DIGITAL ARCHIVING OF SPECIFIC SCIENTIFIC INFORMATION IN THE CZECH REPUBLIC

P. Slavik*, P. Mach, M. Snorek and J. Koutnik

** Dept. of Computer Science and Engineering, Czech Technical University in Prague, Karlovo nam 13., 121 35 Praha 2, Czech Republic*
Email: slavik|mach|snorek|koutnij@fel.cvut.cz

ABSTRACT

This paper deals with a 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 a description of the 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 the reasons for archiving of this type of information was the fact that these theses contain information about scientific and technological developments in a given period of time. Such information might be widely appreciated in future by historians who will investigate the history of science and technology of a certain period of time. The research is oriented towards robust archiving systems that can be used in small-scale applications. These small systems do not offer universal solutions in the field of digital archiving – they solve problems that become urgent in various applications: to save current digital documents in the form that could be transferable to general archiving systems developed later. The described implementation is a pilot practical solution to this problem. The approach described in the paper will allow the user to archive also documents that contain non-textual information.

Keywords: Digital archiving, Thesis, Long term preservation

1 DIGITAL ARCHIVING IN THE CZECH REPUBLIC

Classical archiving as a scientific discipline on one hand and routine archiving 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 exist 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 have been investigated. Besides these activities also problems of archiving of web on national level have been 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 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.

Up to 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. Besides study of available information about digital archiving round the world also series of visits in institutions abroad where digital archiving is routinely performed was realized. These visits included UK, Sweden and The Netherlands.

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 (Mach , Snorek and Slavik, 2002).

The formation of this workplace dealing with long-term preservation of records and documents in digital form will be of a high importance – but not only from the archiving point of view. This pilot project will undoubtedly bring an extensive knowledge important both for the community of users and for community of researchers in the field.

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 represents 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 for digital archiving on national level is used for documents coming from various sources. 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 these theses. Also in this particular case one of the results expected from this project will be an experimental workplace where certain procedures and schemes will be examined.

There exist 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 (ProQuest 2003, Borchert 2002). This requires (besides other issues) 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 mostly 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. 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). Moreover the traditional motivation – wider access to result of student works should be preserved.

Making theses electronically accessible in future is not an unknown problem. There are several software tools developed for this task. Namely EPrints (EPrints software, 2005) which fulfils a task of self-archiving where students submit their work. EPrints provides access to the theses but the software itself is most likely a digital

library than a digital archive. It is focused on easy access to the theses but the long term preservation and readability of the document has not been solved properly. For example it allows to store Microsoft Word documents which are not suitable for long term archiving.

The CTU as an institution produces several hundreds theses per academic year so we can take it as a mid range institution. Institutional repositories are discussed in (Hey, Hitchcock, Carr & Brody, 2005) mostly from the document accessibility point of view. Our approach differs because of its primary focus – digital archiving.

Our approach has been based on the assumption that the archiving system will be stand alone one and no extra effort with its running 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. In other words the robustness of the system was the key issue that influenced the system design. Having collection of documents of this type (with appropriate redundancy) 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.

We designed and implemented a system called DIPREP (DIPloma these REPository), see Figure 1. The DIPREP has following parameters:

- information should be stored in such a form that could be later transferable in more general digital archive (should allow the document migration)
- the effort for submitting documents in the archive should be minimal (students should be able to submit their theses by themselves)
- the security issues (the text should not be modified or stolen after its submitting into archive) 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 text documents will be stored. It is not assumed that executable programs etc. will be archived.

The key decision was selection of formats in which the document will be stored. There exist a lot of various recommendations which formats 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, tables and images. In such a case general recommendation is mostly targeted to Adobe PDF format. In our case a very important fact is that there is no problem to convert theses from traditionally used formats into PDF. At Czech Technical University the theses are written by means 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.

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 the data will be automatically generated by means of university information system. The information is stored in the form defined by Berkeley Database Version 4. 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.

