Пет̄ш̄а национална конференција со меѓународно учес̄ш̄во ЕТАИ'2000
Fifth National Conference With International Participation ETAI'2000
Охрид, Република МАКЕДОНИЈА – Ohrid, Republic of MACEDONIA
21–23. IX 2000

I1-3

SYSTEM FOR DIGITAL PROCESSING, STORAGE AND INTERNET PUBLISHING OF PRINTED TEXTUAL DOCUMENTS

Grcevski Nikola[1], Mihajlov Dragan[2] , Gorgevic Dejan[2],

[1]SEMOS, MK-1000Skopje, Republic of Macedonia, ngrcevski@yahoo.com
[2] Faculty of Electrical Engineering, Ss. Cyril and Methodius University
P.O. Box 574, MK-1001 Skopje, Republic of Macedonia

Abstract-- **Written sources of information are generally the best way to store data or other kind of human created materials and creations, like songs, music etc. From very beginnings of human civilization, people realized that valuable and important "things" should be written, so that they are remembered. Civilization created alphabets, letters and many machines afterwards to support the written material production.**

The later growth of information technology lead to the need of transformation of these printed materials in suitable digitized form, and by that giving them extended capabilities of manipulation and use of stored information. By this old and long time prepared and stored valuable information gets new dimensions of flexibility and usability, creating electronic digital archives. These archives offer many important advantages over standard printed libraries and archives like, ease of copying, ease of availability, the possibility of creating search engines, large indexes etc.

Index terms**-- documents, digital processing, storage and internet publishing**

1. INTRODUCTION

The new age of digital culture increased the need of electronic document storage and retrieval not only to those documents that are produced in our time, but also for documents that are written many years ago, which are ultimate source of information and knowledge. Digital transformation of these written documents is time consuming and very resourceful operation, involving a lot of human resources and however, can lead in lot of potential errors during the transformation. The problems involved with this transformation lead us in development of efficient system for digitizing written documents, which is also flexible and can be adopted for many different kinds of documents.

Digitizing of documents not only involves just simple transformation from printed to electronic form, but also can be supported by additional effort in document classification and index creation. These additional elements, which accompany the process of digital transformation, are also considered in this paper and they are however the most complex part of the process, which requires human interaction and their adoption and adjustment to different kind of documents or materials, is often impossible.

Another important aspect of this digital transformation is archiving of the documents into large and massive digital archives and libraries, their organization and the response time to document request. The information retrieval techniques that should be considered for implementation of large databases of indexed documents, which may contain graphics along with the text, often require extensive planning and careful implementation.

The phase of document transformation and digital library creation takes several steps.

2. DOCUMENT TRANSFORMATION

The first step in written document transformation is document content analysis in order to colect information that will be used for searching purposes and information retrieval. From the contents of the documents important features are extracted, then classified and this classification will be used later when storing the digitized documents in the document archive.

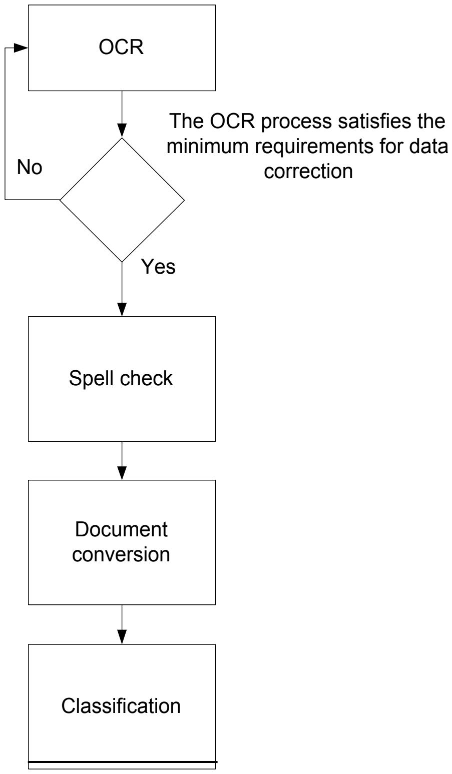
OCR
The OCR process satisfies the minimum requirements for data correction
No
Yes
Spell check
Document conversion
Classification

Fig.1. The workflow model of the system for electronic document transformation

After the analyse of the document content the next step in document transformation is Optical Character Recognition (OCR) of the printed material. This is one of the most important steps in the entire operation, because the quality of this procedure directly effects the success of the electronic archive. If the recognition process produces many errors the correction process, which follows, will be time consuming and resourceful operation. This process is mainly dependent of the scanning equipment, which is used for scanning images of the printed material. The resolution of the scanned image has crucial effect on the recognition engine. Even the most quality recognition software needs at least 300 dpi for OCR reading. Sometimes, generally when archiving older documents, the print quality of the materials is in bad condition or has poor print marks. Applying higher scanning resolution in those cases is necessary so that OCR program can function correctly. The higher the scanning resolution is the slower the process of scanning will be. Another important aspect of the scanning is that high speed scanning machines does not solve the problem of fast document image retrieval, because the scanned materials are often with different light intensity so the OCR program sometimes functions better if the scanning brightness is lower and sometimes if it is higher. For the scanning purposes the flatbed scanner usually is the best choice because page by page scanning, which is faster, requires the materials to be separated from the book covers and by that having the risk of destroying the original content.

