Digital archiving of scientific information – Czech experience

P. Slavik, P. Mach, M. Snorek
Czech Technical University in Prague
Prague, Czech Republic
Slavik|mach|snorek@fel.cvut.cz

Abstract
This paper deals with description of activities in the Czech Republic related to digital archiving. First of all the general situation in the field is described in order to give insight in the state of art in the field in the Czech Republic. The key part of this paper deals with description of design and implementation of a pilot system that should serve for digital archiving of scientific information of certain kind – MSc and PhD theses at Czech Technical University in Prague. One of reasons for archiving of this type of information was the fact that these theses contain information about scientific and technological developments in given period of time. Such an information might be widely appreciated in future by historians who will investigate the history of science and technology of certain period of time. Another important reason for the research performed was the fact that there was an urgent need for practical experiments accompanying thus theoretical research in the field of digital archiving performed at CTU Prague as described below. The approach described is a small scale solution – it allows to solve the urgent needs for digital archiving in small scale before a large scale general solution will appear on the market.

1. Digital archiving in the Czech Republic

Classical archiving as a scientific discipline on one hand and routine activity on the other hand has a long tradition in the Czech Republic. Everyday practice of document archiving runs according to schemes that do not differ significantly from other countries. There exist many archives of various types for different purposes where documents in classical form – this means in paper form – are archived. This situation has changed with the use of computers when documents in digital form became more and more frequent.

Because of historical reasons there was certain technological delay in the Czech Republic from the point of view of massive use of personal computers caused by embargo during cold war etc. This means that the problem of archiving of digital documents was not urgent about fifteen years ago because number of users who could massively produce this type of documents was not so high (in comparison with technologically advanced countries like USA, UK etc.). In the course of nineties the problem of archiving of digital documents was identified and the first activities of this type emerged.

The first activities concentrated at first on creation of database applications working with data about records and documents. Such database systems allowed the users to search documents very quickly not only in particular archive but in a set of archives in general. The introduction of these database systems had very important aspect – archivists got acquainted with computers what eased the next steps that dealt with archiving of digital documents. The next activity that dealt both with archiving and with digital archiving was a number of projects where historical documents existing in paper form were digitized and stored in electronic form (mostly on CD media).

There are research activities in the Czech Republic that are linked up with problems of digital libraries. In the framework of this research also some archiving issues dealing with archiving in libraries (e.g. electronic journals) have been investigated. Besides these activities also problems of archiving of web on national level are investigated.

The problems linked up with digital archiving became subject of interest relatively recently. The need for long-term preservation and accessing the documents that were originally created in digital form came few years ago. This fact resulted from the increasing number of governmental offices and institutions that started to generate documents in digital form. The Ministry of Informatics of the Czech Republic defined recently a policy of circulation of digital documents between central institutions of the Czech Republic. This effort is a part of activities related to introduction of e-government in the Czech Republic. It is obvious that e-government cannot exist without proper strategy for digital archiving.

Till now there is no official institution in the Czech Republic that could be considered as a digital archive. This situation is potentially dangerous as there is a serious threat that some documents in digital form of high importance could be lost. The Ministry of Interior is the institution under which jurisdiction all archives in the Czech Republic belong. The ministry has allocated in years 2001-2002 funding for research project the result of which should have been a national strategy for digital archiving. The research was carried out by a team formed by specialists from Czech Technical University in Prague and by specialists from the Central State Archives in Prague.

The aim of the research was to map situation in governmental institutions (the amount of digital documents produced etc.), to gather the data about the methods used for digital archiving abroad, to describe and evaluate potential suitable methods of long-term preservation of records and documents in digital form. One of expected outcomes of this project was definition of a workplace located in the near future in State Archive where the digital documents will be archived.

The solution suggested has been based on existing international standards like OAIS and others. Due to the fact that the planned workplace will have rather limited extent (about four people will work there in the first stage of this pilot project) it was necessary to adopt some limitations. The idea is that in the course of the time the structure of the workplace and the number of staff will be gradually extended [1].

2. Motivation for specific activities

In the text above national activities in the field of digital archiving were described. It is obvious that digital archiving represent many challenges both from the point of everyday use and from the point of research. Czech Technical University in Prague who participated in this research decided to launch another project dealing with digital archiving. The main motivation for this project was the need to have some kind of test bed where some approaches could be tested on real problems. As stated above the designed workplace is of general nature. In accordance with experience gained from materials and visits abroad it is obvious that there are many applications where some specific solution should be found where specific features of particular applications should be taken into account.

