IFLA WLIC 2019 Athens

Submitted on:26.09.2019

Satellite Meeting: *Artificial Intelligence and its impact on libraries and librarianship*
Date: *22 August 2019*
Location: *Ionian University, Corfu, Greece*

Semantic enrichment on Large Corpora: a case study for Patrologia Graeca

Evagelos Varthis
Department of Archives, Library Science and Museology, Ionian University, Corfu, Greece.
E-mail address: evagelosvar@gmail.com

Marios Poulos
Department of Archives, Library Science and Museology, Ionian University, Corfu, Greece.
E-mail address: mpoulos@ionio.gr

Ilias Giarenis
Department of History, Ionian University, Corfu, Greece.
E-mail address: yarenis@ionio.gr

Sozon Papavlasopoulos
Department of Archives, Library Science and Museology, Ionian University, Corfu, Greece.
E-mail address: sozon@ionio.gr

Abstract:

In this paper a case study for Patrologia Graeca (PG) is presented to reveal the difficulties that arise when a semantic enrichment though interconnections is pursued with specific aim to discuss an alternative cooperative architecture between libraries. The presented case method interconnects the PG with a range of satellite scanned or editable texts to provide a better user navigation experience, easy citation for specific PG page and document interconnections. Although the semantic enrichment can be achieved by using semi-manual methods, this is not recommended due to the fact that these tasks are time consuming and costly. Therefore a new architecture for a better cooperation between libraries and cultural institutions is proposed in order to provide a framework for the application of Artificial Intelligence(AI) and Pattern recognition (PR) techniques for mass enrichment on Large corpora. The proposed framework creates the possibility for the participating libraries to cooperate and provide a high quality commonly accepted version of a specific corpus which is used by independent research teams to apply AI and PR methods and create automatically the structure and the interconnections between documents.

Keywords: Artificial Intelligence, Semantic, Enrichment.

1. Introduction

Nowadays the libraries as well as various cultural institutions have already converted a significant part of text collections into digitized form with the aim to offer an increased accessibility compared to what would be possible through direct contact with the physical medium. However the attempt that has been made for a simpler user access and navigation on digitized texts is still far from being considered satisfactory. The browsing of the digitized collections as well as the locating of information meets many obstacles while the semantic enrichment of scanned texts through interconnections with other scanned or editable texts is practically non-existent. The established practice for indexing is based on the manual tagging of the digitized collections with MARC, MARCXML XML, TEI, JSON, SCOS (Park and Brenza 2015) and other notation formats along with heavy use of Relational Database Management Servers (RDBMS) which results to the introduction of a fairly considerable financial cost.

In the majority of cases the search returns links to a complete book or volume rather than links to specific pages or page sections which would be more desirable for a focused locating of information.

We should not also forget that the digitizing efforts in many cases have been very hasty and as a result only a part of the digitization is suitable for further automated use and processing. Double scanned pages or careless scanning that introduces skewing or warping is common to archives. Moreover digitised versions of the same collection appear with different faults when scanned by different entities which results to an additional complexity for processing. It is also customary during digitization to collect all of the pages in a consolidated archive which provides ease of copying and transferring the archives however does not facilitate the browsing, searching and semantic enrichment for the collections.
Before proceeding with our study it is useful to point out some points for better readability. Throughout this study the term "***semantic-enrichment"*** is interpreted mainly as the interconnections between scanned documents. The term "***satellite-texts"*** is defined as the set that contains certain documents which characterize and are relative to a specific collection. This set has the potential to enrich semantically the collection through proper interconnections.

Taking into account the above issues, a question and concurrently a range of difficulties arise for the Libraries. How is it possible to have an error-free collection with the best possible quality in order to provide automated *semantic enrichment* on large corpora by using their satellite-texts? As it is explained in following sections, a commonly accepted digitised version is a strong requirement to achieve further semantic enrichment by utilizing the advances in pattern recognition techniques and learning algorithms. Moreover, the creation of such commonly accepted high quality digitized collections will lead to a better browsing experience and direct locating, citing and disseminating the information since the automated building of the document structure or indexes will be significantly easier and less error prone.
Our discussion without losing its generalised applicability, since the nature of the difficulties are nearly similar for various corpora, is focused on large-sized historical scanned texts such as Migne's Patrologia Graeca so as to provide specific examples of applications and limitations.

