CULDILE: Cultural Dimensions of Deep Learning, a Document Analysis System for Historical Documents

15TH IAPR INTERNATIONAL WORKSHOP ON DOCUMENT ANALYSIS SYSTEMS
May 22-25, La Rochelle, France

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

[1]Computational Intelligence Laboratory,
Institute of Informatics and Telecommunications,
National Research Center "Demokritos",
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 It includes a user-friendly software platform to analyze, enhance, index and provide access to a large number of historical document pages.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.

The CULDILE project

The CULDILE project focuses on pioneering research activities in historical document image processing aiming to significantly improve access to historical documents and to take away the barriers that stand in the way of the mass digitization of cultural heritage documents. It includes a user-friendly software platform to analyze, enhance, index and provide access to a large number of historical document pages.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) 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.

CULDILE Architecture

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.

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 , formats and standards related to master files, metadata, OCR results, delivery files, guidelines for semantic technologies, linguistic resources and tools packaging.

➢ JISC , guidelines for preparation of collection materials, copyright clearance, creation of metadata, scanning, web delivery, digital archiving and preservation.

➢ National Library of France (BnF) , guidelines for storing and processing digital collections, exploring and sharing resources, metadata management and catalogues.

➢ The National Archives and Records Administration (NARA), USA , recommendations for capture, minimum metadata, formats, naming, storage and quality control.

Platforms for Document Image Visualization

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

System Specifications

➢ The software should be user friendly and permit several levels of access 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 as well as custom metadata defined by the user. Metadata should be global or local.

➢ 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. Retraining 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.

➢ All metadata should be saved in a convenient JSON format while export capabilities to the most famous metadata description schemes should be supported.

Tools for Image Annotation

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

➢ Aletheia , an advanced system for accurate and yet cost-effective analysis, recognition and annotation of scanned documents.

➢ labelme , a graphical image annotation tool using polygons written in Python.

