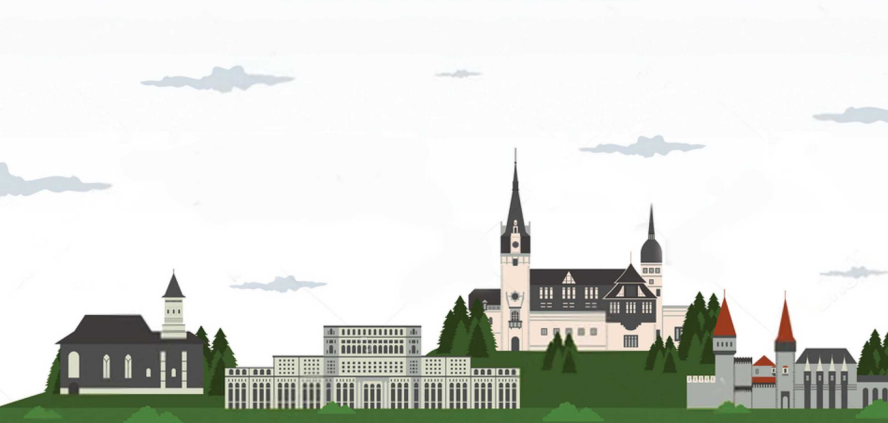


COORDONATORI:
MIHAI DASCĂLU, BOGDAN ȘANDRIC

HERITAGE IN THE DIGITAL ERA

CASES AND BEST PRACTICES FROM ROMANIA



ISTORIE ȘI
STUDII CULTURALE

Comisia Națională a României pentru UNESCO



Mihai DASCĂLU • Bogdan ȘANDRIC
(Coordonatori)

**Heritage in the digital era.
Cases and Best Practices from Romania**

Comisia Națională a României pentru UNESCO



unesco

National Commission
of Romania for UNESCO

Comisia Națională a
României pentru UNESCO

Mihai DASCĂLU • Bogdan ȘANDRIC
(Coordonatori)

Heritage in the digital era

Cases and Best Practices from Romania



Editat de **Pro Universitaria SRL**, editură cu prestigiu recunoscut.
Editura **Pro Universitaria** este acreditată CNCS în domeniul Științelor Umaniste și CNATDCU (lista A2-Panel 4) în domeniul Științelor Sociale.

Copyright © **2021, Editura Pro Universitaria**

Toate drepturile asupra prezentei ediții aparțin **Editurii Pro Universitaria**.
Nicio parte din acest volum (fragment sau componentă grafică) nu poate fi copiată fără acordul scris al **Editurii Pro Universitaria**.

Descrierea CIP a Bibliotecii Naționale a României

Heritage in digital era : cases and best practices from Romania /
coord.: Mihai Dascălu, Bogdan Șandric. - București : Pro Universitaria,
2021

ISBN 978-606-26-1486-7

I. Dascălu, Mihai (coord.)

II. Șandric, Bogdan (coord.)

004

008

Redactor:	Elena Onea
Tehnoredactor:	Liviu Crăciun
Copertă:	Vlad Pătruță



Redacție:

tel.: 0732.320.664

e-mail: editura@prouniversitaria.ro



Editura Pro Universitaria



Librăria Ujmag:

tel.: 0733.673.555; 021.312.22.21

e-mail: comenzi@ujmag.ro

ujmag.ro



Ujmag.ro

Contents

INTELLIT – Modeling Literary Trends in Romanian History.....	7
Laurentiu-Marian Neagu, Irina Toma, Laurentiu Hanganu, Lucian Chisu, Mihai Dascalu, Ștefan Trausan-Matu, Eugen Simion	
Advanced Natural Language Processing Techniques for Restoring Old Romanian Documents	25
Silvia Tomescu, Irina Mitocaru, Gabriel Guțu-Robu, Melania Nițu, Mihai Dascălu, Ștefan Trăușan-Matu	
Computer-assisted methods in historical linguistics	41
Alina Cristea, Anca Dinu, Liviu P. Dinu, Simona Georgescu, Ana Uban	
3D Roma – Sarmizegetusa. Turn on the History. From the data collection on the field to the deployment of a Virtual Museum.....	57
Antal E., Bota C., Ciongradi E., D’Annibale E., Demetrescu C., Dima B., Fanini D., Ferdani	
Heritage at the Crossroads. Digitization of Stone Crosses in Prahova County, Romania	76
Marius Streinu, Oana Borlean, Mihaela Buruiana, Liviu Mihail Iancu, Bogdan Șandric	
Good practices in the ProEuropeana Digital Library https://biblioteca-digitala.ro/	99
Vasile Andrei, Bogdan Șandric, Cosmin Miu, Oana Borlean, Marian Tufaru, Florentina Ghemuț	
The E-culture Project: the Culturalia platform and quantitative ambitions.....	111
Dan Matei, Bogdan Șandric	
The Virtual Genealogical Archive: a pilot digitization project of the parish and civil status registers in the Bucharest and Brașov county archives (Romania) - http://arhgenvirt.ro.....	118
Rafael-Dorian Chelaru	
The Special Fund deposit of the National History Museum of Romania - a history of the Ceausescu era in gifts received from abroad	131
Cristina Păiușan-Nuică	

INTELLIT – Modeling Literary Trends in Romanian History

Laurențiu-Marian Neagu¹, Irina Toma¹, Laurentiu Hanganu², Lucian Chisu²,
Mihai Dascalu¹, Ștefan Trausan-Matu¹, Eugen Simion²

(¹ University Politehnica of Bucharest)

(² Institute of History and Literary Theory “G. Călinescu”)

Every national literature contains compositions in verse or prose that reflect the history and express the collective cultural identity of a nation. Thus, it is essential for each member of the community to be acquainted with the most important national writers and their works. With the development of online environments, there is a growing need to enhance the teaching process, to shift to e-Learning environments by providing online materials, statistical studies, and methods for evaluating the acquired knowledge. All information on writers and literary entities that contributed to the Romanian literature is available in the General Dictionary of Romanian Literature (DGLR), a project developed by The Romanian Academy in a series of 7 published volumes. Together with the writers and literary entities, another comprehensive work developed by The Romanian Academy presents the important literature-related events which happened after World War II nationwide and is available in the Chronology of Romanian Literary Life (CVLR). Both DGLR and CVLR are brought to the digital format through the INTELLIT web platform which, besides the extracted general information, offers also visualizations for quantitative analyses, such as: writers' demographic information (birthplaces map, birth and death years plots), interactive map with trips for the most important writers, semantic distances or similarities between writers and clustering based on their descriptions from DGLR and the evolution of literary trends after World War II.

1. Introduction

Literature is a key element in defining a community's culture. The national literary heritage is defined as the totality of works produced by all writers, including poets, prose writers, playwrights, publicists, or cultural journalists who brought spiritual value to the society. Heritage literature helps individuals understand themselves, while playing a fundamental role in guiding the community members through the process of acquiring cultural values and learning cultural codes. The Romanian Academy successfully created a unified and comprehensive perspective of all writers that published in Romanian in the *General Dictionary of Romanian Literature* (DGLR) (Simion, 2004-2009). Besides writers, the dictionary includes information about publications, concepts, literary movements, anonymous writings, literary groups and institutions, translations into and from Romanian. In 2016, two volumes from a second edition of the DGLR were published, containing up-to-date information about the Romanian literature (Simion, 2016-2018). In addition, the *Chronology of Romanian Literary Life* (CVLR) (Simion and Grigor, 2010-2012)(Simion and Chișu, 2017-2020) is another project developed by the Romanian Academy as a complement to the DGLR. This work contains literature events that occurred after World War II and had an impact upon the national culture.

As technology evolves, people are more drawn to online environments. Resources, such as library databases and book archives are already available online. According to a statistical study (Statista, 2020a), in 2020 approximately 118 million Europeans bought eBooks and it is expected to increase up to 127 million by 2025. In Romania, these numbers are considerably lower – less than 6% individuals purchase books, magazines, and e-learning materials online (Statista, 2020b). However, 7 out of 10 Romanians download eBooks free of charge, without registration (Statista, 2019). The DGLR and CVLR follow the identified trend and are moved to the online environment by digitalizing their entire content and posting it online in an easily accessible manner through the INTELLIT platform¹. Besides presenting the information extracted from DGLR and CVLR, the aim of the project is to provide an overview of the national literature through quantitative studies. All analyses and the INTELLIT web platform are available online, free of charge.