This paper's discussion is organised as follows: In section 3, we provide a brief review for the Optical Character Recognition (OCR) and Word-Spotting (WS) methodologies by using either Neural Networks (NNs) frameworks or Image Features (IF) of the scanned texts.. In section 3, the PG collection and the importance of exposing this collection on the Web domain is described. In section 4, we present a detailed analysis of the difficulties that we faced on PG when we tried to accomplish at first a semantic enrichment through interconnections by using OCR on some of the satellite-texts of PG to get only the printed numbers of the pages. Secondly we tried to provide easy navigation and direct locating of information beyond the peculiarities that exists in the structure of PG pages. These peculiarities are also explained. In section 5, based on the analysis of the previous section, we present and describe a proposed architecture for libraries or cultural institutions in order to overcome the aforementioned difficulties we met on PG and to provide a general framework that can be used independently by research teams for the automated application of learning algorithms and pattern recognition techniques for semantic enrichment, creating page indexes, citing provision and better navigation experience.

2. Image Recognition Techniques

2.1 OCR and Word-Spotting differences

OCR systems aim to recognise a document at a character lever and the systems with most prominent role are the proprietary AbbyyFineReader and the open access Tesseract. Both of them are trained on a quite large range of fonts and languages and utilise as part of their recognition engines some kind of NNs in order to improve further the recognition rate. Particularly, Tesseract engine since the version 4.0 introduces Long short-term memory (LSTM) NNs in addition to the old recognition engine. In general, it is known that Latin scanned texts can be quite successfully recognized by OCR methods while on the other hand texts such as ancient Greek or manuscripts of various languages do not have so good results mainly because of the characters specificity, unknown printed fonts or degradation of the printed texts.

Due to the above OCR limitations, today there is a shift from OCR to word spotting (Ahmed, Al-Khatib, and Mahmoud 2017), that is the recognition of whole words in image texts in order to create indexes per page, by using techniques invariant with respect to rotation, scaling and translation. Word-spotting has been actively studied lately for handwritten texts (Giotis, Gerogiannis, and Nikou 2014)(Giotis et al. 2015) due to the great difficulties that these texts present, however it is not less applicable to digitized historical printed texts such as Patrologia Graeca (PG) which also provides difficulties for the OCR because of its polytonic script (Sfikas et al. 2015).

2.2 Neural Networks

NNs try to mimic the behaviour of the human brain and generally can cover a wide range of applications such as recognizing images, voice, text and patterns in general. The great advancement in the field of a special type of NNs the Convolution Neural Networks (CNNs) or alternatively Deep Learning (DL), opens up new possibilities for libraries in order to convert the multitude of scanned texts into better editable formats, provide search indexes for the scanned texts and a better user experience. However, DL beyond the great successes for recognising faces, animals and numbers or characters (Liu et al. 2017) its application in libraries is still in embryonic state.

As described in (Giotis et al. 2017) building a framework for a CNN to recognise images is not an easy task. The modelling of the CNN itself is trivial, however to achieve concrete results, big data with corresponding labels are required to train this CNN. Although the building of such data implies considerable cost, the successful application of CNNs on a specific data set does not guarantee its applicability to other datasets. New labelled data are required or even a new CNN model structure and therefore a desired generalised solution does not seem feasible at the moment.

2.3 Image Features

IF techniques rely on the spatial characteristics of the document in order to recognize shapes words or letters (Belongie, Malik, and Puzicha 2002) (Erdem and Tari 2010). IF techniques are very attractive because they do not require the vast amount of labelled data as required by NNs.

Various WS techniques are proposed utilising IF for handwritten and printed documents (Giotis et al. 2017) (Ahmed, Al-Khatib, and Mahmoud 2017). However there is not a unified technique for application to an extensive set of collections due to the particularity of the printed characters or the handwritten style or even the degradation of the collection, therefore a special approach is also required for each case.
Discussing the above limitations of both NNs and IF based on OCR and WS techniques, in no way implies that they are not applicable for semantic enrichment on scanned collections, however it is wise to keep in mind as aforementioned that their applicability is valid per case of the collection.

For the rest of this study and for better readability when there is a reference to the terms OCR or WS then that implies that either some kind of NNs or IF is used behind their implementation.

