A Web-based Semantic Navigation System for Migne's Patrologia Graeca based on OCR extracted Page and Volume Numbers from the Table of Contents of Dorotheos Scholarios

EVAGELOS VARTHIS
Department of Library Science
Ionian University
I. Theotoki 72, Kerkira 491 00
GREECE
evagelosvar@gmail.com

MARIOS POULOS
Department of Library Science
Ionian University
I. Theotoki 72, Kerkira 491 00
GREECE
mpoulos@ionio.gr

ILIAS GIARENIS
Department of History
Ionian University
I. Theotoki 72, Kerkira 491 00
GREECE
yarenis@ionio.gr

SOZON PAPAVLASOPOULOS
Department of Library Science
Ionian University
I. Theotoki 72, Kerkira 491 00
GREECE
sozon@ionio.gr

Abstract: - In this paper, the prototype of a new tool is presented for the navigation of a 19th century collection of Greek authors. This collection is published by Jacques Paul Migne and it is known today as Patrologia Graeca (PG). The project aspires to interconnect this vast amount of about 120000 scanned pages with the scanned Table of Contents (TOC) published by D.Scholarios in 1879. The D.Scholarios's work contain summaries for the chapters and sub-chapters of PG, having next to them the corresponding volume and page number of the location in the PG. Using Optical Character Recognition (OCR) and pattern recognition techniques, we extract from D.Scholarios's work the appropriate information in order to create links to the specific pages of PG. Our aim is to provide a Web Interface in which D.Scholarios's work is used as a semantic compass for PG about the subjects it covers. The complete system consists by three main sections. A REST API backbone service for the scanned images of PG. OCR and pattern recognition techniques for extracting the volume and the page information from the scanned pages of D.Scholarios. A Web interface presenting the TOC by D.Scholarios with the appropriate functionality. The originality of our system lies in the interconnection of two different scanned texts for semantic enrichment and browsing convenience, especially if one is nearly 120000 pages and the other about 600 pages.

Key-Words: - Migne's Patrologia Graeca, Dorotheos Scholarios, Rest API; Web Interface, Semantic Web.

1 Introduction

This article presents an ongoing work on how to get a better way of exploring, as well as exploiting the semantic information contained in the scanned corpus of Migne's Patrologia Graeca (PG). PG is an epic Collection of works by east christian fathers over a period of 1400 years. This collection consists of 166 volumes (bound as 161) and exists in digitized form from various sources, indicatively we refer [1][5], however, only a fraction exists in unstructured edited texts. The transformation of the PG scanned volumes in edited form, has been done mainly by Thesaurus Linguae Graecae (TLG) using extensive writing. The work of TLG contains with a loose estimation nearly 20% of the complete Patrologia Graeca (compared only with the Greek texts), while the rest 80% still exists, only in scanned images, More specific, the works of TLG found in [2] contains 140 authors and 1524 works compared to 658 authors and 4,287 works identified by Perseus Digital Library (PDL)[3].

The PDL also provides works of PG, however, is far less comprehensive than TLG [4]. A Greek archbishop, D. Scholarios published in Greek, a table of contents for PG named "Κλείδα Πατρολογίας" [5][6] with the authors and subjects (Athens, 1879). He also published a more advanced interlinking, between specific words and authors (Athens, 1883) named "Ταμείον Πατρολογίας" [5][6]. Three decades later, in 1912, Garnier Frères in Paris, published a PG index volume, edited by Ferdinand Cavallera in Latin [5]. The works of D.Scholarios and F.Cavallera, are very important on semantic level. Specifically, for the work of D.Scholarios, as the user reads a topic in his work next to it, exists the page number of PG to locate the information. It is worth to note that, even for a scholar related in the field of PG, the finding of a single page is not an easy task. Having the printed TOC of D. Scholarios, it is necessary to have all the 161 volumes of PG in order to locate the information. This implies not negligible financial cost. In the case of an interested user who wants from time to time to find information about specific subjects, there are also difficulties. He should search the Web, find the specific volume, then download it and finally locate the page. This is a time consuming and tedious task. In general, we can say that this vast amount of information gathered over 14 centuries, is not easily accessible from a semantic point of view. Specifically, there is not a navigation system on the Web Domain that helps to find semantic issues about PG and easily locate them. Several studies for automated OCR techniques to extract text with various methods are presented in [7][8][9]. However, in the case of D.Scholarios, a rich set of approximately 600 pages of semantic topics, is already prepared by D.Scholarios. Due to the above reasons, we decided to build a system for easy location of any page of PG, combined with the work of D.Scholarios. The enrichment of PG with this interconnection will be semantically very important.