¹ <https://intellit.ro/>

The current chapter is structured as follows. After introducing the DGLR, the CVLR and the INTELLIT web platform, the following section presents similar recent works on text processing, with several analyses and types of visualizations. Next is the Method section, which describes the used corpus, types of processed documents, the database structure and processing pipelines. Afterwards, the Results section presents the experimental plots, graphs, maps, timeline views for quantitative analyses, a 3D interactive view for the semantic distances or similarities, and a subset of results for the evolution of trends across time. The last section outlines the conclusions, and future work directions are proposed.

2. State of the art

Most literature analyses available in the online environment target individual elements from literature, such as novels, poetry, or literary movements. The more complex ones compare writers or genres, but few target the literature as a whole, as it is difficult to cover its evolution in a single work. (Moretti, 2000) introduced a new perspective of analyzing literature from a quantitative point of view, called distant reading. The method requires the abstraction of the actual text moving the focus on the interconnectivity between elements, defining shapes, relations, structures, forms, and models that (Moretti, 2005) represented in a visual manner: graphs, maps, and trees. Moretti used graphs to represent quantitative history, such as the number of new novels per year, but also to identify trends, and correlations between literatures. For example, in mapping novelistic trends, the target is to identify historical patterns in the evolution of literature. In terms of maps, events happening in British novels are transposed in a geographical map to emphasize the main points of interest and to identify the pattern of events. Lastly, trees were used to depict the evolutionary theory transposed in literature. Moretti used British detective fiction to identify divergencies and convergencies in the paths that detectives follow, concluding that convergence develops only when the distance to the original paths is not significant. Moretti's work was continued and strengthened through a collection of Pamphlets (Moretti, 2016, Algee-Hewitt, 2021, Allison et al., 2011) developed at Stanford Literary Lab using automatic tools for textual analysis.

Following Moretti's distant reading concept, several studies were conducted on different literatures. (Lansdall-Welfare et al., 2017) use textual analysis techniques to evaluate a corpus formed by regional

newspapers from the United Kingdom, spanning on approximately 150 years. Information such as historical events (e.g., wars, coronations, epidemics) was extracted using simple content analysis, searching for specific keywords. For example, the frequency of concepts such as “fight”, “war”, “troop”, and “soldier” accurately identified wartime frames. Deeper analyses targeted several concepts, such as values and beliefs, politics, social change, and popular culture. Through these analyses the authors were able to test existing hypotheses such as the decline of “Victorian values”.

Another quantitative study by (Reagan et al., 2016) extracts a set of six core emotional arcs from a corpus of 1327 British stories. The purpose of the study was to identify which of these emotions are reflected in today’s publications, and how they impact the popularity of the publication measured through the number of downloads. Each emotional arc is represented through several points denoting the level of emotion (rise – positive emotion, fall – negative emotion) and is associated a name. The identified emotional arcs include rags to riches (rise), tragedy (fall), man in a hole (fall-rise), Icarus (rise-fall), Cinderella (rise-fall-rise) and Oedipus (fall-rise-fall).

Quantitative studies evaluate literature in numbers and percentages. A study conducted by (Sinykin et al., 2019) targets economics, genre, and race in the American post-war novels. The results of the study highlight women use less economic terms than males, while African Americans use less economic terms than Caucasians. Other studies evaluate the similarities between authors’ writing styles (Eder, 2017), or analyze the relationship between the writer’s gender and book genres (Thelwall, 2017).

The recent advances in Natural Language Processing (NLP) empower the binding of the educational field, and more precise, the literature field, with Machine Learning algorithms. The semantic distance between two texts can be computed in several ways, as it was presented in an empirical evaluation performed by (Lee et al., 2005). Relatedness can be computed with simple word-based, keyword-based or n-gram measures, or with more complicated approaches, such as the Latent Semantic Analysis (LSA) (Dumais, 2004). (Lilleberg et al., 2015) have explored other approaches for extracting features for documents and they have presented word2vec (Mikolov et al., 2013), term-frequency, inverse document-frequency (Tf-Idf) or both methods combined, using, or ignoring stop words. As the combination of both word2vec and Tf-Idf has been proven to be the most efficient in Lilleberg study, the current experiment also followed this solution.

