An integrated system for creating a Digital Library from Newspaper Archives

B. Gatos, S.L. Mantzaris and N.Gouraros

Department of Digital Technologies, Lambrakis Press Archives,
8, Heyden Str., 104 34 Athens, Greece, bgat@dolnet.gr

Abstract. Newspapers are considered to be the first draft of history, while at the same time, are part of a country's cultural heritage. By converting newspaper archives to digital resources we achieve digital preservation in terms of preventing paper deterioration as well as providing full utilization of the archives by all interested parties. In this paper, we present a series of applications pertaining to the retro-conversion of newspapers, i.e. the conversion of newspaper pages into digital resources, as well as to the transformation of the printed material to an accessible digital archive. These applications constitute an integrated system that provides solutions to problems related to digitization, verification and quality control of newspaper images, manual and automatic article clipping, and, finally, information retrieval in intranet and internet environment.

1. Introduction

Lambrakis Press S.A. is the largest and most experienced press house in Greece currently publishing 3 newspapers and 11 magazines. The Lambrakis Press Archives (LPA) is a sector of Lambrakis Press S.A. that deals with the conservation and digitization of the printed material as well as the design and the development of an archival digital library consisting of the digitized printed material. We have developed a workflow for newspaper pages preservation and digital library construction that is designed to fit the needs of any newspaper archive worldwide.

Our main task is to create a Newspaper digital archive that can provide full access to all articles of the newspaper issues. Firstly, the newspaper material is gathered and prepared for digitization. Then, a Newspaper Issues Catalogue is created by examining all archival material. After checking for missing, incomplete or deformed issues, we proceed with digitization, image preprocessing and a visual check of all digitized material. The preprocessing task mainly involves image filtering for the improvement of image quality, as well as skew correction in order to restore horizontal image status. Then, we proceed to article clipping and indexing by marking all articles and inserting all necessary article metadata. When necessary, an archival update procedure is applied. The final task of our workflow concerns the construction of a web-search module that will provide access to all articles in intranet and internet environment. In this paper, we will focus on article clipping which is implemented manually or automatically, as well as on the web-search module construction.

2. Manual Article Clipping

At the manual clipping module, all articles are marked and indexed electronically. The main steps involved are: (a) the user marks the articles of a page using isothetic polygons (polygons having only horizontal and vertical edges), (b) for each of these articles all the necessary cataloging information is given and stored in a database, and (c) article continuities are defined as links between two or more article parts that are located on different pages.

The manual clipping module is at a production phase (every day approximately 1000 articles are clipped and indexed, in 84 human hours). After article marking, the user inserts all the metadata that describes the article (title, over-title, subtitle, running headlines, category, authors etc.).

3. Automatic Article Clipping

We have implemented a new technique for automatic newspaper page segmentation based on gradual extraction of newspaper image components in the following order: Lines, images and drawings, background lines, special symbols, text and title blocks ([1], [2], [3]).

In order to restrict all different components inside simple geometric shapes we use isothetic polygons with minimum number of vertices. In this way we achieve simplicity of description and efficiency of storage. These polygons are defined by a recursive formula, where the resulting areas are calculated by successive additions and subtractions of simple rectangular blocks ([4]). By using isothetic polygons, we solve the problem of restricting text and inset segments as well as restricting slightly skewed text columns.

Individual articles are traced and automatically recognized using suitable optical character recognition techniques. For article tracking, we follow a novel rule based approach, which exploits the segment relationships that exist in the page layout format of newspaper pages ([2],[3]). A novel technique is used for the linking of the textual parts of an article that can be found on different pages of a newspaper issue ([5]). The automatic clipping module is at a testing - evaluation phase. We expect to conclude soon for a supervised automatic clipping module that will increase production 300%.

4. Web Search Module

An article can be located using a combination of metadata information (created during the clipping and cataloguing of articles), and textual information (full text or titles). Our corporate users investigation has shown that the search mechanism should satisfy the following requirements: (a) To support users with different search mentalities. (b)To support users of different levels of sophistication. Some users are novice searchers while others have significant experience using internet search tools. (c)To provide simple ways in order to help users express their information need. (d) To provide a simple and unified way to present the results of the search.

We believe that a broader spectrum of users would have similar requirements. Other issues that we have taken into account in order to design our search engine are: (a) The articles have a great thematic variety. (b) The articles cover a broad time period of over 100 years. This has implications on the language, for example the stems of some words have morphological differences. (c) Easily updating available information. (d) Portability and use of standard tools.

Our metadata are stored in an RDBMS (SQL Server 7.0) while a low resolution version of images is stored on filesystems. Every requested image is transformed to pdf and watermarked on the fly. For each requested article, a pdf image having the article outlined is presented to the user. Figure 1 presents our newspaper web search module. A user can enter information in any of the fields (i.e. date, page number, author, text. etc). To support more elaborate user queries, an advanced search dialog box is currently under preparation.