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 that are left out). This textual form also allows the user to search 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 bitmap. This format is very simple - 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.

Both formats selected (text, bitmap) are very simple. This means that in future there will be always some way to create some tools by means of which it will be possible to read the contents of these files. Other data formats

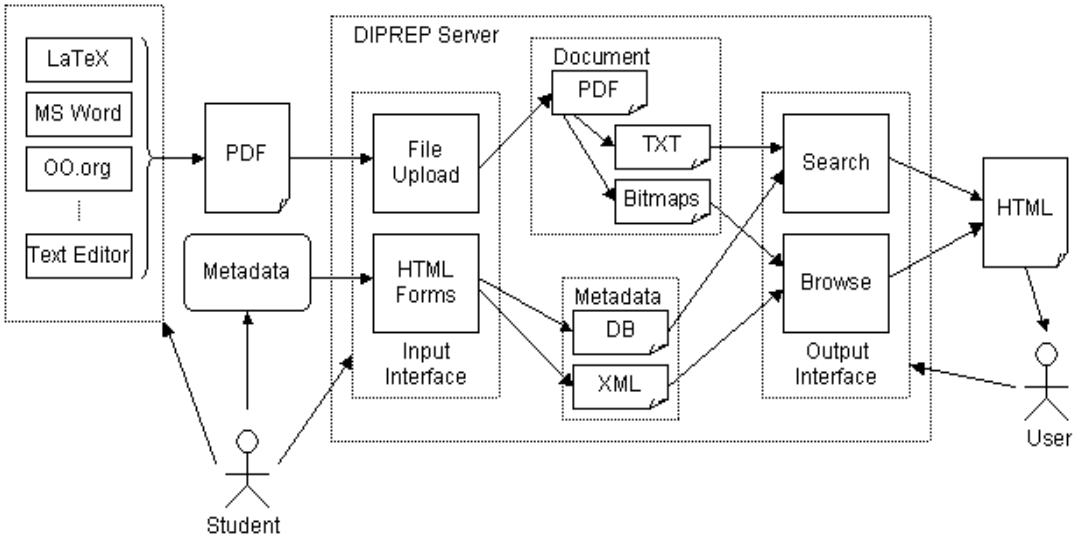

Figure 1. Document Life Cycle in DIPREP

used are mostly rather complex – sometimes their structure is not publicly available. This fact creates a serious obstacle in situations when information recovery is needed.

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, see Figure 2. The simplicity of formats might allow the user 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. This means that also metadata are stored in two independent formats (redundancy and robustness of metadata is thus guaranteed).

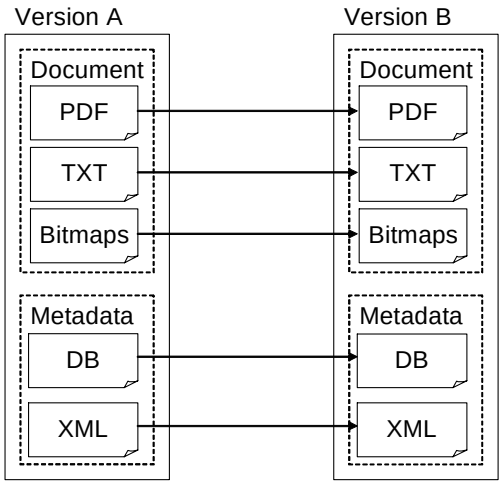

Figure 2. Migration scheme in DIPREP

Generation of metadata is linked up with subsystem that is responsible for user authentication. It was necessary to design and implement specific interface between the archiving system and university information system where all students are registered. This means that the data about individual students who access the archiving system (when submitting their theses into system) are derived (after confirmation that the person trying to access the system is eligible) from database that is part of university information system.