3. Patrologia Graeca

PG is an epic Collection of works by east Christian fathers over a period of 1400 years (Papadopoulos 1982). This collection consists of 166 volumes (bound as 161) and exists in digitized form from various sources, indicatively we refer (Google 2004)(Archive.org 1996), 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 while the rest 80% still exists, only in scanned images. More specific, the works of TLG found in (TLG 2000) contains 140 authors and 1524 works compared to 658 authors and 4,287 works identified by Perseus Digital Library (PDL) (Perseus 1987).

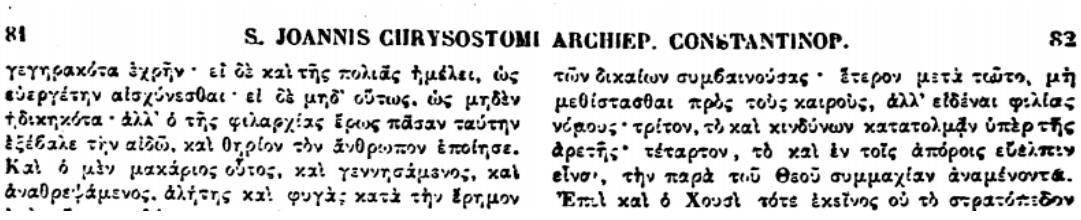
The PDL also provides works of PG, however, is far less comprehensive than TLG. A Greek archbishop, Dorotheos Scholarios (DS) published in Greek, a detailed Table of Contents (TOC) for PG named *"Κλείδα Πατρολογίας"* (Anemi 2006) with authors, small summaries and titles of chapters (Athens, 1879). He also published a more advanced interlinking, between specific words and authors (published in Athens, 1883) named *"Ταμείον Πατρολογίας"*.
Three decades later, in 1912, Garnier Frères in Paris, published a PG index volume, edited by Ferdinand Cavallera (FC) in Latin (Archive.org 1996). The works of DS and FC, are very

important on semantic level and semantic enrichment. Specifically in the work of DS as the user reads a topic in DS's work, next to the topic exists the page number of PG to locate the information. In general, by utilizing such characteristics on corpora a great semantic upgrade is provided by the use of the semantically created interconnections.
The interconnection of roughly120000 pages of PG with nearly 600 pages of DS's summaries is not at all a negligible issue.

4. Applied Methodology of Semantic Enrichment on Patrologia Graeca

4.1 Problem Description

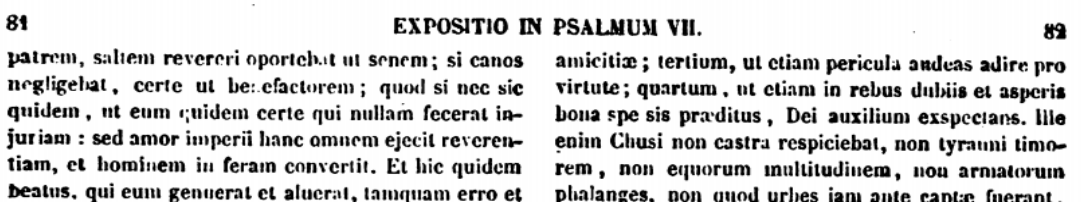
So far various versions of the same scanned collection exist throughout libraries or cultural institutions with differences in quality which is dependent by the physical medium condition and the scanning processes. More specifically for PG, the most well known digitation as mention in Section 1, has been done by Google (Google 2004) via Google books project, where the same volume is scanned more than one time as found in (Patristica.net 2019) links. These scans have variable quality between them as well as between the pages of each volume. Each volume is offered as a separate PDF file with a rough average number of 700 pages.
During the exploration of these PDFs we discovered a number of obstacles that prevent us from a further processing of these volumes until these difficulties are overcome. For example, double pages are found in the volumes by careless scanning or an effort to re-scan a badly scanned page. Skewing and warping is also custom phenomenon. Adding the degraded paper, ink dispersion and fade symptoms the situation becomes gradually more complicated. However the list does not end here, the internal structure of PG pages have some more peculiarities which are time consuming to solve, such as: The pages have two columns and have double numbering as it is shown in Figure 1 while some PG pages have the same numbering, one for the Greek and one for the Latin translation, see Figure 1 and Figure 2. Also some commentary pages in specific volumes are numbered by Latin number format, see Figure 3. The above issues should be taken into account, since we are interested to build a navigation system for PG as well as citing mechanism per page and semantic enrichment.