Our proposed system consists by three main stages.

A) Building a Rest API and a Web Interface for the presentation of Patrologia Graeca with capability to locate on the Web Domain a specific single page by giving a unique Universal Resource Identifier (URI).

B) Applying OCR techniques to the work of D. Scholarios in order to retrieve the page numbers that link to specific pages to PG.

C) Building a Web Interface of the table of contents of D.Scholarios that helps the user to navigate from the interface directly to the specific Patrologia Graeca page, when he finds a semantic subject that it is interesting for him. The above stages are independent of each other and can be

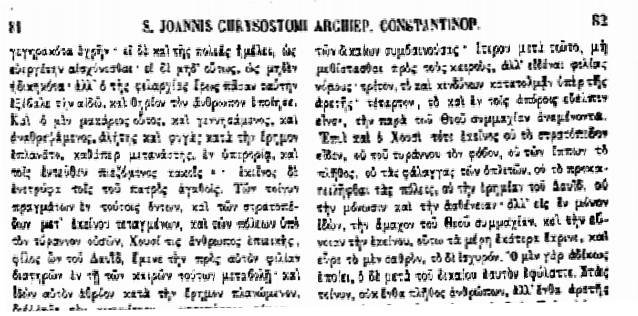
81 S. JOANNIS CHRYSOSTOMI ARCHIEP. CONSTANTINOP. 82

Figure 1. Image of the PG-page with column numbers 81-82 with Greek text

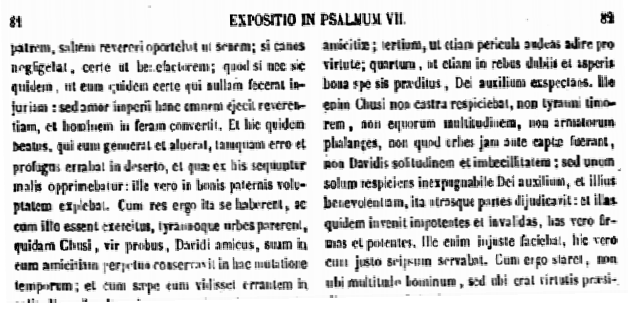
81 EXPOSITIO IN PSALMUM VII. 82

Figure 2. Image of the PG-page with column numbers 81-82 with Latin text

handled in parallel. In the following sections (Section 2, Section 3 and Section 4), we describe the implementation achieved per stage, the difficulties that arise, and what parts in our project are still incomplete. In Section 5, we discus the final results and possible considerations for future work.

2 Building a Rest API and Web Interface for the presentation of Patrologia Graeca

In order to being able to locate every single page of PG by an URI, a tokenization at the page level is needed. We initially collected the volumes in Portable Document Format (PDF) from various sources, such as the Google Project Libray, Archive.org, Harvard University and Hathitrust.org and concentrated them in the following repositories [10][11][12]. It is worth to note that the PG collection has some volumes bound as one (16, 86, 87). The method to split the PDF volumes and convert them to Joint Photographic Experts Group (JPEG) images can be achieved by using common Linux tools, such as Pdfseparate, Pdftoppm and Imagemagick.

We put all the pages of a specific volume in a separate folder with name, the number of the specific volume. The bound volumes are numbered with the number followed by a suffix letter (16a,16b,16c,86a,86b,87a, 87b,87c).

The crucial problem however, is how to name the pages. PG has a particularity regarding the numbering of pages. Every single page has two columns and each column has its own numbering. However this is not a strict rule. There are volumes like '55' and many more where there are continues pages with the same double numbering. One for Greek and one for the Latin translation, see Figure 1, Figure 2. There is no specific pattern that this particularity follows, so semi-manual checking is needed in order to capture these particularities of the pages. The right method for numbering is crucial in order to create an unique URI for every page in PG. Our proposed schema for the URI is shown in (1). When there is a double numbering, one candidate method, is to call the pages with the addition of suffix letters such as, in (2) and (3).

