The Lib2Life Platform – Processing, Indexing and Semantic Search for Old Romanian Documents

Irina Mitocaru, Gabriel Guțu-Robu, Melania Nițu, Mihai Dascălu, Ștefan Trăușan-Matu

University Politehnica of Bucharest

Splaiul Independentei 313, 060042, Bucharest, Romania

{irina.mitocaru, suzana_melania.nitu}@stud.acs.pub.ro
{gabriel.gutu, mihai.dascalu, stefan.trausan}@upb.ro

Silvia Tomescu

Carol I Central University Library Bucharest

Boteanu 1, 010027, Bucharest, Romania

silvia.tomescu@bcub.ro

Gabriela Florescu

National Institute for Research and Development in Informatics ICI Bucharest

Maresal Averescu 8-10, 011455, Bucharest, Romania

gabriela.florescu@ici.ro

DOI:10.37789/rochi.2020.1.1.3

ABSTRACT

Preserving the cultural heritage of a nation throughout generations is essential in a continuously developing society. This paper introduces the Lib2Life platform powered by advanced Natural Language Processing techniques, focusing on the processing, indexing and semantic search of old documents from the Central University Libraries in Romania. Our platform enables the upload and text pre-processing of scanned documents by librarians, who can afterwards manually correct the extracted content and corresponding metadata. In addition, Lib2Life ensures the exploration of the collection of books using a semantic search engine to retrieve documents fitted to the users' interests. The platform was evaluated using an usability questionnaire which pinpointed out that Lib2Life is a modern and user-friendly smart search engine for old documents written in the Romanian language. Improvements in terms of server response time and functionality were suggested. The platform proved to be intuitive and easy to use, having the potential to become an analytical system incorporating a rich and diverse collection of books.

Author Keywords

Library software; automated text extraction; Natural Language Processing.

ACM Classification Keywords

H.3.7: Information Systems: Information Storage and Retrieval: Digital Libraries.

H.5.1: Information Systems: Information Interfaces and Presentation: Multimedia Information Systems

General Terms

Text analysis; Software usability.

INTRODUCTION

The evolution of technology nowadays allows people to use electronic devices to access digitalized documents. Digitalization becomes mandatory due to the increased ubiquity of electronic devices, and their ease of use compared to traditional reading, as well as document research methods.

The Central University Libraries from Romania are dispersed throughout the country and host large numbers of old documents that are no longer copyrighted. These documents include books, manuscripts, or newspapers, and they serve to better understand life during those times in terms of politics, science, or education. Digital documents safeguard the initial manuscripts from being naturally deteriorated and provide an unlimited lifespan within a virtual environment. Moreover, access to the original documents can be limited to preserve degradation through manual handling. In contrast, digital documents can be stored in a convenient format that does not take too much storage space, such as the PDF. PDF documents have a structured format and contain also metadata, which is used either to properly display the document to the user (such as font types, font colors, or the physical coordinates of the words), or to store other annotations. Digitalization also eliminates distance and time limitations by allowing concurrent access to documents for individuals, regardless of their physical location or their time availability.

This paper describes the Lib2Life platform, which integrates Natural Language Processing (NLP) techniques and allows Romanian Central University libraries to store digitalized documents and provide access to resources to the public. The platform relies on a processing pipeline designed to properly support the extraction of texts from scanned documents. The paper continues with the description of related systems and text extraction applications. Then follows a presentation of our platform, its evaluation, conclusions and future work.

STATE OF THE ART

A well-known web application used for accessing online books is Google Books (http://books.google.com), which includes a large collection of books retrieved from different data sources. Google Books indexes data on their servers and provides access to portions of texts from the books. Links to buy or to allow further reading the entire document are included. Google Books is equipped with a smart search engine that uses NLP techniques to retrieve the most relevant books and corresponding passages, given the user query.

Google also provides a facility named Talk to Books, which compares the user query with every sentence in over 100,000 books to find responses that would most likely be related with that text or that could be an answer to the user's query. The suitable documents are then retrieved and shown in bold to the user, together with portions of text to provide a better contextualization. This approach came from the idea of mimicking a real conversation by using billions of lines of dialogue to teach an AI how human conversations flow. Their model can predict how likely a statement would follow another as a response based on a collection of possible responses [9; 14]. Similarly, the user can search for books in the Lib2Life platform using a free input text field, which is then semantically compared with portions of texts from the indexed books. Moreover, the user has the ability to find similar books using semantic similarity. The process is detailed further on in the Method section.

At national level, existing solutions to access digitalized documents from Central University Libraries (CULs) are outdated and provide rather limited functionalities (e.g., no search within the actual documents). For example, the Carol I Central University Library relies on Vubis [1] to save various metadata and Restitutio (http://restitutio.bcub.ro/), which is based on DSpace [15]. CUL Cluj has created their online catalog (http://aleph.bcucluj.ro:8991) designed to help people find the physical location of a book easier when that desired book is available in the library. Users have to provide search metadata, such as the book's identifier, the name of the author, the book's title, the publishing house, or the publishing year. Multiple databases are available, like book catalogues or bibliographies. Results can be further restricted by using specific filters, such as the language, time constraints (specific publication period), or the book's publication domain. BCU Cluj also has a digital library (http://dspace.bcucluj.ro) based on DSpace that provides online access to books contained in the physical library. The platform is accessible through a user-friendly and more intuitive interface than the interface provided by the Vubis catalog system. Each book can be read online using a PDF reader incorporated in the browser, or it can be downloaded on user's computer. The intuitiveness behind the incorporated filters and the multitude of options also served as an inspiration for developing the Lib2life platform, which incorporates a similar approach of filtering criteria for the semantic search engine.

