CULDILE: Cultural Dimensions of Deep Learning, A Document Analysis System for Historical Documents

B. Gatos[1], G. Sfikas[1], P. Kaddas[1,2] and G. Retsinas[1]

[1]Institute of Informatics and Telecommunications,
National Centre for Scientific Research "Demokritos", GR 153 10, Athens, Greece
{bgat,sfikas,pkaddas, georgeretsi}@iit.demokritos.gr
[2]Department of Informatics and Telecommunications,
University of Athens, GR 157 84, Athens, Greece

Abstract. In this paper, an overview of the Greek National project CULDILE (CULtural DImensions of deep Learning) is presented. It includes a user-friendly software platform to analyze, enhance, index and provide access to a large number of historical document pages. CULDILE platform includes functionality for image pre-processing (image binarization and enhancement, page split etc.), automatic metadata extraction (e.g. detect the existence of handwritten or machine-printed text, tables, seals, signatures etc.), document classification and keyword spotting. The focus of this paper is on the specifications and architecture of CULDILE as well as on relevant general practices and tools.

Keywords: Historical Document Image Processing, Document Image Pre-processing, Document Metadata Extraction, Document Classification, Keyword Spotting.

1 Introduction

The CULDILE project[1] focuses on pioneering research activities in historical document image processing aiming to significantly improve access to historical documents and to take away the barriers that stand in the way of the mass digitization of cultural heritage documents. It includes a user-friendly software platform to analyze, enhance, index and provide access to a large number of historical document pages. A private new dataset from the library of the Piraeus Bank Group Cultural Foundation (PIOP)[2] is used to provide a first proof-of-concept. CULDILE platform includes functionality for image pre-processing (e.g. image binarization and enhancement, page split), automatic metadata extraction (e.g. detect number of columns, the existence of handwritten or machine-printed text, ornamental symbols, seals, signatures), document classification and keyword spotting (search by a keyword marked by the user). In this paper, we give an overview of relevant general practices and tools as well as of CULDILE specifications and architecture.

[1] http://culdile.bookscanner.gr
[2] https://www.piop.gr/el/vivliothiki.aspx

2 General practices and tools

At a first step we recorded all general practices and tools relevant to CULDILE research activities. This includes guidelines for digitization, tools for image annotation, platforms for document image visualization and metadata description schemes.

2.1 Guidelines for Digitization

Recommendations for selecting a particular format or standard for the digitization-related activities can be found in the following sources:

- IMPACT Centre of Competence[3], formats and standards related to master files, metadata, OCR results, delivery files, guidelines for semantic technologies, linguistic resources and tools packaging.
- JISC[4], guidelines for preparation of collection materials, copyright clearance, creation of metadata, scanning, web delivery, digital archiving and preservation.
- National Library of France (BnF)[5], guidelines for storing and processing digital collections, exploring and sharing resources, metadata management and catalogues.
- The National Archives and Records Administration (NARA), USA[6], recommendations for capture, minimum metadata, formats, naming, storage and quality control.

2.2 Tools for Image Annotation

Existing tools that can be used for document image annotation include:

- Aletheia[7], an advanced system for accurate and yet cost-effective analysis, recognition and annotation of scanned documents.
- labelme[8], a graphical image annotation tool using polygons written in Python.
- Computer Vision Annotation Tool (CVAT)[9], an interactive video and image annotation tool for computer vision.

2.3 Platforms for Document Image Visualization

Platforms that provide access to the page images of book and manuscripts include Open Library[10], Google Books[11], Many Books[12] and National Library of Greece, e-Reading Room[13].

---

[3] https://www.digitisation.eu
[4] https://digitisation.jiscinvolve.org/wp/
[5] https://www.bnf.fr/en
[6] https://www.archives.gov
[7] https://www.primaresearch.org/tools/Aletheia
[8] https://github.com/wkentaro/labelme
[9] https://github.com/openvinotoolkit/cvat
[10] https://openlibrary.org
[11] https://books.google.com
[12] https://manybooks.net
[13] https://ereading.nlg.gr/en/

2.4 Metadata Description Schemes

Metadata are usually described following schemes such as:

- Encoded Archival Description (EAD)[14]
- Dublin Core[15]
- Metadata Object Description Schema (MODS)[16]
- Machine Readable Cataloguing (MARC)[17]
- Metadata Encoding & Transmission Standard (METS)[18]
- CIDOC Conceptual Reference Model (CRM)[19]

3 System Specifications

In order to design the CULDILE platform, we took into consideration a long list of specifications that resulted after discussing with all involved partners (people from archives, industry and the research community). The most important are the following:

- The software should be user friendly and permit several levels of access (guest, authorized user, moderator, validator and admin) in a web-based and multi-threaded environment.