81 S. JOANNIS CHRYSOSTOMI ARCHIEP. CONSTANTINOP. 82

γεγηρακότα ἐχρῆν· εἰ δὲ καὶ τῆς πολιᾶς ἠμέλει, ὡς εὐεργέτην αἰσχύνεσθαι· εἰ δὲ μηδ' οὕτως, ὡς μηδὲν ἠδικηκότα· ἀλλ' ὁ τῆς φιλαρχίας ἔρως πᾶσαν ταύτην ἐξέβαλε τὴν αἰδῶ, καὶ θηρίον τὸν ἄνθρωπον ἐποίησε. Καὶ ὁ μὲν μακάριος οὗτος, καὶ γεννησάμενος, καὶ ἀναθρεψάμενος, ἁλήτης καὶ φυγὰς κατὰ τὴν ἔρημον

τῶν δικαίων συμβαινούσας· ἕτερον μετὰ τοῦτο, μὴ μεθίστασθαι πρὸς τοὺς καιροὺς, ἀλλ' εἶδέναι φιλίας νόμους· τρίτον, τὸ καὶ κινδύνων κατατολμᾷν ὑπὲρ τῆς ἀρετῆς· τέταρτον, τὸ καὶ ἐν τοῖς ἀπόροις εὐέλπιν εἶναι, τὴν παρὰ τοῦ Θεοῦ συμμαχίαν ἀναμένοντα. Ἐπὶ καὶ ὁ Χουσὶ τότε ἐκεῖνος οὐ τὸ στρατόπεδον

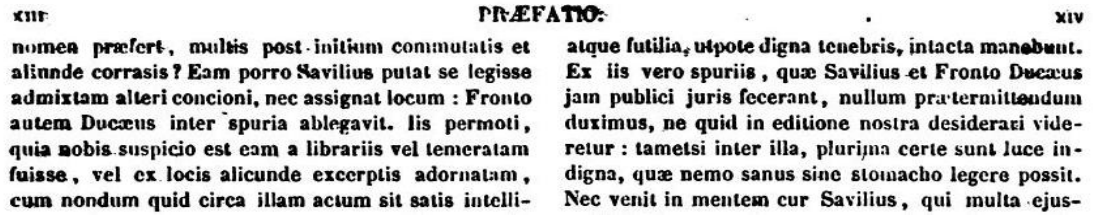
Figure 1. Double numbering of PG page with Greek text.


81 EXPOSITIO IN PSALMUM VII. 82

patrem, saltem revereri oportebat ut senem; si canos negligebat, certe ut benefactorem; quod si nec sic quidem, ut eum quidem certe qui nullam fecerat injuriam: sed amor imperii hanc omnem ejecit reverentiam, et hominem in feram convertit. Et hic quidem beatus, qui eum genuerat et aluerat, tamquam erro et

amicitiæ; tertium, ut etiam pericula audeas adire pro virtute; quartum, ut etiam in rebus dubiis et asperis bona spe sis præditus, Dei auxilium exspectans. Ille enim Chusi non castra respiciebat, non tyranni timorem, non equorum multitudinem, non armatorum phalanges, non quod urbes jam ante captæ fuerant

Figure 2. Double numbering of PG page with Latin text.



xiii PRÆFATIO. xiv

nomen præfert, multis post initium commutatis et alinnde corrasis? Eam porro Savilius putat se legisse admixtam alteri concioni, nec assignat locum : Fronto autem Ducæus inter spuria ablegavit. Iis permoti, quia nobis suspicio est eam a librariis vel temeratam fuisse, vel ex locis alicunde excerptis adornatam, cum nondum quid circa illam actum sit satis intelli-

atque futilia, utpote digna tenebris, intacta manebunt. Ex iis vero spuriis, quæ Savilius et Fronto Ducæus jam publici juris fecerant, nullum prætermittendum duximus, ne quid in editione nostra desiderari videretur : tametsi inter illa, plurima certe sunt luce indigna, quæ nemo sanus sine stomacho legere possit. Nec venit in mentem cur Savilius, qui multa ejus-

Figure 3. Commentary PG page with Latin numbering.