Correction of the OCR produced material. In this step the digitized form of the document is put trough a spell checking program, which is used to correct the error that are produced by the OCR software or that were present in the printed material. No matter how good the OCR program recognizes the letters in the printed material, some errors will eventually occur mostly because of the poor printing quality. Spell checking of the document will find the misguessed letters from the OCR program and will provide a list of possible replacements that user can choose to replace. The grammar of the text is expected to be correct since no actual changes to the text structure are made.

Document conversion. Previous step is generally carried out in some text processor like Microsoft Word. The file format of these documents is not suitable for storage in databases, because it lacks the flexibility needed for Internet publishing of the documents. In order to convert the documents in the HTML format (for the Internet), special programs sweep the document database, automatically converting the prepared documents. These programs function as batch procedures and are started in the off-hours so that does not take significant processing time and slow down the production process.

The previous three steps can be replaced with a single step of creating an image of the documents in some suitable format rather than going through the entire process of OCR, spell check and conversion. But the advantages of the materials digitized using the previously described process are worth the effort. If the materials mainly contain textual information the character-based databases can contain much greater number of documents than the image based, using the same space. Another thing, probably more important is the text based search capability opposed to the database field search procedures that are the only way of document search and retrieval with image based libraries. On the other hand, if the system is to be published on the Internet the download speed of the documents will be important factor to the database functionality. It is well known that downloading an image, bit by bit, will be much slower in this case than the OCR processed text.

3. DOCUMENT ARCHIVING

The prepared documents, if left as they are, have no actual function, beacause in this form they are not classified and can not be easily retrieved on demand. Using the data colected in the first step of the transformation process, apropriate RDB data model can be defined for database classification and sorting.

The prepared document data is then classified in groups, subgroups forming the database directory. Using the parameters and classification the documents are stored in to the database with the additional properties assigned to them. These additional parameters are valuable information to search engines and are the source of directory creation. The choice of the appropriate database engine is dependent of the amount of data to be collected and stored. The system that we developed uses Microsoft SQL Server as DBMS.

For the archiving purposes additional application modules were created to classify the texts and create RDB table data. The actual documents that were transformed from written to electronic form were not stored in the RDB, but the relative paths from the application directory. By this any further document corrections and upgrade does not have any effect on the database data or structure except if the corrections imply a change in the document properties that are attached to the document.

The system concept is to be used as a search engine for the transformed documents. Moving toward the new standards for search engines an Internet organized and enabled database is the most convenient way to do the job. For the purpose of the system we used Microsoft IIS Server 4.0 and Microsoft Index Server 4.0. For the WWW based application can access the database we used Active Server Pages scripts to retrieve the database directory structure and perform the search process.

Indexing of the documents was done in two ways, by using the document attached properties and by using the Microsoft Index Server to index the contents of the documents. Indexing of the contents of the documents is a complex and demanding task, which Microsoft Index Server does automatically. By this, there are two different ways to search for the documents, by using the properties supplied for each document and by searching a keyword somewhere inside the document content. The content indexing can be avoided, but then the content search will be extremely slower, especially when there are multiple connection to the document database. The attached properties of the documents are also the source for directory creation.

The effectiveness of the system is measured according its manipulation and availability capabilities. The process of extensive data and procedure modeling preceded the process of electronic document archive development in order to minimize manual data processing. Production system modeling process is vital to the determination of necessary parameters for the search engine.

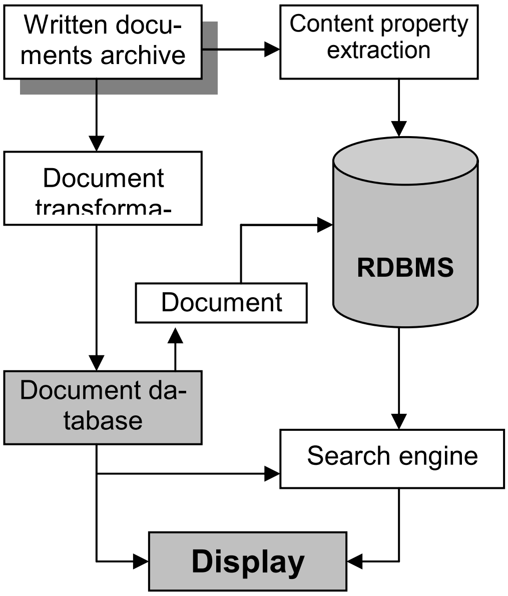

Fig. 2. Workflow diagram of archive structure and functioning

4. CONCLUSIONS

Text based electronic transformation of written document has many advantages over the image archives, although it takes more time and resources.

The transformed documents are of little or no use if not classified and stored in archives, which are search and directory enabled.

The new era of distributed computing implies WWW implementation of the search engines and the electronic archive.

5. REFERENCES

[1]. Enhancement and Restoration of Digital Documents : Statistical Design of Nonlinear Algorithms by Robert P. Loce, Edward R. Dougherty. Hardcover (January 1997).

[2]. Managing Gigabytes : Compressing and Indexing Documents and Images (Morgan Kaufmann Series in Multimedia Information and Systems) by Ian H. Witten, et al. Hardcover (May 1999).

[3]. Managing Gigabytes : Compressing and Indexing Documents and Images
by Ian H. Witten, et al. Hardcover (October 1994).