Such a specific application in university environment is archiving of MSc and PhD theses. Such a system will allow us to perform experiments of various types on sufficiently large set of data. On the other hand such a system will solve an urgent problem of archiving of these theses. Also in this case one of the results of this project will be an experimental workplace where certain procedures and schemes will be examined.

There are several solutions abroad that deal with theses archiving. These solutions handle this problem as a topic that is solved as a special activity in the framework of digital library [2], [3]. This requires (besides other problems) consideration of library environment and thus integration of theses archiving into the organization scheme of the library as a whole. This approach requires in many cases also additional staff and some other expenses.

One of important motivations for this project was the fact that MSc theses at CTU Prague in general are not subject for archiving (PhD theses are archived). From the historical point of view these theses mirror the state of art in various scientific and technical disciplines. They might be in future very valuable source of information for historians that will investigate various topics in technology development. That is why this type of information should be preserved (long-time preservation).

Our approach has been based on the assumption that the archiving system will be stand alone one and no extra effort with exploitation of the system will be required. From the research point of view the main attention has been paid to establishing redundancy of the information archived. This means that the variety of data stored should guarantee in the future access to these documents. Having collection of documents of this type it will be possible to perform experiments and tests by means of which it will be possible to establish relevance of the methods used for archiving.

The parameters for the solution are as follows:

- information should be stored in such a form that could be later transferable in more general digital archive (should allow document migration)
- the effort for inserting documents in the archive should be minimal (students should be able to insert their theses by themselves)
- the security issues (the text should not be modified or stolen) should be solved on proper level
- searching in the archive should be easy
- the system could be used as a sort of test bed for archiving methods developed (without any threat to the contents of documents stored).

It is obvious from the above listed parameters that the system will be relatively simple without an extensive demand for additional staff. The research has been concentrated on suitability of various formats for archiving. It is necessary to stress that only the thesis will be stored. It is not assumed that e.g. executable computer programs will be archived.

The key decision was selection of formats in which the documents will be stored. There exist a lot of various recommendations which format should be used. The recommendation is highly dependent on the content of the document. In our case we will store the documents that will contain text and images. In such a case general recommendation is mostly targeted to Adobe PDF format. In our case a very important role will play the fact that there is no problem to convert theses from traditionally used formats into PDF. At Czech Technical

University the theses are written by using of LaTeX or Open Office.org or Microsoft Office tools. According to various statements the PDF is a prospective format and its use in the field of archiving is usually recommended.

3. Implementation issues

Besides the document itself it is necessary to store also metadata. In our case the metadata serve for identification of a student who is the author of the thesis. Most of these data will be automatically generated by means of university information system. The information is stored in the form defined by Berkeley Database Version 2. The main advantage is easy readability and easy migration of metadata. This form also allows us to convert the information stored in XML format that might be in case of necessity transformed into HTML page in case that the info should be available by means of web.