Specific scheme for the access the information stored has been designed (and already 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, see Figure 3. In this solution an important fact is that the contents of the document (theses) is displayed in raster format (where the watermark has been merged). This means that the key representation (PDF format) is not accessible and thus the danger of the misuse of the document has been minimized. Moreover the raster image is displayed in reduced quality – more than sufficient for reading but not sufficient for information extraction in digital form.

This approach does not exclude the possibility to create a specific access procedure for wider public. This procedure should take into account security and legal aspects (like copyright etc.) related to the access to this sort of documents. Design and implementation of such a procedure should be part of the future work.

3 EXPERIENCE GAINED AND FUTURE WORK

There exists a pilot implementation of the system described. The first practical experience has been gained in June 2004 when students submitted 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. could emerge.

One of the goals of the research conducted was to develop an easy and cheap strategy for digital archiving in small scale. This means that there are currently many applications where digital documents are being created but no simple (or relatively simple) solution is available. Such a situation might result in the loss of many digital documents currently existing. The solution designed and implemented offers a possible way to archive digital documents (at least for some particular applications) with sufficient redundancy what will make possible in the future to store these documents in some proper form in the framework of a large scale system where more general solution will be offered.

It was stressed in the introductory part of this paper that the system developed has besides its traditional function (public access to theses) also a research aspect. Vision about future experiments has been given. There is also another important aspect not mentioned: education of experts in the field of digital archiving. Up to now there was no special course in this field taught at Czech Technical University in Prague. A new study program is currently under preparation. This program should also include course oriented towards digital archiving. The system designed and implemented should serve in future as a tool where some practical aspects of digital archiving should be demonstrated in practical way.

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.

One of the improvements we are working on is also an archiving of not purely textual documents. There are many types of heterogeneous data besides regular figures, tables etc. contained in archived theses in PDF format. The text of the document is the most important information to be preserved in the theses. Nevertheless the thesis usually consists of other parts like software developed as a part of the thesis, presentation slides for a thesis defense, any multimedia material (like video recorded during experiments etc.) that could not fit into the PDF document. This means that the result of the thesis is not fully covered by the text document stored in archive so there is a motivation to store such information in appropriate formats as well. The good starting point for our research was an archiving of defense material – presentation slides.

either learning or recalling process. We pushed the design of the language to be close to these abstract descriptions as much as possible.

2.1 Implementation Notes

The SiMoNNe implementation contains an incremental compiler of the SiMoNNe Language into a special neural byte code which is interpreted by a virtual machine or which can be translated to native machine code of the target platform. Currently we have two implementations. The first is in Common Lisp and takes advantages of accelerated compilation and easy development of the SiMoNNe Language in Lisp. The second and more robust one is in Java. This implementation exploits advantages of portability.

Additionally, our implementation of the simulator exploits following features:

- **Remote access** to the simulator is provided by splitting the text communication stream with a TCP/IP layer. Afterwards, the SiMoNNe server can be hosted on a powerful server and the simulations can be performed remotely.
- **Graphical User Interface.** Since the SiMoNNe Language is well defined bi-directional interface, one can connect a GUI to the simulator. The GUI generates language sentences and sends them to the simulator. During the simulation the GUI reads text output of the simulator and represents it in the graphical form. The GUI is currently under the development.

3 USAGE AND TESTING

Following example SiMoNNe sessions show a subset of SiMoNNe features. The SiMoNNe was used for simulation of GOLOKO [8] neural network and artificial retina model. We use SiMoNNe for simulation of Categorizing and Learning Modules (CALM) [9].

3.1 Modular Neural Networks

Let us take CALM as an example of a modular neural network. First of all we define one of the basic building block - a neuron. Following example session shows how to define and run a modular neural network. Let us start with definition of a simple module (neuron, see Fig.1):

```
moduledef V_neuron begin
  output={0.0};
  up=0.5;
  flat=-1.0;
  k=0.05;
```

The module definition has two parts. First part is executed while a module is being constructed in a simulator. We can define any variables in a module definition.

