DIGITAL PRESERVATION:
THE RARE AND UNIQUE'S LONGEVITY

Antonia Arahova & Eleni Mamma
National Library of Greece — tonia@idkaramanlis.gr
Hellenic Ministry of Foreign Affairs — e_mamma@mfa.gr

During the epoch of latest rapid changes in the information environment, the outstanding existence of new technologies makes its appearance and acts with the main aim to support, exploit and distribute information services' content. The principal goal of the poster is to present an important piece of the work of two powerful organizations in Greece which create digital cultural content and conduce to the management and the utilization of valuable information resources. On the first hand, National Library of Greece, the biggest public information entity and on the other hand, Lambrakis Press S.A. which consists one of the biggest private publishing houses in Greece with a great number of publications in its credit. What the poster shows is the implementation of two new soft wares that enforce the full management of old newspapers and magazines using technologies such as OCR and digital documentation. These soft wares deal with problems, which have to do with the rarity and the decay of old and valuable printed material. Furthermore, the can be implemented to different technological platforms and they have the capacity to elaborate with multilingual texts which need to be preserved and accessed by different kind of users from all over the world.

THE FIRST MODERN (18th and 19th CENTURY) MAGAZINES AND NEWSPAPERS DIGITALLY IN OCR AND METADATA METHOD PRESERVED

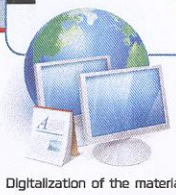
- MELISSA (1819 - 1821)
- EUROPEAN ERANISTIS (1840 - 1843)
- ILISSOS (1868 - 1872)
- ATHINAION (1872 - 1873)
- POLITIKOS ANTHON (1886 - 1887)
- (TEHNI) - ART (1898 - 1899)
- TA NEA (1931 - TODAY)
- TO VIMA (1922 - TODAY)

Projects >>
1. Stride
2. Korais
3. E-preserve newspapers

Basic stages
of a newspaper digital library creation

Newspapers collections

- "quantity"
- "quality"
- conservation of collections from detriment problems
- digital indexing of the newspaper leaves
- collections digitized through a specific programme of images scanning
- articles cataloguing
- insert all of the essential metadata
- emerged data elaboration
- Digital objects' repositories
- articles indexes
- articles indexes
- data management bases
- search and retrieval systems developed
- Ultimate purposes: the digital material's release in the Internet, the end users access to this material and finally the organized articles' retrieval, so as the full text, the OCR text and the metadata to be offered to the users on time, and as much as possible with greater speed, precision and completeness



Essential steps of the newspapers and magazines digital library technical process

Digitalization of the material includes six high level steps, made through the consequent quality and security control of data

Step 1: Scanning

The creation of the digital collection both from the printed material and the microfilm is performed. Images records in jpeg and tiff format through high digital analysis are produced. Later, these images are copied and archived as a product of this specific work.

Step 2: Article clipping

Article clipping segmentation leads to the creation of individual article images with the aim to be retrieved one by one from the whole page. During this phase, all articles are marked and catalogued electronically in order the end user navigates through all pages and has the possibility to retrieve the whole or the part of the requested article.

Step 3: Metadata

Step 4: Optical Character Recognition (OCR)

Using suitable and "smart" optical character recognition techniques, the individual articles are traced and automatically recognized. For article tracking, a novel rule based approach is being followed, which exploits the segment relationships that exist in the page layout format of the newspaper and magazines pages.

Step 5: Search

The search approach is to provide end users with a variety of search tools which will help them to cover in an effective way their information needs. The search, made both in articles' metadata and the content (Full text and/or keywords), is a combination of a Boolean conjuction and a ranked combination, so as to offer a single list of results. For the achievment of all these above mentioned, the search mechanism exhibits some characteristics such as:
The most extensive coverage of the Greek Language - various types known.
End users' support of various levels of searching sophistication.
Provision of simple ways so as the results of searching to be presented
Greek and Multilingual language support over the newspaper articles, which cover a broad time period of almost 200 years.

Step 6: Access

The end users have the capability of accessing to the whole content and look to the requested article, which is marked in an appropriate way.

Technical Description

Scope:
- suitable infrastructure knowhow
- Method of optical recognition
- Bibliographic alphebatization
- Accessible and exploitable material

Method:
- National Agricultural Text Digitizing Project
- Page-images
- ASCII files
- Bibliographic data
- Optical Character Recognition in different types of writing
- Special software for the management of the database

Project' s Purpose

- Rescue of old material from the deterioration
- Improvement of provided services in the public
- Decongestion of stocking space

Results

- Right organisation of infrequent material's transport in optical means of wide use
- Reject of faults of choice and resolution of all technical problems
- Exploitation and safeguarding in the primary form - rescue irreplaceable historical and cultural archives
- Growth software interface for the recognition of characters of Greek alphabet

Software

- System Software
- Applications' Software
- System' s Functionality Software
- Databases' creation Software:
- Completed as for the available operations
- Friendly for the user

a) From any form of information (ASCII, images) the user can be led automatically to the other
b) The search in ASCII files will become with base word, root of word, combination of algebra Boole, with possibility of restriction of breadth of search with use of criteria as year, date, language e.t.c.)

Chances - Possibilities

1. Mixture Greek and Latin characters
2. Support of various types of writing
3. Creation of special symbols
4. Localisation and change of word, department of word and line of characters
5. Command on automatic layout
6. On-line Help
7. Simultaneous production of different bases - segregation of files
8. Creation on each printed LOG file that will contain summary, dates of beginning, code operator e.t.c.)
9. Batch-job Page-images
10. Automatic segregation of page in text and graphic
11. Automatic localisation of likely errors
12. Creation of contacts between concrete forms (eg forms that include 1 article in continuities)
13. Recuperation of information
14. Thumb through Electronic Pages of Form

>> Importance

- Existence of additional functional rungs for the recognition crumbled or clusters of characters
- Distribution of cultural archival material
- Special dinamically prolongative database
- Benefits from completed and also organised electronic information' services
- Recuperation and distribution of infrequent Modern Greek forms

IFLA WORLD CONFERENCE 2007, DURBAN SOUTH AFRICA