- Facilities for document image viewing on page or book/manuscript level should be provided as well as searching based on filters using metadata information.

- A list of pre-defined metadata should be supported (e.g. document category, existence of handwritten or machine-printed text, ornamental symbols, tables, seals, signatures) as well as custom metadata defined by the user. Metadata should be global (concern the whole document page, e.g. number of columns, letter color, existence of tables or images) or local (concern a certain part of the page defined by a polygon, e.g. an area containing handwritten text or a signature, see Fig.1a).

- All metadata should be filled in or edited by the user while all initial entries (values or/and defining polygons) should be automatically calculated by document image processing and deep learning methods that will be implemented and integrated in the platform. Re-training of these methods should be also provided based on selected existing data.

- A list of image pre-processing capabilities (e.g. image enhancement, page split) should be provided.

- Search by a keyword marked by the user should be also supported (query by example keyword spotting).

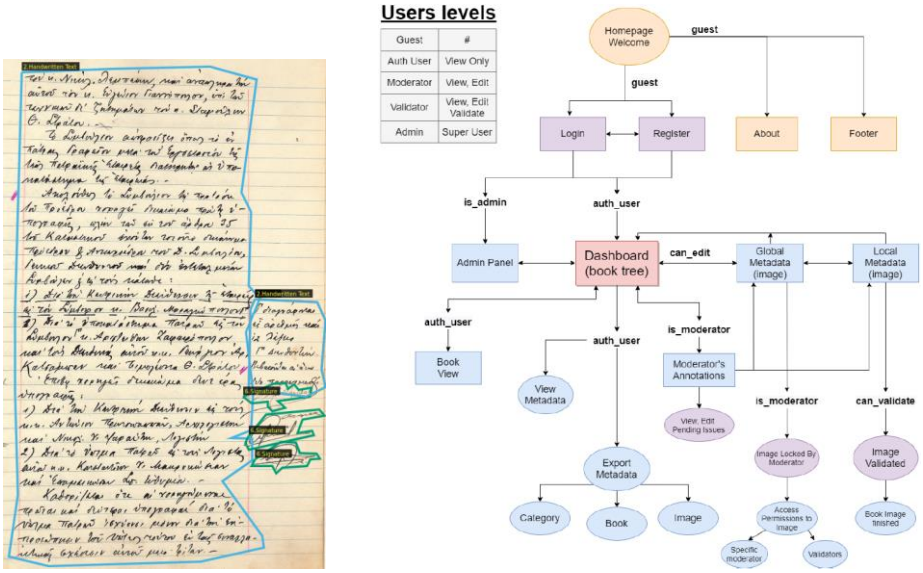
- All metadata should be saved in a convenient JSON format while export capabilities to the most famous metadata description schemes (see 2.4) should be supported.

[14] https://www.loc.gov/ead/
[15] https://dublincore.org
[16] http://www.loc.gov/standards/mods/
[17] https://www.loc.gov/marc/
[18] https://www.loc.gov/standards/mets/
[19] https://www.cidoc-crm.org

4 CULDILE Architecture

The architecture of the CULDILE platform is demonstrated in Fig. 1b. Different levels of access are supported with the following functionality:

- Guest: Access only at general information about the platform
- Authorized User: Read only rights, search and view data through the dashboard tree view, page browsing using thumbnails and book view, metadata view and export facilities.
- Moderator: authorized user with permissions to edit data, check and edit all global and local metadata, keep a record of actions done or pending, lock pages during processing.
- Validator: checks and validates all actions done by moderators, verifies all data that will be provided to all authorized users.
- Admin: Full access to all platform functionality, an admin panel is used to monitor all processes and actions.



(a) (b)

Fig. 1. (a) Example of local metadata defined by polygons. (b) CULDILE architecture overview

Acknowledgment

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH-CREATE-INNOVATE (project code: T1EΔK-03785) as well as by the program of Industrial Scholarships of Stavros Niarchos Foundation.