The current study is in line with the previously presented analysis through the quantitative analysis performed on the DGLR and CVLR. Also, it expands on various previous studies that considered only part of DGLR and a smaller sample of the entire CVLR that are introduced in the follow-up sections. All these analyses are available to the public in the form of visual graphs, maps, and special representations, in an online web platform, free of charge.

3. Method

Our analysis is centered on providing several visualizations of experiments conducted on the DGLR and CVLR corpora. Data about Romanian writers, their writings, other literary entities, was extracted from DGLR, while literary events that happened after World War II were from CVLR. All information was provided by the researchers of the Romanian Academy under three datasets. The first dataset is represented by files in Adobe InDesign² format, covering the content from DGLR. The second dataset contains detailed information about the canonical writers (60 writers) and is in Microsoft Word format. The last dataset contains the content from CVLR, and it is also structured in Microsoft Word format. Before extracting the information from the three datasets, additional pre-processing was necessary. A first step was to manually inspect the provided corpus, then parse the provided files, and store relevant data in an accessible format.

3.1 Corpus

The DGLR dataset, provided in ready-to-print format (.INDD files), included 87 files, containing information about writers and literary entities from the Romanian literature. The provided information covers all available writers and is structured in alphabetical order, in correspondence to the published printed format. Currently, only the volumes covering letters A-P are available in the printed format. The rest of the information is expected to be distributed to the public in the last quarter of 2021. For the current analysis, letters Q-Z were provided as a non-revised version, therefore requiring additional text pre-processing. In the pre-processing steps several special entities were found, starting with non-alphabetic characters, such as “13 Martie”, “75 HP”, “1944”, or even punctuation (e.g., “?”). These entities were removed from the analysis as they were not referring to writers or literary events.

² <https://www.adobe.com/ro/products/indesign.html>

Together with the ready-to-print documents, the DGLR team provided a set of Microsoft Word documents with detailed information about 60 most important writers who contributed to the Romanian Literature, also called canonical writers. These documents include the writer's life chronology, his literary activity chronology, quotes from well-known literary critics, quotes from significant works and the titles of his representative writings. All this information is also displayed in the INTELLIT web platform.

The third used corpus is the CVLR dataset. It contains the literary events that happened between 1944-2000. The remaining timeframe is currently under development by the Romanian Academy research team. Documents containing CVLR dataset were provided in Microsoft Word format, and, for each year, one or more documents were associated. The current experiments cover the periods between 1949-1959, 1964-1967 and 1990-2000, summing a total of 24 years. Years 1957 and 1999 were excluded from the analysis as they were not published when current experiment has been performed.

All three corpora were stored in an Elasticsearch (Gormley and Tong, 2015) instance, which is fast for data retrieval on large amounts of texts and suitable for analytical purposes. Data from DGLR in Adobe InDesign format was exported to HTML, which is structured for easy parsing operations and keeps the special writing styles used by the researchers from the Romanian Academy: bold format to specify the current writers' name, or italic format to specify a work's title. For each Adobe InDesign file, two associated files were exported: HTML and its corresponding styling file, in CSS format. The HTML file and its associated CSS were merged, using an inline-CSS strategy to inject specific formatting options directly in the HTML elements; the resulted HTML file was indexed in Elasticsearch. Two categories of entities were identified from the HTML files, based on their specific class identifiers: writers (with two sub-categories: major and minor, stored in an Elasticsearch index called *index-authors*), and literary entities (publications, associations, literary movements, stored in *index-publications*). Several fields were extracted for each type of entity to be used for future analyses. The name, year of birth, year of death, birthplace, place of death, a list of professions, description from DGLR, publishing years, list of publications, and critical quotes were extracted for authors. Only the name and description were available for literary entities. The total number of entities exported and parsed in the current experiment included 4188 writers and 2099 literary entities.

The canonical writers' additional documents were exported from the Microsoft Word format (.docx) to a HTML format, similarly to the previous pre-processing. The generated HTML files already included the styling from the initial Word files. Data extracted from canonical writers' documents was indexed in the *index-authors* node, with new sub-fields for the specific writers: *life_chronology*, *literary activity_chronology*, *representative_writings*, *literary work_quotes*.