However, if the URI is called without the suffix, then the both pages have to be presented to the end user.

http://server.com/volume/{num}/page/{num1-num2} (1)
http://server.com/volume/{num}/page/{num1-num2}–{a} (2)
http://server.com/volume/{num}/page/{num1-num2}–{b} (3)

Based on this approach, the URI in (1) for the page 81-82 returns as response a JavaScript Object Notation (JSON) text, see Figure 6, containing both URLs of the images 81-82-a.jpg and 81-82-b.jpg. The JSON acts as a proxy to map correctly any particularities.

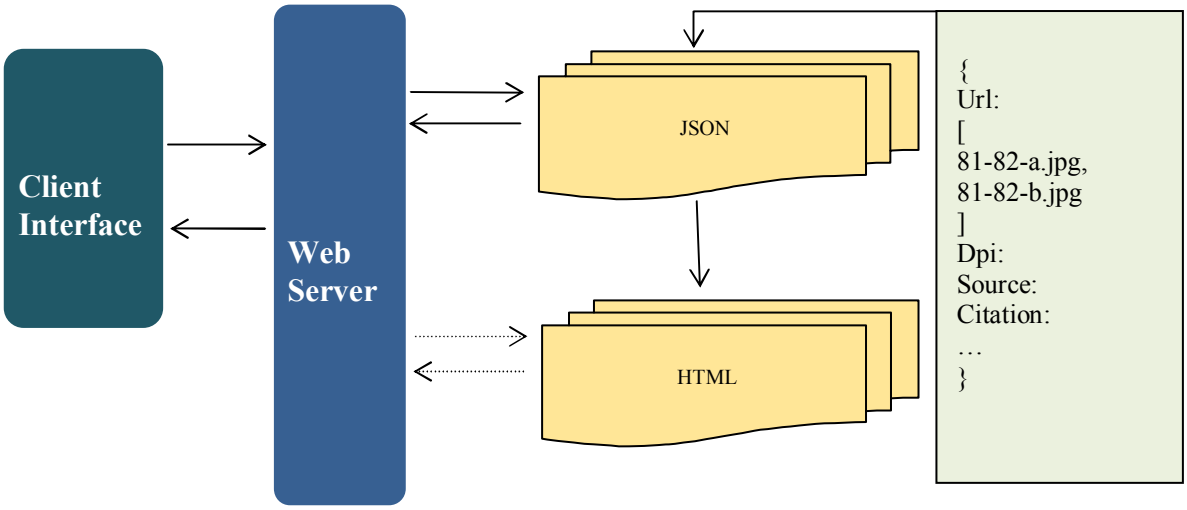

Figure 3. Overview of the Navigation REST API. The HTML file calls the JSON file to present the correctly matched PG image-page as well as to provide additional information.

The JSON files can contain additional information about the specific PG page, such as Dpi, Source, Citation or any other if it is required. In fact, having all the pages in unique URIs as in (1), an additional Web Interface is build specifically for PG. This interface can be used independently from our project. Moreover, various alternative client interfaces can be build by anyone interested since the architecture is open and accessible by using this simple Rest API.

The creation of such a repository of PG pages on the Web domain with the specific format has many advantages. Anyone can call any single page for viewing, downloading or even applying automatic OCR techniques to extract information in a simple memorizable way. Also, the URI or the direct image link of any page can be shared through various methods, email, citation etc. for a better communication or collaboration, something that is currently not easily achieved. Additionally, the JSON format is easily processed by any Web based application.

Figure 5. The extracted OCR text from a page of D. Scholarios's TOC.

Figure 6. The original image from a page of D. Scholarios's TOC.