4.2 Proper Mapping of Patrologia Graeca Pages to the Physical Medium

The proper numbering and mapping of the digitised pages to the physical medium page of the original published collection is extremely important for the direct locating of information by the users. The physical medium numbering is followed by all the satellite-texts and consist the base for the interconnections. The methods we followed were mostly semi-manuals by careful examination of each volume. The PG volumes firstly are separated at page level and secondly the unwanted double pages are discarded. The proper mapping is created by taking into account the peculiarities of the pages structure and by using a JSON format which can be handled naturally (as build-in feature) by the JavaScript frameworks we used. The created clean set of pages of PG is called by using a Universal Resource Identifier (URI) in a memorable preferred way.

The URI is of the form: *http://myserver/patrologia/Volume/Column*. Having the correct mapping, a navigation system and a citation mechanism is easy to be build for the PG while on the same time it is a strong prerequisite for the second phase which is the semantic enrichment.

4.3 Semantic Enrichment of Patrologia Graeca Pages

Any document can be modelled as shown in Figure 4, having flows of semantic enrichment, since some documents enrich PG while some other are enriched by PG and vice versa. Discovering these relations and having the ability to navigate from one piece of information to another that resides to another part of the same collection or to another part of an external collection greatly increases the semantic weight for each collection and offers a better navigation experience of scholars and simple users. PG has semantic cross-connections with a large amount of documents. First of all is the bible and after that the various scholars publications that comment to the multitude of the PG authors after the PG was published. In this study we focus on three anthologies that add significant semantic enrichment to PG without the need to tag any of the PG pages or to use any other statistical method such as Bag-of-Words (BoW), Vector Space Model (VSM) etc. These three anthologies are:

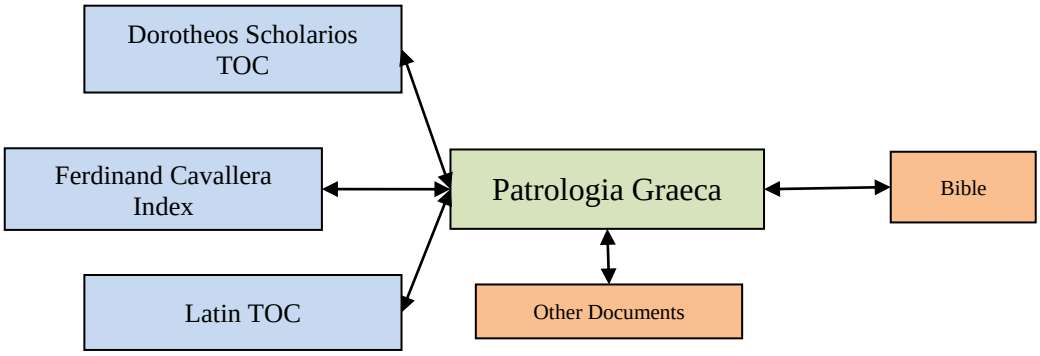


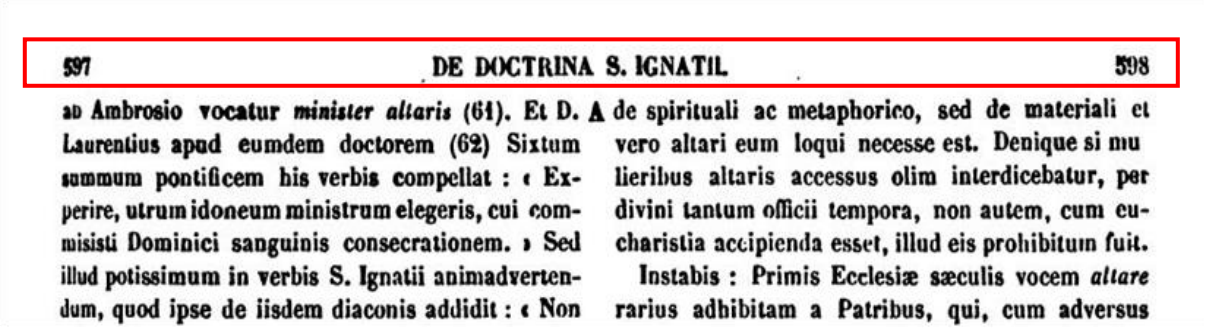
Figure 4: Semantic flows in documents

- The Table of Contents (TOC) by Dorotheos Scholarios (DS)
- The Latin TOC for PG (greatly simplified compared with that of DS)
- The Index created by Ferninard Cavalera (FC)

These indexes and TOCs act as an entrance to the vast amount of information kept behind PG corpus. Specifically, the TOC of DS is named *"κλείδα Πατρολογίας"* (key of Patrologia) because as a key unlocks the semantic information hidden behind the large number of PG pages. The above anthologies provide a brief summary for each chapter of PG and next to the summary resides the page number of PG to locate the information. Scholars for nearly one and a half century had these TOCs and Indexes as compass to find semantic information in PG.

