisses, au niveau de la signalisation de la thèse, de la responsabilité de l'archivage et de la promotion des nouveaux services offerts. Etant donné le nombre d'instances et de personnes concernées au niveau de l'université il est important de veiller à bien atteindre les facultés et les services administratifs et d'offrir une information de qualité. Tout un chemin de la thèse électronique est à élaborer et il est souhaitable de viser dès l'introduction de la publication électronique à diminuer les versions papier à un minimum (par exemple trois exemplaires).

Etant donné que les solutions techniques à choisir pourraient être sensiblement les mêmes, il devrait être possible pour les universités et les bibliothèques de travailler en commun au niveau romand, voire national et pourquoi pas international en se joignant à un autre projet ou initiative? Dans cet esprit des contacts ont déjà été pris au niveau suisse et avec quelques universités à l'étranger. D'autre part il importe de bénéficier d'un soutien officiel au projet afin de pouvoir être actifs à tous les niveaux.

Marie-Pierre Gillieron-Graber
Site Web et Matières
SEBIB / Uni Dufour
24, rue Général-Dufour
CH - 1211 Genève 4
Marie-Pierre.Gillieron@adm.unige.ch

ULRICH NIEDERER, ULRICH WEIGEL,
MARIE-PIERRE GILLIERON-GARBER & KARL BÖHLER

DIGITAL DISSERTATIONS AT THE ETH-BIBLIOTHEK ZURICH

1. *Initial Situation*

In 1999 there were approximately 11,600 students enrolled at the ETH (Swiss Federal Institute of Technology) Zurich and of these about 2,000 were doctoral students. Of the latter 400-500 complete their studies each year with the publication of a doctoral dissertation. The rules of the ETH require them to hand in to the ETH-Bibliothek only four copies of their dissertation (one in loose-leaf form) and with this their obligation to publish is fullfilled. The authors are free to decide whether and how they are going to publish additional copies of the dissertation. Up to the present time the bound copies are catalogued and shelved conventionally. Microfiche copies are made from the loose-leaf version for exchange purposes and this version is then archived as a reference copy.

A few years ago we saw the first dissertations on CD-ROM. Right from the start we acquired dissertations for subsequent loan, e.g. from UMI (now Bell+Howell) by using their facility for downloading them in PDF format. It was therefore logical for us to consider something similar for our dissertations at the ETH. Occasionally we have been asked whether we could accept ETH dissertations produced only on CD-ROM, or even as an HTML file, as has been usual at some American universities for years and where it is sometimes even obligatory. Considering the possibilities of publication in multi-media form (videos, 3-D simulations, sound tracks, animations, programs, hyperlinks etc.) and the internet, the publication of dissertations in a modern digital form is being increasingly called for, naturally in full text, browsable and on the internet.

We therefore made a systematic analysis of the available digital dissertations on the web and tested a few relevant softwares. At the same time we also asked ourselves how the average user outside the circle of technical enthusiasts would accept dissertations in digital form.

2. *Empirical Analysis of the Characteristics of Digital Publications*

Publishing documents in electronic form undeniably opens up new dimensions. Nevertheless the experience of the last few years shows that many developments, instead of being in advance of their time, were not focussed on the demand. The ETH-Bibliothek has been collecting digital documents since the nineteen-eighties and can report not only interesting but also negative experiences, especially concerning data preservation and accessibility of these software-dependent documents. After the initial euphoria over the new ways in which documents could be published we questioned pragmatically the practicability and the consequences. The following aspects are not meant to be a negative catalogue but indicate points that we wanted, from past experience, to clarify more exactly. Besides, many things work wonderfully at an experimental level; small blemishes are tolerable. But when large quantities are involved (and such are to be expected) these small blemishes become significant problems.

Standards:

We assume that also in the electronic age the ETH-Bibliothek will continue to be responsible for the central cataloguing of the ETH's own dissertations. However, the library cannot be expected to take on the task of converting the many and varied softwares used into a single format. There is also little point in insisting on the use throughout the ETH of a complicated text processing system (e.g. LaTex) specially conceived for chemists and mathematicians with their formulas and special signs. The many doctoral students who can manage perfectly well with text, illustrations and tables would find such a specialized system irritating.

The available options for electronic publication are also not unproblematic from the user's point of view, as experience with CD-ROM documents has clearly shown. Quite apart from the versions made by private firms, differing releases and Macintosh- or Windows-compatibility, it is mostly not realized that text (namely with MS Word) is by no means everywhere the same, because the formatting is related to the printer used by the author. When the CD-ROM is loaded into a PC with a printer having different settings then the formatting of the pages changes; titles, columns, illustrations, fonts etc. can be moved or even grouped differently – absolutely unacceptable for an academic publication. Another point is, that one language may not be everywhere the same. The German speaking part of Switzerland, for example, uses different keyboard standards so that with the formatting of texts from the Federal Republic of Germany a lot of wrong signs and functions are generated (and of course the reverse is also true).

A further condition is the familiarity of the reader with the text system used. We must assume that in a few years time MS Word, PostScript, PDF, LaTEx, HTML, XML etc. will also no longer be used. Anyone who has been obliged to work with older documents in MS Word 4, WP 3, WordStar, Volkswriter and others after an interval of several years can appreciate this consideration.

If one is no longer familiar with the keyboard- and other macro-commands that were once learned by heart, one will hardly be able to open the document due to lack of knowledge of the commands. We only install old softwares in extreme emergency (perhaps on old PCs, as many do not run at all on PCs with Windows 95 or NT). Nor do emulation programs help as they logically also simulate the old forgotten commands.