Recommender systems are frequently applied in domains like online shopping and entertainment to predict user preferences. However, the same approach can be used to recommend books by relying on user profiles [13]. Their proposed recommended system takes into account multiple aspects for matching a book with a user. The first aspect consists of matching other users' interests, while the second refers to considering the temporal dimension. Specifically, the temporal dimension relies on the fact that user's preferences change over time.

The previously presented systems helped us into shaping the Lib2Life platform based on the requirements of librarians. Lib2Life provides three major functionalities. First, the system's focus is to centralize documents from multiple sources, more specifically from Romanian Central University Libraries, offering a single point of access for their entire data. Second, Lib2Life incorporates a multitude of NLP approaches, like a semantic search engine. Third, the platform also relies on an ontology for properly categorizing documents.

METHOD

Corpora

The Central University Libraries in Romania built up a collection of about 2,000 scanned documents written in Romanian. This dataset consists mostly of books dated in the 19th or 20th century scanned using high resolution scanners. However, part of the collection was not in a proper format for applying Optical Character Recognition (OCR) due to human errors or scanner issues. In addition, the OCR process encountered problems due to the limited resolution of the documents in some cases, or their degraded physical format.

The OCR process relies on the Tesseract API [16] applied on scanned documents before uploading them to the Lib2Life platform. The API was adapted to allow a good precision in detecting characters and to compress the file into a rather small size PDF document in the end. This requirement also resulted after an iterative process of understanding limitations and improving the general workflow. Large documents (i.e., tens of megabytes) were time consuming when uploaded to the Lib2Life platform, when processing them, and also when accessing them via the integrated PDF viewer in the web browser. Currently, the size of an OCR-ized document is about 10-15 MB.

The target of Lib2Life is to share the cultural heritage of millions of processed historical pages existent in CULs. Nevertheless, the Carol I Central University currently hosts about 2.4 millions of volumes (http://www.bcub.ro/colectii). However, part of the collection contains documents with publishing rights which cannot be used in our platform.

Architecture

The Lib2Life platform contains a web portal that allows librarians and users to interact with the system. The architecture of the platform is presented in Figure 1. The backend of the system integrates several digital services, for example: 1) document categorization; 2) semantic search; 3) semantic recommendations of similar documents. Assigning a category to the document is performed after uploading it and setting its corresponding metadata. The service is based on the Lib2Life ontology [7], which incorporates several domains and the relations established between them. The second service includes the facility to search the indexed documents using filtering criteria or keywords. Semantic algorithms are used to find documents matching the user's keywords. The third service consists of semantic recommendations – similar documents with the accessed one are provided. Both document search and semantic recommendations rely on Elasticsearch indexing [6]. Elasticsearch (https://www.elastic.co) is a non-relational database that stores and indexes the documents' metadata and their content using the JSON format. Elasticsearch provides fast and easy to use queries, filters, and aggregations mechanisms, which were incorporated in the search functionalities provided by the Lib2Life platform.

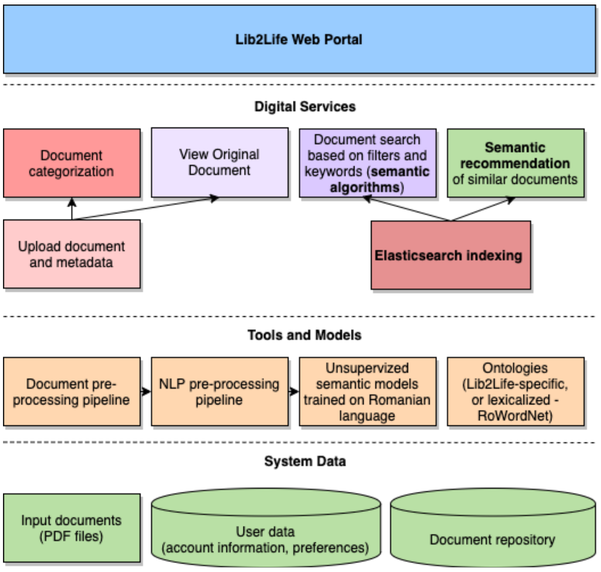

Lib2Life Web Portal
Digital Services
Document categorization
View Original Document
Document search based on filters and keywords (semantic algorithms)
Semantic recommendation of similar documents
Upload document and metadata
Elasticsearch indexing
Tools and Models
Document pre-processing pipeline
NLP pre-processing pipeline
Unsupervised semantic models trained on Romanian language
Ontologies (Lib2Life-specific, or lexicalized - RoWordNet)
System Data
Input documents (PDF files)
User data (account information, preferences)
Document repository

Figure 1. Lib2Life platform architecture.

Tools and models from the ReaderBench framework (http://readerbench.com) [8] are used to provide the previous services. The document pre-processing pipeline refers to the extraction of metadata from the OCR-ized document. The NLP pre-processing pipeline consists of several steps, such as tokenization, part of speech tagging, and lemmatization. The considered unsupervised semantic models include Latent Semantic Analysis [4], Latent Dirichlet Allocation [2], and word2vec [11]. These models are trained on language-specific corpora of documents.

The last layer from the Lib2Life architecture considers data modeling, namely: OCR-ized PDF documents, personal users' data used to interact with the Lib2Life web portal, as well as information related to documents indexed in Elasticsearch.

Document Pre-Processing Workflow

Prior to indexing documents in Elasticsearch, text preprocessing steps are applied. The input data consists of old scanned books on which Optical Character Recognition (OCR) is applied. The OCR process brought challenges in order to allow proper extraction of texts. These included different font types and sizes identified in the same section or line of text, different styles for headers and footers in the same document, disruption of paragraphs, improper page breaks, loss of content structure, or misinterpretation of certain characters and hyphenated words. Currently existing systems are not designed to work with OCR-ized PDFs [12], raising challenges while trying to properly restructure the recognized text. The identified issues imposed the necessity of a workflow that can identify and correlate section titles with their content, recognize paragraphs boundaries, merge hyphenated words and accurately identify and extract images or tables. The Lib2Life document processing workflow (see Figure 2) is designed to index documents into Elasticsearch and facilitate the search for relevant resources based on keywords.