<text top="55" left="45" width="24" height="14" font="0">1 3 4</text>

<text top="56" left="166" width="373" height="13" font="1">ΑΙΩΝ Δ'. — ΔΙΔΥΜΟΣ ΑΑΕ2ΑΝΔΡΕΥΣ — ΤΟΜΟΣ 3 9</text>

Figure 4. XML fragment with the coordinates of the text boxes

However, two important issues arose during the random checking of the downloaded PDFs. The PDFs produced during the original scanning by various sources [1][5] had repetitions of some pages in an arbitrary way. This adds extra difficulty and semi-manual methods are required to get rid of the unwanted pages. Also the quality of the PDF pages is not always high. Even in a high quality scanned volume, there are often low quality scanned pages. This also requires an additional work to collect better pages. The latter is crucial if we want our system to offer in parallel the ability of downloading via the Rest Api specific pages, for further processing.

3 Applying OCR Techniques to the work of Dorotheos Scholarios

Figure 6 shows a fragment of the TOC that D. Scholarios published. Our aim as already mentioned, is to use this TOC for a semantic navigation to PG. More specifically, the TOC is transformed into a structure of web pages so that when the user clicks to the specific page number of a volume to be redirected to the exact page and volume of PG. In order to capture the page numbers of PG, the Tesseract v3.5 [13][14] and Abbyy-FineReader-12 [15]

OCR tools are used. The results obtained by using Abbyy without even training were very satisfactory. On the other hand the Tesseract gave us poor results even using the ancient Greek language training pack [16]. Probably the kind of the fonts used in D.Schoralios work, affected the ability of Tesseract and additional training for the specific fonts is needed. As a consequence, we continued the OCR only with the Abbyy-FineReader. The OCR extracted text, is shown in Figure 5, next to the original image. The extracted text by Tesseract can be easily exported to a hocr file having the coordinates of the boxes. On the contrary, the Abbyy-FineReader (Corporate version) does not offer direct export of the coordinates. A work around this issue it is described below:

Firstly the OCR is applied, exporting the text to a double layer PDF, then the text is extracted from the PDF to an Extensible Markup Language (XML) file using the Linux tool Pdftohtml with the XML flag enabled. The produced XML file holds the information of the coordinates as shown in Figure 4, and so, it is easy to reproduce a HTML file with the accurate position of the boxes using instead HTML Content Division (DIV) and Scalable Vector Graphics (SVG) tags.

After the creation of the HTML files, various pattern recognition techniques can be used on them in order to find when the volume changes. The simplest method we adopted to find the changing point, is when the word "ΠΕΡΙΕΧΟΜΕΝΑ" (contents) or part of it appears to the HTML pages. At this point a hidden div tag with value equal to the number of the specific volume is inserted. We used as a mark point, the word "ΠΕΡΙΕΧΟΜΕΝΑ" for two reasons:

A) It is one of the first starting words when the contents for a specific volume change in the TOC.

B) It is recognized quite well by the Abbyy-FineReader reader, nearly by 90%, so only a little further manual correction and adjustment is needed.

4 Web Interface for D. Scholarios's TOC

In the third stage, the HTML files of the web interface are built from the XML files after the transition from the text-boxes to DIV tags with the corresponding coordinates. The SVG tag is used to stretch the text and capture all the space inside the DIV as well as to have a greater control to the presentation of the text. Afterwards, we code in JavaScript a small script that is called via the HTML files in order to capture dynamically the value of the hidden DIV tag, the value of the page number and produce the correct link dynamically. Both the DIV boxes and the original image are shown in Figure 5 and Figure 6 side by side. The OCR extracted text has some errors, so the original image helps the user to understand the topic he is reading. The links are colored red and reside in the edited text.

However, another candidate option for the interface is the following: The OCR extracted text is transparent on top of the image and we show only the border of the text boxes with the links, see Figure 7. Both methods have been implemented, however the final decision, will be taken after a pilot testing of the interface.