The Latin TOC has been converted to editable form and found in (DocumentaCatholicaOmnia n.d.) while the Ferninard Cavalera's Index exist recently in XML format by Perseus Library. The TOC of DS, which is the richest in semantic subjects, is still in scanned images (Anemi 2006), however an attempt is made by the authors of this paper to recognise the numbers of the PG pages from the work of DS by using OCR techniques and create clickable hyperlinks on the images for automatic transferring to the specific page of PG. Two OCR frameworks were used to get the numbers: a) AbbyyFineReader (AbbyyFineReader 14 2019) and b) Tesseract (Tesseract 4.0 2019) with ancient Greek support. Preliminary tests showed that AbbyyFineReader performed satisfactory for the number recognition which we were interested compared to the poor Tesseract performance probably mainly due to the unknown fonts that had to handle. The above presented system is a work in progress (1/3 of PG volumes has been transformed to correct mapping structure), however the main functionality of the project can be tested (Varthis et al. 2019).

The authors have also interlinked the Latin TOC (Varthis et al. 2019) and strive towards to apply the same for the Index of FC in the near future. The main concept of the system is that the user can now read the indexes or the TOCs for an interesting subject and locate directly the specific information to the specific page. Moreover scholars have now the ability to reference or cite the specific page by its URI as it is in the physical medium.



597 DE DOCTRINA S. IGNATII. 598

ab Ambrosio vocatur *minister altaris* (61). Et D. Laurentius apud eumdem doctorem (62) Sixtum summum pontificem his verbis compellat : « Experire, utrum idoneum ministrum elegeris, cui commisisti Dominici sanguinis consecrationem. » Sed illud potissimum in verbis S. Ignatii animadvertendum, quod ipse de iisdem diaconis addidit : « Non de spirituali ac metaphorico, sed de materiali et vero altari eum loqui necesse est. Denique si mulieribus altaris accessus olim interdicebatur, per divini tantum officii tempora, non autem, cum eucharistia accipienda esset, illud eis prohibitum fuit. Instabis : Primis Ecclesiæ sæculis vocem *altare* rarius adhibitam a Patribus, qui, cum adversus

Figure 5: Page section of PG that contains the candidate pattern for recognition.

These satellite texts, as a future extension, could be also searched on their editable parts for a better user experience and extract semantic knowledge more efficiently. Summarising we can argue that, the semantic enrichment and navigation through interconnections although it gives tremendous semantic benefits on scanned documents, based on our experience, is a rather complicated issue and new frameworks, architectures and automated procedures should for sure to be developed.

5. Cooperative Architecture for Libraries for Semantic enrichment on Large Corpora

Although the difficulties discussed in the previous section for the enrichment of PG can be solved semi-manually as already presented for the specific case, they can also be addressed in order to apply on them NNs or IF techniques in an automated way. As we see in Figure 5, the upper section of each page follows a specific pattern that gives the columns numbers and a brief title for the page.

The utilisation of such patterns on specific page sections by the use of OCR or WS techniques can lead to the building of the structure of PG or other corpora which -as was mentioned earlier- is essential for the navigation, citing and consequently the semantic enrichment. However, such kind of applications require a new architecture for cooperation between libraries as well as between Libraries and research teams, at least for the collections that are on the public domain. The overview of this architecture is shown in Figure 6. The central point of the architecture is the repository that holds the best scanned version of PG. This repository can be updated only by the contributing libraries with the aim to include the best quality version for each page of PG. The PG has been already separated for each volume at page level and the minimum update can be done at page level.