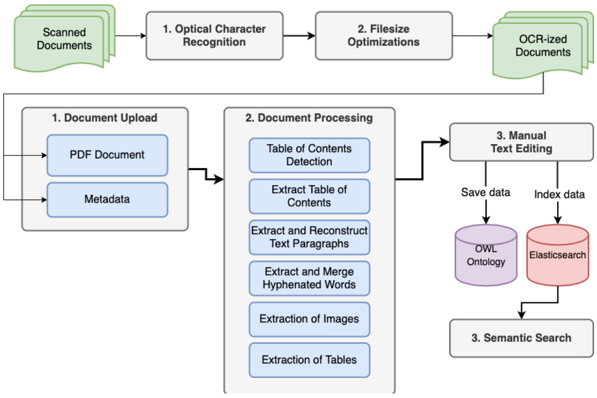
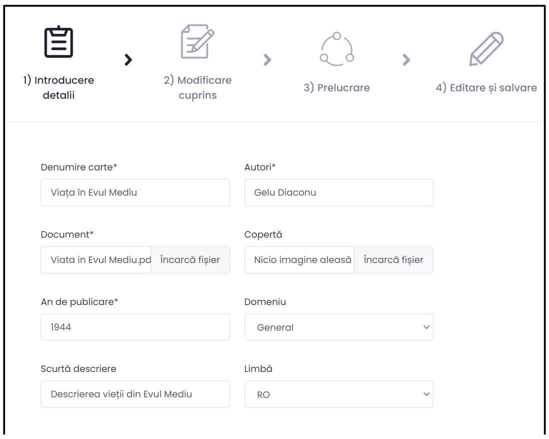


Figure 2. Lib2Life document processing workflow.

Documents are parsed line by line, identifying relevant sections and metadata within the document such as section titles, section headings, paragraphs, images, tables, and the table of contents. Paragraph boundaries are reconstructed, and hyphenated words are merged. The document title, the author, and the publishing year are extracted from the first page (if they are available) and are passed to the processing phase. The librarian has the facility to manually introduce a document's metadata when uploading the file – see Figure 3.

After the initial upload, three steps are performed: 1) detection and extraction of the table of contents and corresponding text; 2) information extraction using NLP techniques and heuristics; and 3) manual text editing.



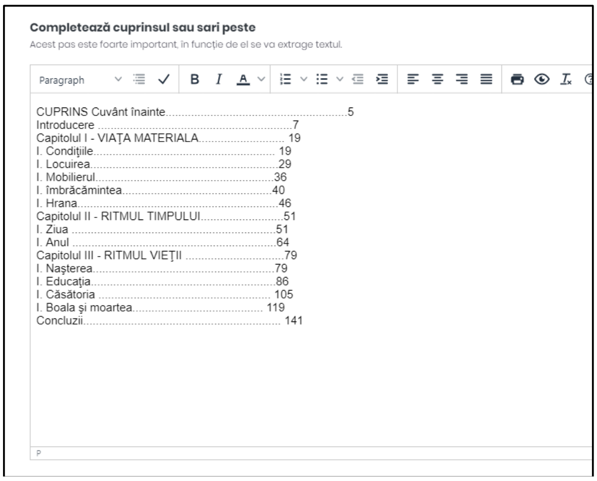
Denumire carte*: Viața în Evul Mediu
Autori*: Gelu Diaconu
Document*: Viata in Evul Mediu.pd Încarcă fișier
Copertă: Nicio imagine aleasă Încarcă fișier
An de publicare*: 1944
Domeniu: General
Scurtă descriere: Descrierea vieții din Evul Mediu
Limbă: RO

Figure 3. Document uploading and corresponding metadata.

Table of Contents Extraction

Two approaches were considered for extracting the Table of Contents (TOC): 1) correlating section titles with their content; 2) finding the predominant font type. The TOC extraction is performed by identifying the first page of the TOC within the first or last pages of a document. Specific words from the Romanian dictionary, such as „cuprins", „tabela", or „tabla de materii", are looked for. The OCR-ized text may contain errors like white spaces or symbols or may be split on several lines, which imposed additional validations using regular expressions. Mapping the section title with its corresponding page number was done in accordance with lines ending with digits. Additionally, empirical values were set for the number of lines ending with digits (more than 3), the max number of pages for TOC (10 pages). TOC entries are then parsed using regular expressions to extract the section title and the associated page range. Figure 4 shows an extracted TOC.

For documents that lack the presence of TOC, the second approach is used in order to identify sections and paragraphs based on the most common font available in the document.



Completează cuprinsul sau sari peste

Acest pas este foarte important, în funcție de el se va extrage textul.