5. References

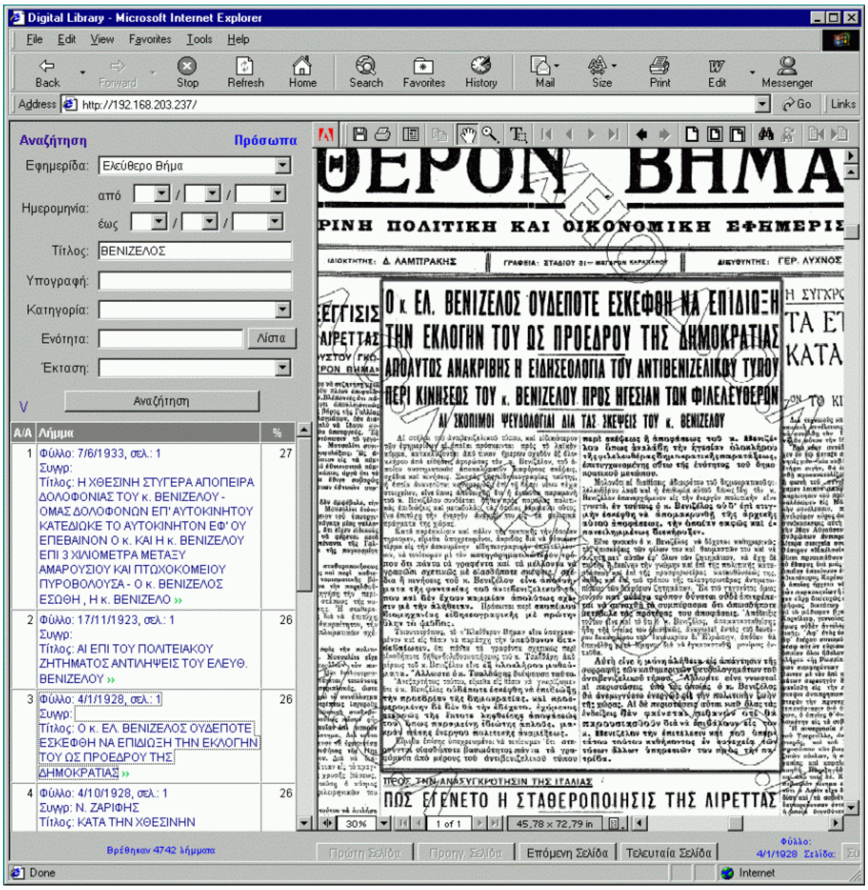
[1] B. Gatos, N. Gouraros, S. Mantzaris, S. Perantonis, A. Tsigris, P. Tzavelis and N. Vassilas, "A new Method for Segmenting Newspaper Articles" , Proc. of the *Second European Conference On Research and Advanced Technology for Digital Libraries (ECDL'98)*, pp. 695-696, Heraklion, Crete, Greece, September 1998.

[2] B. Gatos, S. L. Mantzaris, K. V. Chandrinos, A. Tsigris and S. J. Perantonis, "Integrated Algorithms for Newspaper Page Decomposition and Article Tracking", Proc. of the *Fifth International Conference on Document Analysis and Recognition (ICDAR'99)*, pp. 559-562, Bangalore, India, September 1999.

[3] B. Gatos, S. L. Mantzaris, S. J. Perantonis and A. Tsigris, "Automatic page analysis for the creation of a digital library from newspaper archives", *International Journal on Digital Libraries (IJODL)*, vol. 3(1), pp. 77-84, 2000.

[4] B. Gatos and S. L. Mantzaris, "A novel recursive algorithm for area location using isothetic polygons", Proc. of the *15th International Conference on Pattern Recognition (ICPR2000)*, pp. 496-499, Barcelona, Spain, September 2000.

[5] S.L. Mantzaris, B. Gatos, N. Gouraros and S.J. Perantonis, "Linking Article Parts for the Creation of a Newspaper Digital Library", Proc. of the *Content-Based Multimedia Information Access International Conference (RIAO2000),* pp. 997-1005, Paris, France, April 2000.