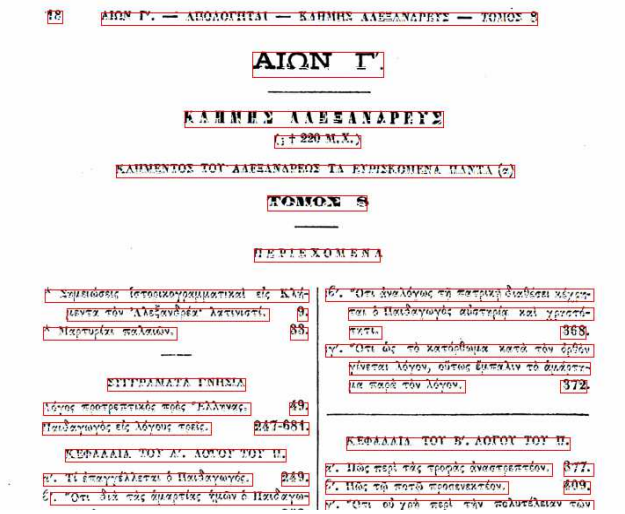

Figure 7. The original image with the appropriate hyperlinks to Patrologia Graeca (alternative interface).

4 Conclusions and Future Work

As pointed out at the beginning of this paper, this is the description of a work in progress. The first results are very encouraging. We provided a method for enriching semantically the epic collection of PG with the interconnection of D.Scolarios's work. The three main sections of our system are explained as well as the key problems and difficulties. The second and third stage of our project, i.e the extracting of the page number and the creation of the Web Interface of D.Scholarios's TOC, require some improvements, however, the main

functionality is almost complete. The prototype of this Web tool, based on these two stages, is uploaded for pilot testing in [17]. The building of the Rest Api (first stage) is the most challenging and time-consuming task. Intensive checking of the scanned volumes and pages is necessary in order to remove both the repetitive and the bad scanned pages . Also crucial is the design of the JSON response so that any Web-based application be able to easily handle the meta-data. After the completion of our project, any user will have the ability to find semantic topics to the vast amount of information gathered over a period of 1400 yeas in PG. Our proposed system is addressed to Greek-speaking scholars or interested users. However, the system can be also extended for Latin-speaking scholars. More specifically, a future consideration will examine the interlinking of PG with the Latina index edited by F.Cavallera.

References:

[1] Google Books Library Homepage, URL: https://books.google.com/.
[2] Ruslan Khazarzar Library, Patrologia Section, URL: http://khazarzar.ske ptik.net/pgm/PG_Migne/.
[3] Perseus Project Homepage, URL: http://www.perseus.tufts.edu/hopper/opensource
[4] Thesaurus Linguae Graeca, Homepage, URL: http://www.tlg.uci.edu/index.prev.php.
[5] Internet Archive Homepage, URL: https://archive.org.
[6] Digital Libary of Modern Greek Studies, https://anemi.lib.uoc.gr/search .
[7] Bruce Robertson, Christoph Dalitz, Fabian Schmitt, Automated Page Layout Simplification of Patrologia Graeca, DATeCH '14 Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, Pages 167-172, Madrid, Spain — May 19 - 20, 2014.
[8] Boschetti F., Romanello M., Babeu A., Bamman D., Crane G. (2009) Improving OCR Accuracy for Classical Critical Editions. In: Agosti M., Borbinha J., Kapidakis S., Papatheodorou C., Tsakonas G. (eds) Research and Advanced Technology for Digital Libraries. ECDL 2009. Lecture Notes in Computer Science, vol 5714. Springer, Berlin, Heidelberg.
[9] Bruce Robertson, Federico Boschetti, Large-Scale Optical Character Recognition of Ancient Greek, Mouseion: Journal of the Classical Association of Canada Volume 14, no. 3, 341-359, 2017.
[10] Collected List of Volumes between 1-50, URL: https://gitlab.com/patrologia/pmg001-050.
[11] Collected List of Volumes between 51-100, URL: https://gitlab.com/patrologia/pmg051-100.
[12] Collected List of Volumes between 101-161, URL: https://gitlab.com/patrologia/pmg101-161.
[13] Smith, R.: An Overview of the Tesseract OCR Engine. In: 9th International Conference on Document Analysis and Recognition, vol. 2, pp. 629–633. IEEE Computer Society, Los Alamitos (2007) Google Scholar.
[14] Tesseract Homepage, URL: https://github.com/tesseract-ocr/tesseract.
[15] Abbyy FineReader Homepage, URL: http://www.abbyy.com.
[16] Ancient Greek language training pack, URL: https://ancientgreekocr.org/2.0/grc.traineddata.
[17] Prototype Web Interface of D. Scholarios's work, URL: http://patrologia.tk/kleida/index.html.