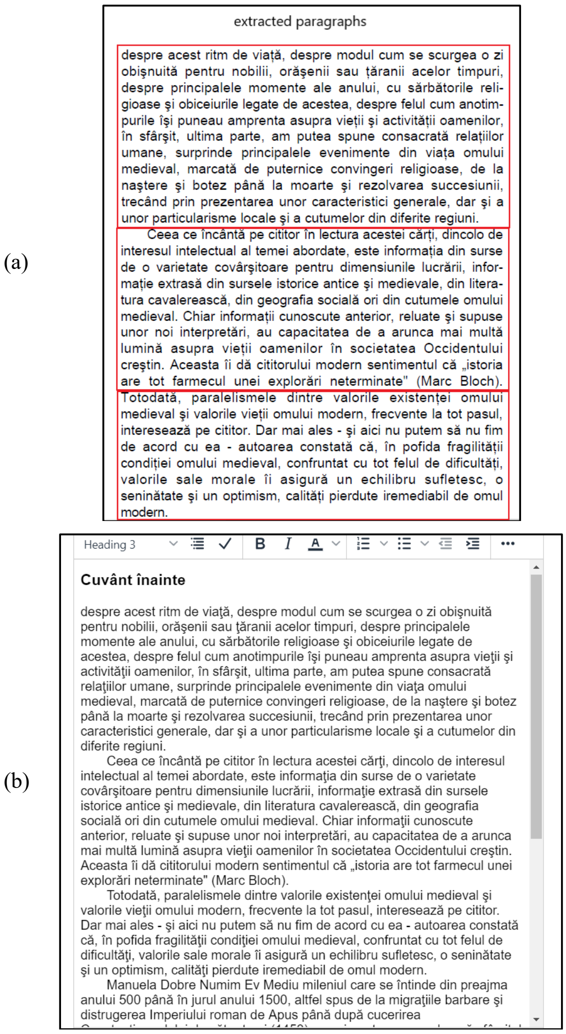
CUPRINS Cuvânt înainte..5
Introducere 7
Capitolul I - VIAŢA MATERIALA........................... 19
I. Condiţiile.. 19
I. Locuirea..29
I. Mobilierul..36
I. îmbrăcămintea...40
I. Hrana...46
Capitolul II - RITMUL TIMPULUI.........................51
I. Ziua 51
I. Anul 64
Capitolul III - RITMUL VIEŢII 79
I. Naşterea..79
I. Educaţia...86
I. Căsătoria 105
I. Boala şi moartea..................................... 119
Concluzii... 141

Figure 4. Table of Contents Extraction

Font name, font size, and text positions are stored in a list that is later on used to identify the type of the text (section title or body content), by comparing each line of text with the predominant font existing within the page. The two models were combined into a robust text extraction algorithm [12], that can easily adapt to most of the PDF document formats. The extracted text is then displayed in a rich text editor, enabling librarians to improve the extracted content by manually modifying the text.

Extracting and Reconstructing Paragraphs

At least one of the following conditions must be satisfied to detect a paragraph: 1) the previous line marks an ending sentence and the current line begins with an uppercase character; 2) the current line starts with a hyphen, depicting a conversation. A comparison of the original text versus the extracted text is shown in Figure 5.



despre acest ritm de viață, despre modul cum se scurgea o zi obişnuită pentru nobilii, orăşenii sau țăranii acelor timpuri, despre principalele momente ale anului, cu sărbătorile religioase şi obiceiurile legate de acestea, despre felul cum anotimpurile îşi puneau amprenta asupra vieții şi activității oamenilor, în sfârşit, ultima parte, am putea spune consacrată relațiilor umane, surprinde principalele evenimente din viața omului medieval, marcată de puternice convingeri religioase, de la naştere şi botez până la moarte şi rezolvarea succesiunii, trecând prin prezentarea unor caracteristici generale, dar şi a unor particularisme locale şi a cutumelor din diferite regiuni.

Ceea ce încântă pe cititor în lectura acestei cărți, dincolo de interesul intelectual al temei abordate, este informația din surse de o varietate covârşitoare pentru dimensiunile lucrării, informație extrasă din sursele istorice antice şi medievale, din literatura cavalerească, din geografia socială ori din cutumele omului medieval. Chiar informații cunoscute anterior, reluate şi supuse unor noi interpretări, au capacitatea de a arunca mai multă lumină asupra vieții oamenilor în societatea Occidentului creştin. Aceasta îi dă cititorului modern sentimentul că „istoria are tot farmecul unei explorări neterminate" (Marc Bloch).

Totodată, paralelismele dintre valorile existenței omului medieval şi valorile vieții omului modern, frecvente la tot pasul, interesează pe cititor. Dar mai ales - şi aici nu putem să nu fim de acord cu ea - autoarea constată că, în pofida fragilității condiției omului medieval, confruntat cu tot felul de dificultăți, valorile sale morale îi asigură un echilibru sufletesc, o seninătate şi un optimism, calități pierdute iremediabil de omul modern.

Cuvânt înainte

despre acest ritm de viaţă, despre modul cum se scurgea o zi obişnuită pentru nobilii, orăşenii sau ţăranii acelor timpuri, despre principalele momente ale anului, cu sărbătorile religioase şi obiceiurile legate de acestea, despre felul cum anotimpurile îşi puneau amprenta asupra vieţii şi activităţii oamenilor, în sfârşit, ultima parte, am putea spune consacrată relaţiilor umane, surprinde principalele evenimente din viaţa omului medieval, marcată de puternice convingeri religioase, de la naştere şi botez până la moarte şi rezolvarea succesiunii, trecând prin prezentarea unor caracteristici generale, dar şi a unor particularisme locale şi a cutumelor din diferite regiuni.

Ceea ce încântă pe cititor în lectura acestei cărţi, dincolo de interesul intelectual al temei abordate, este informaţia din surse de o varietate covârşitoare pentru dimensiunile lucrării, informaţie extrasă din sursele istorice antice şi medievale, din literatura cavalerească, din geografia socială ori din cutumele omului medieval. Chiar informaţii cunoscute anterior, reluate şi supuse unor noi interpretări, au capacitatea de a arunca mai multă lumină asupra vieţii oamenilor în societatea Occidentului creştin. Aceasta îi dă cititorului modern sentimentul că „istoria are tot farmecul unei explorări neterminate" (Marc Bloch).

Totodată, paralelismele dintre valorile existenţei omului medieval şi valorile vieţii omului modern, frecvente la tot pasul, interesează pe cititor. Dar mai ales - şi aici nu putem să nu fim de acord cu ea - autoarea constată că, în pofida fragilităţii condiţiei omului medieval, confruntat cu tot felul de dificultăţi, valorile sale morale îi asigură un echilibru sufletesc, o seninătate şi un optimism, calităţi pierdute iremediabil de omul modern.