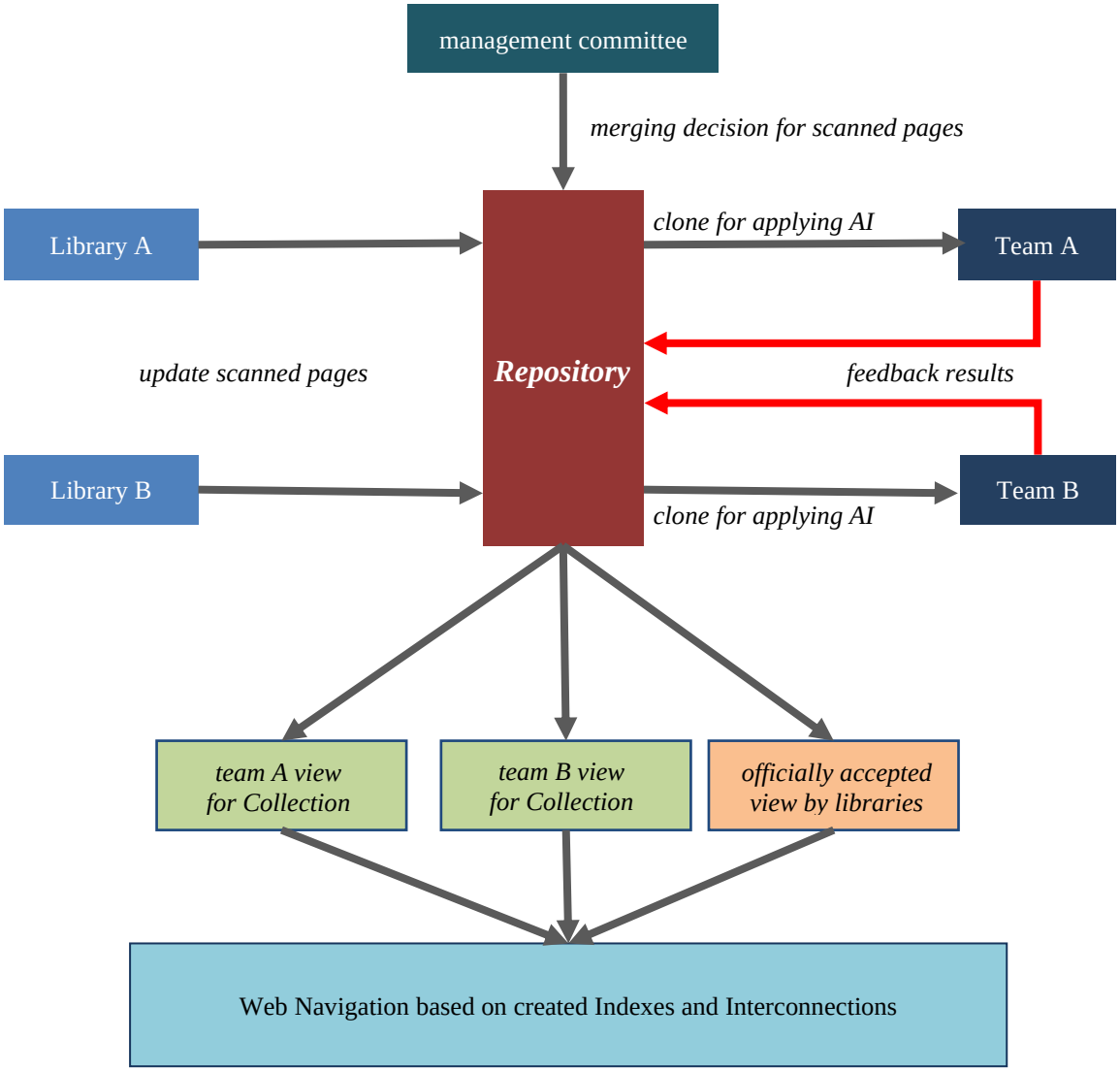


Figure 6: Cooperative Architecture between Libraries, Cultural Institutions and Research Teams.

The management committee decides for merging the current state of updates to the main public branch. The public branch offers the PG version that the user or research teams can at any time explore, navigate or clone. Any research team independently can clone the best version of PG or parts of it at a given time and apply locally any OCR or Word-Spotting technique on the PG. The results are uploaded back to a *"parallel"* repository which belongs to the specific team without re-uploading the files of the download collection since the collection already exists for public use on the Web domain. So the footprint of data is kept small while some very basic feedback results by using OCR or WS techniques can be:

- extracted text from specific pages.
- numbering of pages as they are numbered in the physical medium.
- automatic creation of page indexes, document's structure and TOCs.
- automatic interlinking of pages on the same or external different collections.