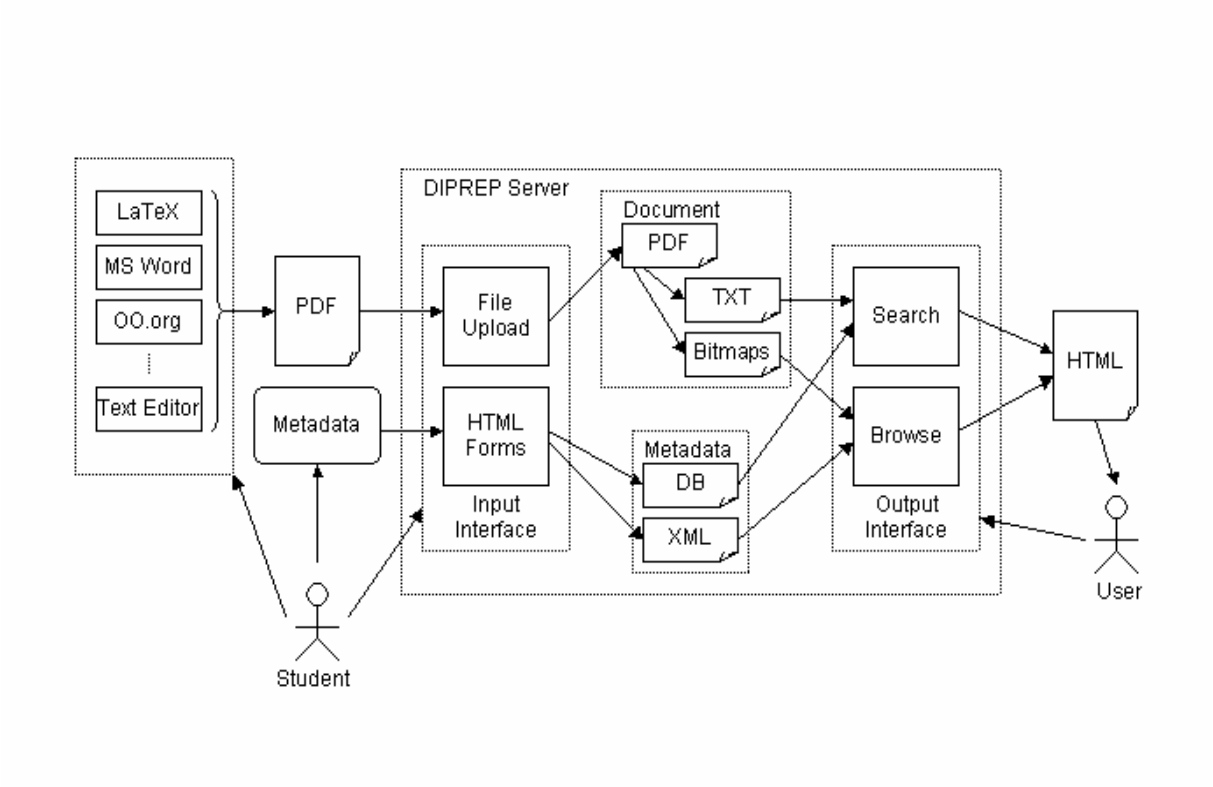

Fig. 1 Document Life Cycle in DIPREP System

In order to avoid problems with document migration in the future it is necessary to archive the documents with some degree of redundancy. One possible form is text file where only textual information is stored (with info about existence of figures). This form also allows us to search in an easy way (from implementation point of view) keywords and other textual information in theses. This file can be easily generated from PDF format. The next format in which the document is stored is a bitmap. This format is very simple so in case that during migration some information was lost (or in case of some disaster in future) we will still have the full information about the particular thesis. Also this form can be easily generated from PDF format.

The document (thesis) preservation is based on migration. Due to simple structure of the system it is relatively easy to take care about migration. As we use three separate forms of documents it is necessary to convert these forms into forms in a new (updated) format. The simplicity of formats might allow in future to create relatively simple readers in case when emulation will be widely used. As for the migration of metadata – also here there are no fundamental problems. The database used is frequently used what gives some hope about continuity of the future usage. In case of some potential problems of this kind we have the second format (XML based) that can be with a small effort transformed into desired metadata format for future use.

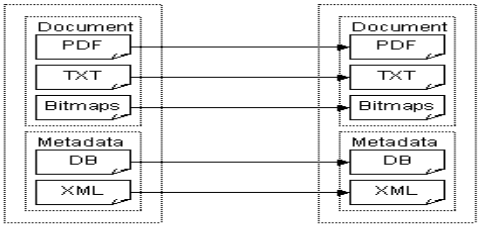

Fig. 2 Migration scheme in DIPREP System

Specific scheme for the access to the information stored has been designed (and partly implemented). The access will be only the local one in university library on dedicated computers only. In such a way we could avoid unauthorized copying of documents stored. Moreover special structure of displayed information was designed where besides other measure a watermark like info with information about the origin of the document is displayed.

4. Conclusion and future work

There already exists a pilot implementation of the system described. The first practical experience has been gained in June 2004 when students submit their theses for defending. The number of students participating in the experiment was about 30. It has been practically proved that the system is easy to use (no student had significant problems with electronic submission into system). Creation of formats that had to be stored was also without any problem. The same holds for creation of metadata. Nevertheless some organizational issues are still pending as well as some improvements in physical data organization. The system has been designed as a “cheap” one without any significant expenses connected with the system exploitation. Also during the implementation of the system the criteria for low costs were considered – the software components used are of open source type. This means that no costs connected with software licenses etc. emerge. This “low cost” solution together with redundancy of the data archived could lead to the application of the approach presented also in other applications where there is an urgent need for archiving of digital documents. The generality of the format used will allow the user to transfer the archived documents later to the more general system used for digital data archiving.

The future work will be oriented towards the development of the procedure that will allow the more flexible access to the information stored. This part of the system should be developed in its final form where security and copyright issues will be handled in a complex way. The tests dealing both with security issues and with archiving methods will be developed and performed.

References:

[1] Mach P., Snorek,M., Slavík P. : Preservation and accessing documents in digital form in long time perspective (Research report – in Czech), CTU in Prague, Praha, 2002. 220 pp.

[2] ProQuest - http://www.il.proquest.com/division/pr/03/20031006.shtml

[3] Australian Digital Thesis Group:

http://ausweb.scu.edu.au/aw02/papers/edited/borchert/paper.htm