Manuela Dobre Numim Ev Mediu mileniul care se întinde din preajma anului 500 până în jurul anului 1500, altfel spus de la migraţiile barbare şi distrugerea Imperiului roman de Apus până după cucerirea

Figure 5. Part of the a) original text contained within the document versus b) text extracted by the algorithm.

Image Extraction

The image extraction task is based on two approaches: 1) parsing image identifiers, and 2) searching for unusual shapes within a page and identifying the number of contained colors. For example, if a page contains only text in the middle top area and in the middle bottom of a specific area, with an irregularly shaped rectangle designed on white, gray, and black colors, then most probably the rectangle is an image, even if no identifier is found. Image and text extraction tasks were joined into one document iteration to reduce the time complexity; thus, image extraction is automatically applied. When parsing the text, three heuristics were applied. First, if a designated word for figure descriptions (e.g., “Figura” or “Fig.” in Romanian) is encountered, the location of the figure is saved. Second, we conducted an experiment on the number of characters existing in a page, which revealed the images were found on pages with less than 200 characters. Thus, the second heuristic compares the similarity of pixels and is applied on pages with no characters: if the pixels on the page are identical, the page is considered to be blank and it is thus skipped. Third, one of the following conditions must be satisfied to mark an entire page as an image: 1) the number of characters is zero and the pixels are different; 2) the character count is less than 200 and the text contains the figure identifier caption. The extracted images are converted to the base 64 format and the text caption is inserted into an HTML tag at the end of each section or at the end of the book, depending on the identified document structure.

Extraction of Tables

The table extraction task raised several challenges due to improper state of the OCR-ized PDF documents. An analysis performed on the collection of documents showed that most tables contained un-aligned lines, missing data, or an irregular structure. Some documents contained hand-written tables, on which the existing APIs are not able to accurately detect table boundaries, or the content itself. The current table extraction algorithm is Nurminen Detection Algorithm from Tabula API (https://github.com/tabulapdf), which showed an average accuracy of 40%. If a TOC is detected, table extraction is applied on each page, storing the coordinates of all tables and mapping the table with the corresponding section. If no TOC found, the details of the detected tables are stored, mapped with page numbers, and appended at the end of the section or of the book, depending on the identified document structure.

In addition, text cleaning steps are applied after extracting text from the document. These steps include removal of empty lines, removal of leading trailing spaces and other delimiter characters, concatenation of hyphenated words, appending white spaces for lines ending without whitespace, skipping pages with less than 400 characters (or 60 words, as in (Foundant, n.d.). Figure 6 shows an example of a page skipped because it contained too few characters.

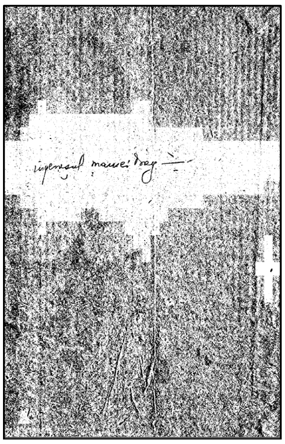
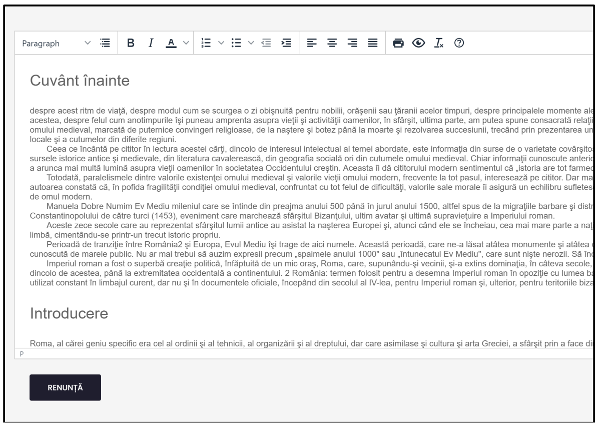


Figure 6. Page skipped because it contains too few characters.

Manual Content Editing

The refined text is sent to the user interface in an editable rich text area using TinyMCE (https://www.tiny.cloud), which enables the user to modify the extracted text before saving the file in Elasticsearch– see Figure 7. The processed text is then converted to JSON and sent to Elasticsearch for indexing. The indexed documents are used in advanced features, such as the keywords-based and semantic search.



