Digital Greek Patristic Catena (DGPC). A brief presentation

Athanasios Paparnakis, Constantinos Domouchtsis

Digital Greek Patristic Catena (DGPC). A brief presentation

Athanasios Paparnakis,
PhD Theol., Ass. Professor, Aristotle University of Thessaloniki, Greece

Constantinos Domouchtsis,
BSc, PhD Theol. candidate, Aristotle University of Thessaloniki

Abstract
The project is to develop a database, which is planned to include all available information on the use of the Bible in the patristic works of Migne's *Patrologia Graeca*. Utilization of the data will be available through a web page equipped with necessary tools for developing data mining techniques and other methods of analysis. The main aim of the project is to revive the *catenae*, the ancient exegetical tool for biblical interpretation.

Keywords
Catena, bible references, patristic authors, biblical exegesis, database, Patrologia Graeca

INTRODUCTION

The project aims to revive with the aid of information technology the ancient hermeneutical tool of the Bible, the *Catena*, by producing an online research tool for scholars from the international community on the field of humanities and especially for disciplines on Christian literature. Our project is planned to combine:

- a full list of the biblical quotations in the works of the Christian authors of the first 15 centuries of Christianity, namely published in J. P. Migne's *Patrologia Graeca* (PG)
- the Bible text in the original languages and translations
- a full corpus of all the relevant Christian texts commenting on those quotations (*catena*)
- a full index of information and bibliographic data on each author and work of the first 15 centuries (*clavis*)

1. THE BACKGROUND AND STATE-OF-THE-ART

The basic idea is similar to the already running project *Biblindex* by Sources Chretiennes that is the state-of-the-art project in the field of the biblical references in patristic literature. Our project is based on a similar dataset, parallely and independently produced, and aims to utilize it in a different way. The core data comes from the work of two professors of the New Testament from the Aristotle University of Thessaloniki, Stergios Sakkos and Pausanias Koutlemanis. By the time when the authors of *Biblia Patristica* (*BiblP*) worked, these two professors also commenced to collect, verify and multiply the references to the Bible from the patristic texts edited in the *PG*. However, unlike the *BiblP* that stopped at the early fifth cen., the two professors processed all volumes of that edition, conducted lists of all the biblical references for each volume and published them in the re-edition of *Patrologia Graeca* by the Centre of Patristic Editions in Athens, Greece (1985-2010). The total sum of the biblical

references in the PG collection of 6.131 works of 685 ecclesiastical authors of the first 15 centuries edited in 225.138 columns of 170 volumes is 360.143. These references, including both quotations and allusions that extend from one word to a whole chapter, have been identified either by introductory formulae (e.g. "the Scripture says"), or by their content, philologiacally analysed in less detail, though, than *BiblP*. The two professors kindly offered this work for further development to our research team that comprises theologians, linguists and computer scientists. Also, there has already been achieved a concensus with the *Biblindex* for a full cooperation, when this project reaches a combatible state.

2. THE PROJECT DEVELOPMENT

The project is to be developed in a way that it will provide all available information on the use of the Bible in the patristic literature with statistical and data mining tools to conduct research on this data. Therefore, it does not focus on processing the texts themselves, but on management of the data concerning the biblical and patristic texts. We processed the original set of data in order to be combatible for insertion to a properly designed database in four steps:

i) The first step was to process the digitized tables of the biblical references in each PG volume and secure the reliability of the data. The following tasks were undertaken: (a) the data of the tables were verified by comparison to the original handwritten records as the producers finalized them. The tables contained seven columns of the following fields in numerical codes: book, chapter and verse of the Bible and volume, column, paragraph anlanguage of the text from PG. (b) Because each record referred either to a single verse or to a set of verses from the Bible, we had to analyse them to single out each verse; this increased the total number of the original records to 873.228. (c) We added new records after digitization of published indices, i.e. 7.300 biblical references from the works of Neophyte the Recluse (12th. cen.) and 5.576 from Gregory Palamas (15th cen.). The final tables constitute the core dataset available for process in a database.

ii) The second step was to design and build a dynamic database, keeping an open eye to future additions and developments, on the one hand, and to the combatibility with similar or other related projects (like *Biblindex*, *Perseus*, *TLG*, *Pinakes* etc.), on the other. Building this database we faced two main challenges: the first one was the normalization of the initial tables so that the Bible will be associated to the PG includinig all the available information. After applying the normalization rules to the third level, the database generated two single code fields that correspond to the unique correlation of each Bible verse to a particular paragraph of a patristic text. These are the nodes where data mining techniques will be applied. We created Indexes over all crucial fields (like foreign keys) in order to increase the searching performance of the database. The second challenge was to associate to this data all the available information concerning the biblical and the patristic works. Therefore we created tables utilizing another three sets of data on (a) the Bible text and subject indexes, (b) the patristic *clavis* and texts (names, dates, places etc.) and (c) several indexes to PG (tables of contents, subjects, names etc.).

iii) The third step was to input data to the database after resolving —as it is expected— numerous compatibility problems concerning the form of the data, the polytonic fonts, adaptation of extant data, digitization of printed material etc. The Bible texts selected were the –so called– Byzantine texts of the Old and New Testaments, as well as the scientific texts of Rahlfs and Nestle-Aland. Useful pieces of information like chapter/unit titles, cross references etc. were also included. Three sets of patristic material were also input in a primary form,

since they are still under process: the Cavallera tables of contents, links to the pdf files of the PG Athens re-edition and nearly 60.000 digitized passages from the patristic tetxs of references from the books of Genesis to 1 Samuel.