Any team has its own offered public view for the navigation on PG, completely separated from those of other teams while the user has the option to choose the public view he uses at his will. The officially adopted public view for the PG by the Libraries will be chosen by the management committee and will change when better results appear. This is essential because in this way a provision of a "live" system takes place that has the potential to be improved by the time. At the same time due to the fact that the compilation is public, there will be significant feedback for the specific NNs and IF techniques that are used, moreover they can be directly compared on the same dataset having significant space for evolving.

6. Conclusions

In this paper we presented a workflow in order to enrich the PG collection by using satellite texts of PG, as a showcase for the difficulties that arise. Although the difficulties found can be solved with the use of semi-manual methods that is not the proper methodology, since is time consuming and costly. An expansion has to be made to the Libraries services with the aim to automate such kind of very significant enrichments. An overview of the proposed architecture was presented in which independent research teams can contribute for the creation of page indexes, collection's page structure (TOCs), citing mechanism, semantic interconnections by using NNs or IF techniques implemented behind OCR and WS methods. The framework architecture can be considered as an evolving eco-system whose efficiency improves over time. All the results feed backed by the teams can be viewed, however the officially adopted view is chosen by the contributing libraries. This interaction between these frameworks provides a stable and significant ground for further comparison between various NNs and IF techniques on the Libraries domain.

References

AbbyyFineReader 14. 2019. “OCR System.” https://www.abbyy.com/en-ca/finereader/.

Ahmed, Rashad, Wasfi G. Al-Khatib, and Sabri Mahmoud. 2017. “A Survey on Handwritten Documents Word Spotting.” *International Journal of Multimedia Information Retrieval*.

Anemi. 2006. “Digital Libary of Modern Greek Studies.” https://anemi.lib.uoc.gr/metadata/8/5/0/metadata-01-0001289.tkl.

Archive.org. 1996. “Internet Archive Homepage.” https://archive.org.

Belongie, S., Jitendra Malik, and J. Puzicha. 2002. “Shape Matching and Object Recognition Using Shape Contexts-14pages.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

DocumentaCatholicaOmnia. “Latin Index of Patrologia Graeca.” *2011*. http://www.documentacatholicaomnia.eu/.

Erdem, Aykut, and Sibel Tari. 2010. “A Similarity-Based Approach for Shape Classification Using Aslan Skeletons.” *Pattern Recognition Letters*.

Giotis, Angelos P., Demetrios P. Gerogiannis, and Christophoros Nikou. 2014. “Word Spotting in Handwritten Text Using Contour-Based Models.” In *Proceedings of International Conference on Frontiers in Handwriting Recognition, ICFHR,*.

Giotis, Angelos P., Giorgos Sfikas, Basilis Gatos, and Christophoros Nikou. 2017. “A Survey of Document Image Word Spotting Techniques.” *Pattern Recognition*.

Giotis, Angelos P., Giorgos Sfikas, Christophoros Nikou, and Basilis Gatos. 2015. “Shape-Based Word Spotting in Handwritten Document Images.” In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR,*.

Google. 2004. “Books Library Project.” https://books.google.com.

Liu, Weibo et al. 2017. “A Survey of Deep Neural Network Architectures and Their

Applications.” *Neurocomputing*.
Papadopoulos, Stylianos. 1982. *Patrology, Vol. 1: Introduction, Second and Third Century (2nd. Ed.)*. Athens.
Park, Jung-ran, and Andrew Brenza. 2015. “Evaluation of Semi-Automatic Metadata Generation Tools: A Survey of the Current State of the Art.” *Information Technology and Libraries*.
Patristica.net. 2019. “PG Volumes List.” http://patristica.net/graeca/.
Perseus. 1987. “Digital Library Homepage.” http://www.perseus.tufts.edu/hopper/.
Sfikas, Giorgos, Angelos P. Giotis, Georgios Louloudis, and Basilis Gatos. 2015. “Using Attributes for Word Spotting and Recognition in Polytonic Greek Documents.” In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR,*.
Tesseract 4.0. 2019. “OCR System.” https://github.com/tesseract-ocr/tesseract.
TLG. 2000. “Thesaurus Linguae Graeca Homepage.” http://www.tlg.uci.edu/index.prev.php.
Varthis, Evagelos, Marios Poulos, Ilias Giarenis, and Sozon Papavlasopoulos. 2019. “Patrologia Graeca, Semantic Enrchment and Navigation.” http://patrologia.tk/kleida/.