Digital Library - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Forward
Stop
Refresh
Home
Search
Favorites
History
Mail
Size
Print
Edit
Messenger
Address
Go
Links
Αναζήτηση
Πρόσωπα
Λίστα
Αναζήτηση
V
ΘΕΡΟΝ ΒΗΜΑ
ΡΙΝΗ ΠΟΛΙΤΙΚΗ ΚΑΙ ΟΙΚΟΝΟΜΙΚΗ ΕΦΗΜΕΡΙΣ
ΙΔΙΟΚΤΗΤΗΣ: Δ. ΛΑΜΠΡΑΚΗΣ
ΔΙΕΥΘΥΝΤΗΣ: ΓΕΡ. ΛΥΧΝΟΣ
Ο κ. ΕΛ. ΒΕΝΙΖΕΛΟΣ ΟΥΔΕΠΟΤΕ ΕΣΚΕΦΘΗ ΝΑ ΕΠΙΔΙΩΞΗ ΤΗΝ ΕΚΛΟΓΗΝ ΤΟΥ ΩΣ ΠΡΟΕΔΡΟΥ ΤΗΣ ΔΗΜΟΚΡΑΤΙΑΣ
ΑΠΟΛΥΤΩΣ ΑΝΑΚΡΙΒΗΣ Η ΕΙΔΗΣΕΟΛΟΓΙΑ ΤΟΥ ΑΝΤΙΒΕΝΙΖΕΛΙΚΟΥ ΤΥΠΟΥ ΠΕΡΙ ΚΙΝΗΣΕΩΣ ΤΟΥ κ. ΒΕΝΙΖΕΛΟΥ ΠΡΟΣ ΗΓΕΣΙΑΝ ΤΩΝ ΦΙΛΕΛΕΥΘΕΡΩΝ
ΑΙ ΣΚΟΠΙΜΟΙ ΨΕΥΔΟΛΟΓΙΑΙ ΔΙΑ ΤΑΣ ΣΚΕΨΕΙΣ ΤΟΥ κ. ΒΕΝΙΖΕΛΟΥ
ΠΡΟΣ ΤΗΝ ΑΝΑΣΥΓΚΡΟΤΗΣΙΝ ΤΗΣ ΙΤΑΛΙΑΣ
ΠΩΣ ΕΓΕΝΕΤΟ Η ΣΤΑΘΕΡΟΠΟΙΗΣΙΣ ΤΗΣ ΛΙΡΕΤΤΑΣ
Πρώτη Σελίδα
Προηγ. Σελίδα
Επόμενη Σελίδα
Τελευταία Σελίδα
Done
Internet

Address: http://192.168.203.237/

Εφημερίδα: Ελεύθερο Βήμα

Ημερομηνία: από / / έως / /

Τίτλος: ΒΕΝΙΖΕΛΟΣ

Υπογραφή:

Κατηγορία:

Ενότητα:

Έκταση:

Α/Α	Λήμμα	%
1	Φύλλο: 7/6/1933, σελ.: 1 Συγγρ: Τίτλος: Η ΧΘΕΣΙΝΗ ΣΤΥΓΕΡΑ ΑΠΟΠΕΙΡΑ ΔΟΛΟΦΟΝΙΑΣ ΤΟΥ κ. ΒΕΝΙΖΕΛΟΥ - ΟΜΑΣ ΔΟΛΟΦΟΝΩΝ ΕΠ' ΑΥΤΟΚΙΝΗΤΟΥ ΚΑΤΕΔΙΩΚΕ ΤΟ ΑΥΤΟΚΙΝΗΤΟΝ ΕΦ' ΟΥ ΕΠΕΒΑΙΝΟΝ Ο κ. ΚΑΙ Η κ. ΒΕΝΙΖΕΛΟΥ ΕΠΙ 3 ΧΙΛΙΟΜΕΤΡΑ ΜΕΤΑΞΥ ΑΜΑΡΟΥΣΙΟΥ ΚΑΙ ΠΤΩΧΟΚΟΜΕΙΟΥ ΠΥΡΟΒΟΛΟΥΣΑ - Ο κ. ΒΕΝΙΖΕΛΟΣ ΕΣΩΘΗ , Η κ. ΒΕΝΙΖΕΛΟ »	27
2	Φύλλο: 17/11/1923, σελ.: 1 Συγγρ: Τίτλος: ΑΙ ΕΠΙ ΤΟΥ ΠΟΛΙΤΕΙΑΚΟΥ ΖΗΤΗΜΑΤΟΣ ΑΝΤΙΛΗΨΕΙΣ ΤΟΥ ΕΛΕΥΘ. ΒΕΝΙΖΕΛΟΥ »	26
3	Φύλλο: 4/1/1928, σελ.: 1 Συγγρ: Τίτλος: Ο κ. ΕΛ. ΒΕΝΙΖΕΛΟΣ ΟΥΔΕΠΟΤΕ ΕΣΚΕΦΘΗ ΝΑ ΕΠΙΔΙΩΞΗ ΤΗΝ ΕΚΛΟΓΗΝ ΤΟΥ ΩΣ ΠΡΟΕΔΡΟΥ ΤΗΣ ΔΗΜΟΚΡΑΤΙΑΣ »	26
4	Φύλλο: 4/10/1928, σελ.: 1 Συγγρ: Ν. ΖΑΡΙΦΗΣ Τίτλος: ΚΑΤΑ ΤΗΝ ΧΘΕΣΙΝΗΝ	26

Βρέθηκαν 4742 λήμματα

30% 1 of 1 45,78 x 72,79 in

Φύλλο: 4/1/1928 Σελίδα:

Figure 1. Web Search Module.