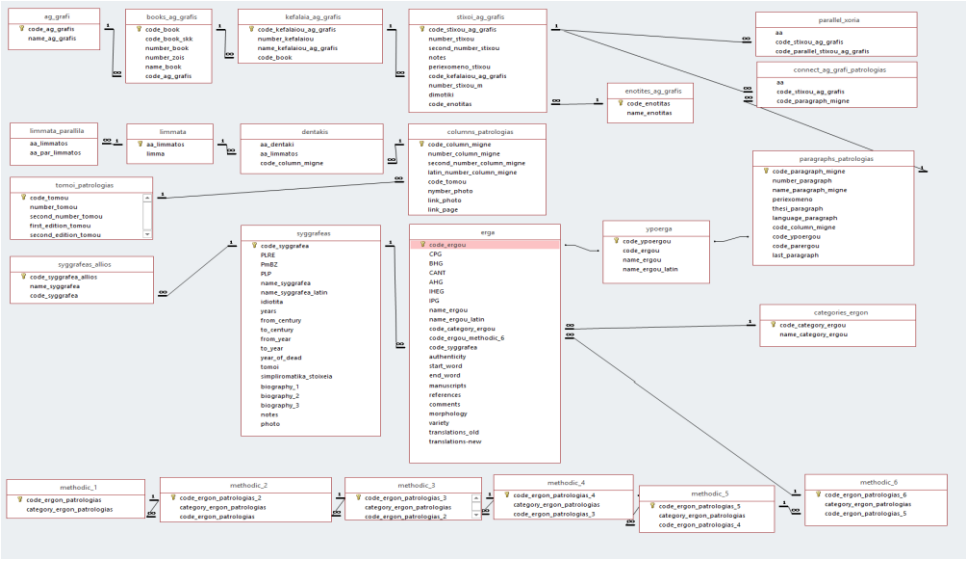

Table 1. Database design

iv) The fourth step was to develop the web page in order to provide access to the material in a multi-user environment with the possibility of feedback and flexible searches. Because we built our database on the relational model, we transferred it to MySQL —after resolving, of course, numerous compatibility problems— and created the web page using PHP, Apache and AJAX technologies. We uploaded the web page to the address http://dgpc.web.auth.gr for pilot testing of performance.

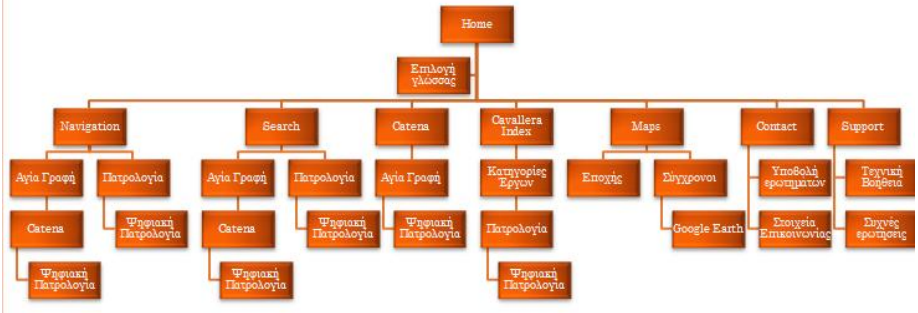

Table 2. Web page design

3. FUTURE PLANS

The main goal of the immediate future is to complete all the absolutely necessary data input, after they will have been verified and tested for functionality in the web page. The database is to be developed in order to hold all the available information about the patristic authors and their works from all major scholarly reference works, like PLP (Prosopographisches Lexikon der Palaiologenzeit), PmBZ (Prosopographie der mittelbyzantinischen Zeit), PLRE (Prosopography of the Later Roman Empire), CPG (Clavis patrum graecorum), BHG (Bibliotheca Hagiographica Graeca), CANT (M. Geerard, Clavis Apocryphorum Novi Testamenti, Turnhout 1992), AHG (Analecta Hymnica Graeca), IHEG (E. Follieri, *Initia Hymnorum Ecclesiae Graecae*, τ. I-V2, Città del Vaticano 1960–1966), IPG (Chr. Baur, *Initia Patrum Graecorum)*.

The final goal is to develop and integrate to the web page intelligent research tools that are flexible enough to meet the needs of scholars who work in the field, in order to analyze all this information in as many different ways as possible. Further to any obvious benefits stemming from developing such a database related to efficient searching and retrieving material, advanced analytics become plausible. Using SQL, we can submit aggregate queries of the form ``list the count of all elements per category'', where the notion of category can be defined dynamically by selecting a set of database attributes (i.e., each distinct combination of attribute values forms a different category) according to ad-hoc user queries. However, more advanced analytics can be enabled through the application of data mining techniques. Data mining aims to extract hidden patterns of knowledge from raw data that are not obvious and, in some case, are not even intuitive. It goes beyond simple database and statistical analysis. More specifically, we plan to explore the added value from the application of two main data mining techniques. First, we aim to extract association rules, which can succinctly describe the fact that certain elements tend to appear as a whole, and the presence of one set of elements increases the probability of appearance of another. For example, to automatically detect that two authors tend to make similar comments and the fact that a reference to a particular text tends to be accompanied (or should accompanied) by another reference to a seemingly unrelated excerpt. Second, we aim to apply clustering techniques and compare their results to those from a laborious manual process, such as the effort by G. Dentakis who produced extensive subject indexes to the patristic texts of PG. If this process is done successfully, data mining clustering would allow the automated and reasonably accurate grouping of records based on several criteria, including their meaning and text content.