Cuvânt înainte

despre acest ritm de viaţă, despre modul cum se scurgea o zi obişnuită pentru nobilii, orăşenii sau ţăranii acelor timpuri, despre principalele momente ale
acestea, despre felul cum anotimpurile îşi puneau amprenta asupra vieţii şi activităţii oamenilor, în sfârşit, ultima parte, am putea spune consacrată relaţi
omului medieval, marcată de puternice convingeri religioase, de la naştere şi botez până la moarte şi rezolvarea succesiunii, trecând prin prezentarea un
locale şi a cutumelor din diferite regiuni.
Ceea ce încântă pe cititor în lectura acestei cărţi, dincolo de interesul intelectual al temei abordate, este informaţia din surse de o varietate covârşito
sursele istorice antice şi medievale, din literatura cavalerească, din geografia socială ori din cutumele omului medieval. Chiar informaţii cunoscute anteric
a arunca mai multă lumină asupra vieţii oamenilor în societatea Occidentului creştin. Aceasta îi dă cititorului modern sentimentul că „istoria are tot farmec
Totodată, paralelismele dintre valorile existenţei omului medieval şi valorile vieţii omului modern, frecvente la tot pasul, interesează pe cititor. Dar ma
autoarea constată că, în pofida fragilităţii condiţiei omului medieval, confruntat cu tot felul de dificultăţi, valorile sale morale îi asigură un echilibru sufletes
de omul modern.
Manuela Dobre Numim Ev Mediu mileniul care se întinde din preajma anului 500 până în jurul anului 1500, altfel spus de la migraţiile barbare şi distr
Constantinopolului de către turci (1453), eveniment care marchează sfârşitul Bizanţului, ultim avatar şi ultimă supravieţuire a Imperiului roman.
Aceste zece secole care au reprezentat sfârşitul lumii antice au asistat la naşterea Europei şi, atunci când ele se încheiau, cea mai mare parte a naţ
limbă, cimentându-se printr-un trecut istoric propriu.
Perioadă de tranziţie între România2 şi Europa, Evul Mediu îşi trage de aici numele. Această perioadă, care ne-a lăsat atâtea monumente şi atâtea
cunoscută de marele public. Nu ar mai trebui să auzim expresii precum „spaimele anului 1000" sau „întunecatul Ev Mediu", care sunt nişte nerozii. Să înc
Imperiul roman a fost o superbă creaţie politică, înfăptuită de un mic oraş, Roma, care, supunându-şi vecinii, şi-a extins dominaţia, în câteva secole,
dincolo de acestea, până la extremitatea occidentală a continentului. 2 România: termen folosit pentru a desemna Imperiul roman în opoziţie cu lumea ba
utilizat constant în limbajul curent, dar nu şi în documentele oficiale, începând din secolul al IV-lea, pentru Imperiul roman şi, ulterior, pentru teritoriile biza

Introducere

Roma, al cărei geniu specific era cel al ordinii şi al tehnicii, al organizării şi al dreptului, dar care asimilase şi cultura şi arta Greciei, a sfârşit prin a face di

Figure 7. Editing the automatically extracted text.

After performing all the document processing steps, the book is saved and available to be accessed by users. Librarians can later on continue with editing the contents of the book and its metadata, while regular users can access its contents, perform searches and access the original PDF document.

Semantic Search

The search facility incorporated in the Lib2Life web portal relies on a semantic search algorithm. Due to the lack of annotations, the algorithm currently uses a K-Nearest Neighbors classifier with semantic distances based on word embeddings from ReaderBench. The workflow for semantic search is shown in Figure 8.

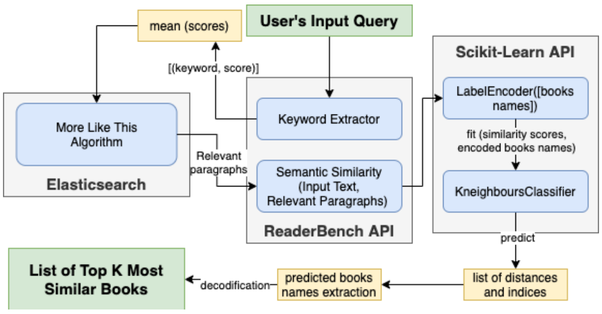

mean (scores)
User's Input Query
[(keyword, score)]
Scikit-Learn API
More Like This Algorithm
Elasticsearch
Keyword Extractor
LabelEncoder([books names])
fit (similarity scores, encoded books names)
Relevant paragraphs
Semantic Similarity (Input Text, Relevant Paragraphs)
KneighboursClassifier
ReaderBench API
predict
List of Top K Most Similar Books
decodification
predicted books names extraction
list of distances and indices

Figure 8. Semantic search algorithm.

The service extracts keywords from the input query text using the Keywords Extraction endpoint provided by the ReaderBench framework. The output consists of lemmatized content words, with their corresponding relevance score. This set of words is used to find the top k-nearest documents using the "More Like This" query incorporated in Elasticsearch. The result of this query consists of a list of paragraphs with at least one of the keywords. The similarity between the paragraph and the query embeddings is afterwards used to predict the top nearest documents. The corresponding document for each paragraph is retrieved, and a list with the most similar documents is returned (see Figure 9).

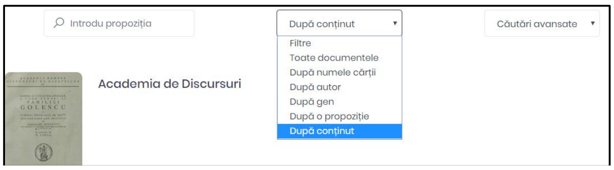


Figure 9. The semantic search functionality besides traditional search mechanisms.

The Web Portal

The Lib2Life platform provides a web interface developed for librarians and readers – see Figure 10. The User Interface (UI) is created using the Angular framework, version 9 (https://angular.io). Specific functionalities were developed for the two main roles: administrators (librarians) and readers. Librarians are able to upload, modify, and delete documents, while readers are only able to access the documents' content. Both librarians and readers can use the keyword-based search with filtering criteria, as well as the semantic similarity function to retrieve similar documents.

The web UI is connected to two application servers, one on Java, and another using the Flask framework developed in Python (https://github.com/pallets/flask). Both servers interact with an Elasticsearch instance for storing and retrieving indexed data. The ReaderBench API is used for performing advanced NLP processing, both for semantic search, and also for text extractions.