Second part of each module, which starts with `stepcode` keyword, contains a code that is executed in one simulation step of the module.

```
  stepcode
    exc=input[0]*up;
    for(i=1; i<length(input); i=i+1)
    exc=exc+flat*input[i];
    output[0]=activation(exc, k, output[0]);
end
```

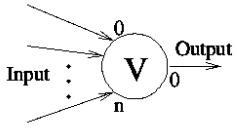

Figure 1. Simple example of a module - neuron.

Activation function used in previous definition has to be stored in the simulator as well. Then we will continue defining CALM module according to Fig.2.

Figure 3. Example page accessed via DIPREP

These presentation slides have one difference to text printable documents – an interactivity that could be exploited by a user during the presentation. Since there is currently no standard for animated presentations (specific format depends on a tool used for creation of the presentation such as Microsoft PowerPoint, OpenOffice.org, Latex, Keynote, Adobe Flash etc.), we have to reduce the interactivity by conversion to PDF format as in case of text documents. Loss of interactivity is a trade for long term readability guaranteed by PDF format.

The presentations usually employ various forms of interactivity like moving text, animations of objects etc. Our current approach is based on the preservation of individual slides (that have static form – no animations etc.). The archiving is done in the same way as archiving of the thesis content (PDF + text + bitmap). Having these documents we can have at least the basic information about the content of such a presentation together with some kind of information of dynamic behavior of the whole presentation.

Possible improvements will be based on construction of an accompanying static file where besides the transition between individual slides also dynamic changes inside of an individual slide will be made. In other words: having a series of static slides (stored in formats mentioned above) we can get also information about the individual changes in individual slides. These changes will be represented as a series of PDF documents that will represent individual stages of a development in the framework of one particular slide. The accompanying static file will contain the information about time necessary to perform these individual changes.

Technically, the PDF format allows us to define e.g. transition styles between slides but such feature is not supported by all PDF processing and viewing tools. Besides a creation of an interactive presentation in Latex produces a sequence of slides that constitute the animation.

Not only presentations accompany thesis. A lot of theses contain accompanying source codes that were developed in order to fulfill the task. Archiving of such source codes without preserving their future usage (compilation and running) is not satisfactory. In case of platform independent languages such as Java we expect that future usage might be possible. Archiving of such data leads to heterogeneity in the archive and making it readable and usable in future is one of our future tasks.

4 CONCLUSION

We have designed and developed a system called DIPREP which easily archives students' final theses. Thesis is not the only document type created in the framework of any university. In case of CTU (as a mid-range institution) many other classes of documents are being created such as progress reports, articles, papers, grant applications, and students' project reports etc.. All these documents need to be archived. The DIPREP system can be relatively easily modified and used for archiving of such documents in our institution. The main contribution of this paper is investigation of methods that will allow the user in small institutions to have readable documents (of a special type) after certain period of time – the issue that is still not properly handled in general in this particular context.

5 REFERENCES

Borchert, M. (2002) Australian Digital Theses Program - Promoting Australian Postgraduate Student Research to the World. Retrieved 20 December, 2004 from the World Wide Web:
http://ausweb.scu.edu.au/aw02/papers/edisted/borchert/paper.htm

Eprints Free Software (2005) Homepage of EPrint Free Software. World Wide Web:
http://www.eprints.org/software

Hey J., Hitchcock S., Carr L. & Brody T. (2005) Preservation for Institutional Repositories: practical and invisible. *Proc. of the PV2005*. Edinburgh, Scotland: The Royal Society.

Mach P., Snorek M., Slavik P. (2002) Preservation and accessing documents in digital form in long time perspective, Research report – in Czech. Praha, Czech Republic: CTU in Prague.

ProQuest (2003) ProQuest preserves academic heritage with new dissertation archiving program. Retrieved 20 December, 2004 from the *ProQuest* website:
http://www.il.proquest.com/division/pr/03/20031005.shtml