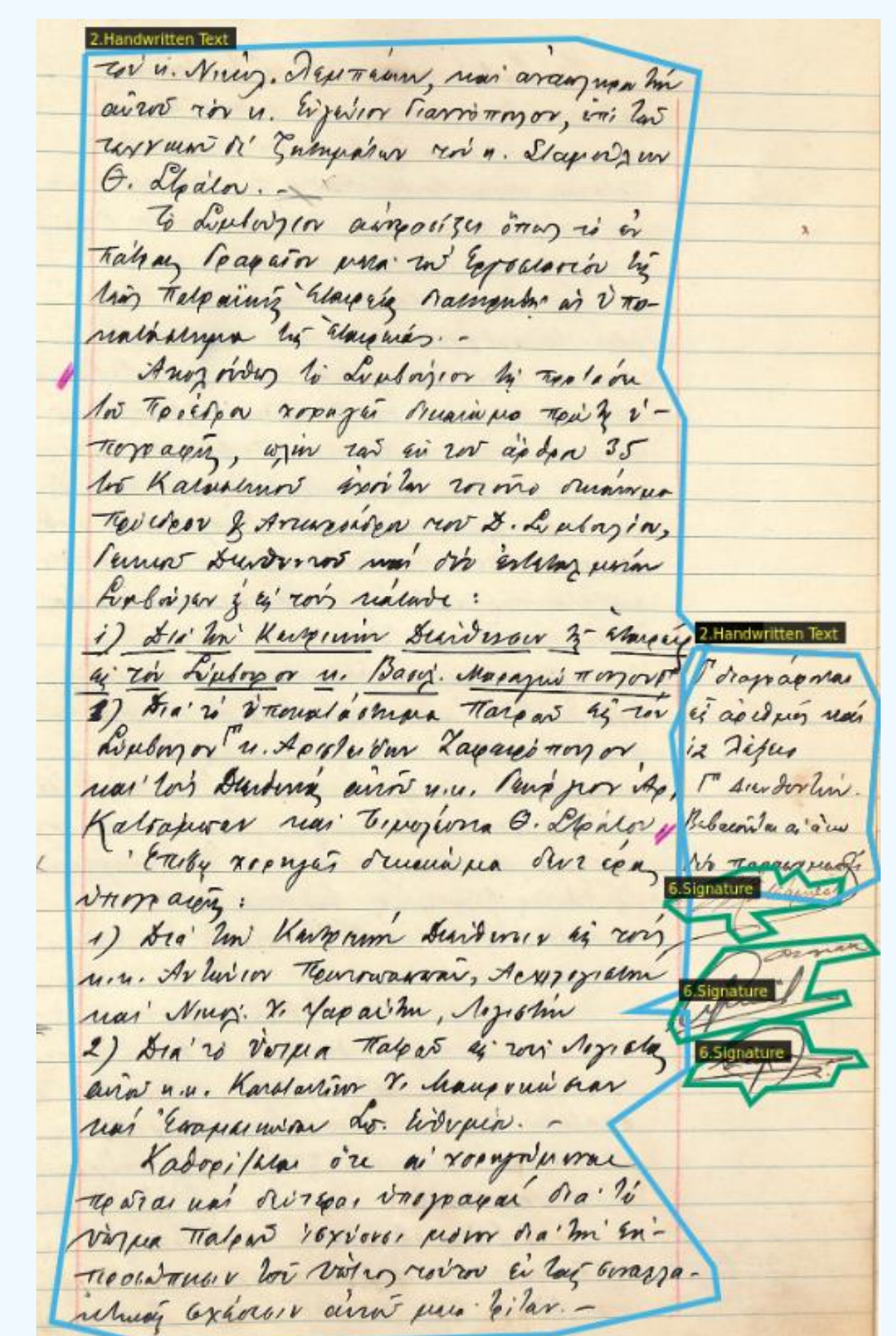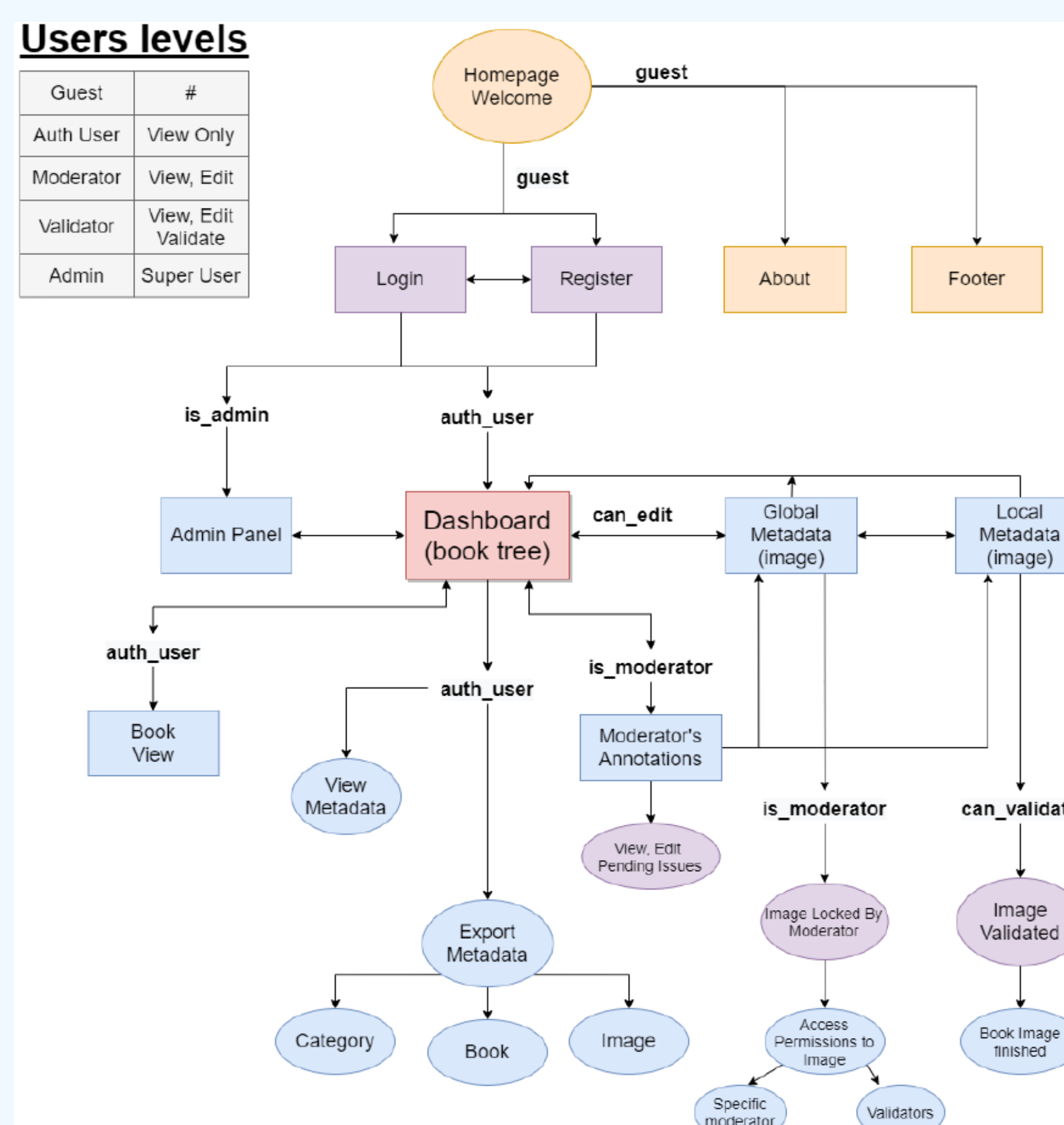➢ Computer Vision Annotation Tool (CVAT) , an interactive video and image annotation tool for computer vision.

Metadata Description Schemes

Metadata are usually described following schemes such as:

➢ Encoded Archival Description (EAD)

➢ Dublin Core

➢ Metadata Object Description Schema (MODS)

➢ Machine Readable Cataloguing (MARC)

➢ Metadata Encoding & Transmission Standard (METS)

➢ CIDOC Conceptual Reference Model (CRM)