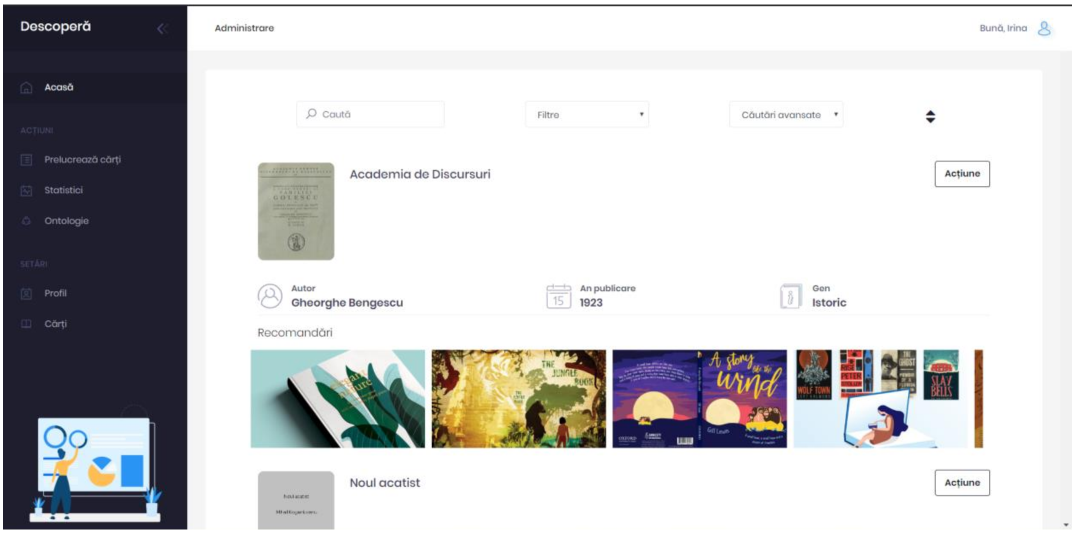


Figure 10. The Lib2Life dashboard.

The web portal incorporates an ontology viewer based on Protégé [10] – WebVOWL (http://vowl.visualdataweb.org/webvowl.html). The underlying ontology is used for text categorization and for better contextualizing the covered domains [7]. Figure 11 introduces the Lib2Life ontology viewer depicting a knowledge domain, namely Linguistics and Philology together with its corresponding subclasses.

RESULTS

A survey was conducted, and a questionnaire was distributed to evaluate the initial version of our platform. Twenty-three users aged between 20 and 50 years old, with a background ranging from students to Ph.D., working in a wide range of activity domains, were asked various questions about the platform and their experience with it. Demographic data showed that 54% of the respondents were aged 20-30, 21%

were aged 30-40, while 20% were aged higher than 40 years old. In terms of gender, 59% of users were women, while 41% were men. The education background included: high school – 4%, bachelor – 17%, master's degree – 54%, and Ph.D. – 25%. The activity domain ranged from IT in general – 34%, research in NLP – 29%, education – 29%, medicine – 4%, and graphic design – 4%. The users had to answer 14 Likert scale (1 – strongly disagree; 5 – fully agree) questions, which are presented in Table 1. Users found the web application intuitive, easy to use, and with a pleasing design. The questionnaire included four open-ended questions, allowing users to write opinions in natural language about what they liked or disliked, what features they missed, and what improvements should be performed to the application.

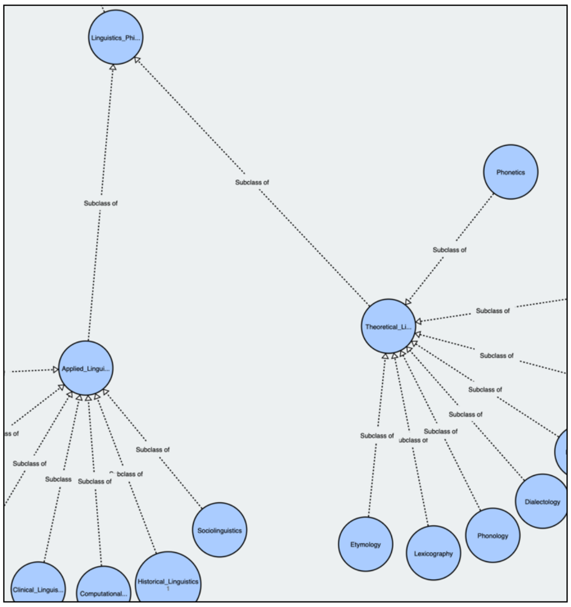


Figure 11. The Lib2Life ontology viewer.

Based on the answers to the open-ended questions, we found that the full document visualization page needs to be modified, both in terms of performance and of functionality. Users also requested having access to additional filtering criteria for browsing the collection of books. In addition, users encountered some errors and requested optimizations in terms of response time. Moreover, the implementation of a personalized bookshelf was also suggested.

A system limitation was met when text extraction algorithms did not work properly for some of the scanned documents. The solution relied on iteratively improving the OCR process. However, issues caused by improper scanning could not be always addressed.

In addition, the Lib2Life system did not differentiate amongst document types. Using another iterative process, the text extraction algorithm had to be constantly adapted based on characteristics shown by each new analyzed document. However, different categories or domains of documents may exhibit category-specific characteristics. Thus, we will consider improving the text extraction process by taking into account the domain of the document.

CONCLUSIONS AND FUTURE WORK

The Lib2Life platform aims to empower virtual restoration of historical documents owned by Central University Libraries in Romania by providing access to the digitalized documents in an online environment. Lib2Life currently stores about 100 documents provided by partner libraries which were manually corrected and used for testing.

Table 1. Feedback Questions.

#	Question	M (SD)
1	The Lib2Life application is intuitive.	4.57 (0.51)
2	The Lib2Life application is easy to use.	4.87 (0.34)
3	I could use the Lib2Life application without encountering errors.	4.13 (0.97)
4	The uploading and correction processes are easy to use.	4.52 (0.68)
5	The uploading and correction processes cover all the necessities.	4.43 (0.68)
6	The documents' visualization page is intuitive.	4.74 (0.69)
7	The search engine and the filters work as expected.	4.22 (0.80)
8	The search engine and the filters are intuitive.	4.22 (0.80)
9	The information included in the statistics page is useful.	4.90 (0.30)
10	The ontology is useful.	4.87 (0.34)
11	The Lib2Life application has an intuitive design and is suited to a system dedicated to libraries.	4.82 (0.50)
12	I would like to see more statistics about available documents.	4.43 (.843)
13	I would like to have more filters at my disposal on the dashboard page.	4.57 (.788)
14	I would like to be able to access documents from multiple domains and multiple languages.	4.91 (.288)

Lib2Life is a novel platform that includes useful filters for searching books, as well as the opportunity to read a document in PDF format. Users can also explore the entire domain ontology. The semantic search algorithm allows readers to find the most relevant documents for a query, or to be provided with similar documents to a selected one.