We have had rather negative experiences with multiple reformatting, which will certainly be necessary in a few years time. For example it is claimed that the frequently mentioned SGML can be relatively easily transformed into HTML and in a further step reformatted into PDF. Our own experiments, however, showed clearly that the fidelity to the original suffers. Illustrations and tables cannot be decoded. Words or even whole passages of text are changed according to the internal dictionary; moreover links are lost. Test runs have shown that it is much simpler and faster to scan in the unbound paper reference copy and process it directly into PDF. A further aspect from the user's point of view: A voluminous document structured by HTML is certainly impressive, but hardly anyone wants to read it. Various test persons were rather unanimous in their preference for the PDF version, which allows one to read, browse and print out page by page in the familiar way.

Full Text Search:

A strong argument for electronic publications is full text searching. For documents with standardized formatting this is not a problem, for those with mixed formatting it is fraught with complications. Moreover one may pose the quite heretical question: what is really gained? With a relatively limited test quantity one is pleased with each hit, with several hundred or even thousand documents one does acheive a respectable number of hits, but there is also much empty information noise. One would therefore have to attach in addition an intelligent search engine, the care of which would require new qualified personnel.

Citability:

From the scientific point of view it continues to be necessary to guarantee the citability of electronic publications and their content. It can become complicated, for at present the pages are still primarily cited, a criterion that no longer applies with electronic publications, when browser and editor provide completely different representations. Moreover we learned from colleagues in other libraries that printed and electronic versions are often markedly different. Therefore several authors who have written on this subject do not exclude the holding of a paper copy as a sort of „master reference".

Storing, Shelving, Archiving:
It is well known that electronic storage space is by no means inexpensive. The greater the quantity of data, the more its storage will cost. The advantage of full text searching is only valid when all the data is held in an active data bank. Therefore at the present time only servers can be used and not external fixed storage devices (tapes, discs). Digitized versions eat up a lot of time and storage space in relation to the quantity of data. On a CD-ROM (650 MB) there is space for about 38 dissertations in PDF format. Our dissertations for the year 1999, with so far 325 titles scanned in, take up altogether (i.e. including master files) 17 GB. One must also bear in mind that especially with dissertations a few „top" titles are frequently requested whilst most of the others „gather dust" undisturbed. In conventional book stacks this is less critical than on a server, which constantly has to carry the entire holdings as ballast.

Seen from a management point of view, conventional book storage units, microfiche- and disc-archives cost money, but the expenditure is certainly considerably less than the cost for EDP specialists. Only a few people in our library are involved in the handling of ETH dissertations. The results of informal exchange of information lead us to conclude that by offering dissertations in digitized form we could scarcely achieve a saving of a single member of staff; on the contrary the need for EDP personnel will increase. Considering the fact that the data must constantly be maintained and converted, a future reduction of staff on the EDP side is unlikely.

3. Present Practice

The reference copy is and remains the printed version, which the author must hand in to the vice-chancellor's office. Such a copy is unchangable, for the award of a doctorate is based on a fixed text at a moment in time and not on a possible dynamic document containing unobtrusive improvements. The supposed advantages of a completely electronic reference copy are, looked at more closely, not very convincing. No one was able to give us a plausible assurance that he would never have occasion to print out parts of his dissertation – therefore he could also make a paper version; if necessary, as is already done, with multi-media supplements (video, CD, DVD). Naturally he is free nevertheless to provide us in addition to the paper version a complete version in any kind of electronic form.

Before we send the loose-leaf reference copy to an outside source to be filmed we scan it in (form feed scanner Canon 3020, resolution 300 dpi; a colour scanner is at present not authorized because of the cost).

The individual images come in PCX format on MOD (magneto-optical disks), likewise the PDF versions created by Adobe Acrobat Capture (image and hidden text). The latter are copied for use on CD-R, the MODs go into the archive as master files. If new formatting should become necessary (e.g. replacement of the PDF standard in a few years time) the PCX files can be drawn upon.

In our library system Aleph 500, the PDF version is an order option in the catalogue entry for the dissertation.The digital version is recognizable in the catalogue, apart from the call number „DISK...", by the technical note „Format PDF; [nn] MB". The first experiments made in fulfilling orders for digital documents online were disappointing: all the files are several MB large and were often automatically blocked by many webmasters at the receiving end. On the other hand sending them on loan as CD-R (multisession) has proved satisfactory. The discs, which can be reused many times, are for copyright reasons not sold but lent and they must be returned.

4. The Next Steps

The PDF format used permits a usable but not very precise word search. For the reasons already given we are not considering full text searching of the entire holdings of digitalized dissertations. We are however creating a data bank of summaries.These are, by definition, a distillation of the work itself. It is a rule at the ETH that every dissertation must have an abstract in English and in one of the Swiss official languages (German, French or Italian).

In a special process the PCX images of the abstracts are transformed into ASCII-files using OCR (ABBYY FineReader). If necessary subject specialists convert special signs, formulas or index numbers, which have not been correctly recognized, into defined search values. At the moment we are still testing links from the catalogue entry for the dissertation to the full text of the abstracts. Moreover, all the abstracts will be cumulated into a general folder which serves as a special option for full text searching in the summaries of all

ETH dissertations. By means of links one returns from the list of hits to the alphabetical catalogue.

5. Conclusion

The ETH-Bibliothek has decided on a very pragmatic solution, which in our opinion meets the requirements of a wide circle of users. In any case the acceptance has been good. Spot checks have shown that despite the availability of a paper version the PDF version is frequently ordered. The technical outlay is small. All stages of the work could be integrated into existing routines without a need for extra staff. We are eagerly waiting to see how the planned full text searching of abstracts will be received by the public.

Dr. Karl Böhler
Digital Media and Data Preservation
ETH-Bibliothek
Rämistrasse 101
CH-8092 Zürich, Switzerland
boehler@library.ethz.ch

Translation: Michael Beckett