A usability questionnaire distributed to multiple users showed that the application is useful and includes a convenient semantic search functionality. The questionnaire argued that Lib2Life stands as a suitable software application to enable individuals to access digitalized historical documents. However, users requested improved response times, reducing error messages, and fixing server-related issues. In addition, a simpler user interface, but with more in-depth search criteria was also suggested.

The particularities of the old Romanian language used in the indexed historical documents should be also further explored. Namely, archaic words and structures should be considered in ReaderBench as the current version only supports contemporary language. This can be performed with the help of the eDTLR dictionary [3], which is an electronic Romanian dictionary containing more than 175,000 words. The temporal dimension should also be taken into account in a future version of the system, namely understanding how user preferences evolve over time, and followed by adjusting their recommendations.

Lib2Life enables librarians to build up a repository for their collection of documents. The resulting collection may be used for performing analyses focused on the evolution of literature across time. Example analyses include correlations to major historical events, changes in writing styles [5], as well as exploring inter-textual links between documents.

ACKNOWLEDGMENTS

This work was supported by a grant of the Romanian Ministry of Research and Innovation, CCCDI - UEFISCDI, project number PN-III-P1-1.2-PCCDI-2017-0689 / „Lib2Life - Revitalizarea bibliotecilor si a patrimoniului cultural prin tehnologii avansate” / "Revitalizing Libraries and Cultural Heritage through Advanced Technologies", within PNCDI III.

REFERENCES

1. Alewaeters, G., 1982. VUBIS: A user-friendly online system. *Information Technology and Libraries 1*, 3, 206-221.
2. Blei, D.M., Ng, A.Y., and Jordan, M.I., 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, 4-5, 993–1022.
3. Cristea, D., Răschip, M., Forăscu, C., Haja, G., Florescu, C., Aldea, B., and Dănilă, E., 2007. The Digital Form of the Thesaurus Dictionary of the Romanian Language. In Proceedings of the 4th International IEEE Conference SpeDIEEE, 195-206.
4. Crossley, S.A., Dascalu, M., and McNamara, D.S., 2017. How important is size? An Investigation of Corpus Size and Meaning in both Latent Semantic Analysis and Latent Dirichlet Allocation. In Proceedings of the 30th Int. Florida Artificial Intelligence Research Society Conf. (Marco Island, FL), AAAI, 293–296.
5. Gifu, D., Dascalu, M., Trausan-Matu, S., and Allen, L.K., 2016. Time Evolution of Writing Styles in Romanian Language. In Proceedings of the 28th Int. Conf. on Tools with Artificial Intelligence (ICTAI 2016) (San Jose, CA), IEEE, 1048–1054.
6. Gormley, C. and Tong, Z., 2015. *Elasticsearch: The definitive guide: A distributed real-time search and analytics engine*. O'Reilly Media, Inc.
7. Gutu-Robu, G., Ruseti, S., Tomescu, S.-A., Dascalu, M., and Trausan-Matu, S., 2020. Designing an Ontology for Knowledge-Based Processing in Romanian University Libraries. In Proceedings of the The 16th International Scientific Conference eLearning and Software for Education (Bucharest).
8. Gutu-Robu, G., Sirbu, M.-D., Paraschiv, I.C., Dascalu, M., Dessus, P., and Trausan-Matu, S., 2018. Liftoff - ReaderBench introduces new online functionalities. *Romanian Journal of Human - Computer Interaction 11*, 1, 76–91.
9. Hämäläinen, W. and Vinni, M., 2006. Comparison of machine learning methods for intelligent tutoring systems. In Proceedings of the Int. Conf. in Intelligent Tutoring Systems (Jhongli, Taiwan), Springer, 525–534.
10. Knublauch, H., Fergerson, R.W., Noy, N.F., and Musen, M.A., 2004. The Protégé OWL plugin: An open development environment for semantic web applications. In Proceedings of the International Semantic Web ConferenceSpringer, 229–243.
11. Mikolov, T., Chen, K., Corrado, G., and Dean, J., 2013. Efficient Estimation of Word Representation in Vector Space. In Proceedings of the Workshop at ICLR (Scottsdale, AZ).
12. Nitu, M., Dascalu, M., Dascalu, M.-I., Cotet, T.-M., and Tomescu, S., 2019. Reconstructing Scanned Documents for Full-text Indexing to Empower Digital Library Services. In Proceedings of the 12th Int. Workshop on Social and Personal Computing for Web-Supported Learning Communities (SPeL 2019) held in conjunction with the 18th Int. Conf. on Web-based Learning (ICWL 2019) (Magdeburg, Germany), Springer, 183–190.
13. Rana, C. and Jain, S.K., 2012. Building a Book Recommender system using time based content filtering. *WSEAS Transactions on Computers 11*, 2, 27-33.
14. Sebastiani, F., 2002. Machine learning in automated text categorization. *ACM Comput. Surv. 34*, 1, 1-47. DOI= http://dx.doi.org/10.1145/505282.505283.
15. Smith, M., Barton, M., Bass, M., Branschofsky, M., McClellan, G., Stuve, D., Tansley, R., and Walker, J.H., 2003. DSpace: An open source dynamic digital repository.
16. Smith, R., 2007. An overview of the Tesseract OCR engine. In Proceedings of the Ninth international conference on document analysis and recognition (ICDAR 2007)IEEE, 629-633.