Data from the CVLR corpus was exported from the Microsoft Word format (.docx) to a simple text format (.txt), without keeping the special formatting (bold and italic) used in the initial text. Each year contains a list of literary events, chronologically ordered and, for each event, two fields are stored: event date and description. The CVLR dataset was stored in the *index-events* index, and each event is represented as a separate document in the Elasticsearch database. The CVLR corpus was used in a topic modelling experiment presented in the following sections.

3.2 Experiments

Several analyses have been performed based on the information presented in the INTELLIT corpus, both from a general quantitative perspective of case studies and visualizations (Neagu et al., 2020b), but also from a computational, NLP perspective (Neagu et al., 2019a, Neagu et al., 2019b). Considering the general quantitative analysis, the following experiments are of interest:

- The evolution of writers' related data through time – the collection of experiments contains graphical representations of the number of writers born each year, the birth location of writers, their death age (if available), the number of publications and active writers per year;
- The literary entities through time – the collection of experiments contains graphical representations of the number of publications per year and the number of active publications per year;
- Canonical writers – the collection of experiments contains graphical representations of the writers' active publication period, the cities visited by the writers and a publication timeline.

Two NLP-driven experiments were conducted: the first one targeting DGLR (Neagu et al., 2020a), which now presents the results using full dictionary corpora (initial work included only letters A-D), and a second one targeting CVLR (Neagu et al., 2019b), where the results are reinterpreted.

The first experiments aim to generate interactive three-dimensional visualizations with all Romanian writers, literary entities, and their corresponding distances, based on specific features extracted using data stored in the Elasticsearch server: similar text descriptions in DGLR in terms of semantic models, timeframes, professions, and biographic references. For better visualization, clustering methods, such as K-Nearest Neighbors, were applied for grouping similar writers and entities together. The DGLR writers' descriptions were concatenated with the writers' quotes and Tf-Idf has been applied to the resulted data, using the scikit-learn implementation in Python³. Spacy⁴ was used for all the language specific tasks, such as tokenization, lemmatization and stop-words removal. Afterwards, the generated Tf-Idf matrix was used to weight each word embedding given by fastText (Athiwaratkun et al., 2018), pre-trained with 300 dimensions, averaged across each DGLR description and quotes. A word embedding is a learned representation for text, where words that have the same meaning have a similar representation. The next step consists of reducing the space dimensionality, from 300 to 17 dimensions through Principal Component Analysis (PCA). A second PCA was applied on the remaining dimensions obtaining a 3 dimensions space, which was afterwards plotted using TensorBoard toolkit⁵. To obtain a more accurate distance function between writers, several other features were integrated:

- *Average publishing year* – for each writer, it was computed an average value of the publishing years, that were afterwards scaled using a min-max normalization. The distance function is defined as the absolute difference between the calculated values;
- *Writers' professions* – were extracted from Elasticsearch and further processed to reduce the number of similar professions. The distance function calculated for this feature is represented by the Jaccard index (Leydesdorff, 2008) applied on pairs of writers;
- *Critical quotes* – for each writer, the list of people who wrote critical quotes was extracted and afterwards the lists were compared. This distance function is also based on the Jaccard index.

³ https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

⁴ <https://spacy.io/>

⁵ <https://www.tensorflow.org/tensorboard>

The resulting distance function was computed by summing the distance functions from the embeddings with the features' distance functions. Next, a distance matrix was computed for the writers in the corpus based on the new distance function. After this step, the unsupervised neighbors' implementation from scikit-learn in Python⁶ was used for determining the writers nearest neighbors.

The second NLP experiment used the CVLR corpus to capture the topics' evolution across time and to model how specific historical timeframes, after World War II, determined the main discussion topics in Romania. A text pre-processing step was also implemented, which involved the removal of several words from the text: non-alphabetic characters, words with less than 3 characters, words with high frequency (appearing in more than 80% of the documents) or stops words. Afterwards, the named entities were identified and grouped together, trained on the RONEC corpus⁷ with a standard spaCy model for Named Entity Recognition. At the end of this phase, only content words were kept, and used for topic modeling. Latent Dirichlet Analysis (Blei et al., 2003) was used for topic modeling, and multiple models were trained to find the optimal coherence score – sum of the pair wise scores on the underlying words used to describe a topic. The best coherence score for each timeframe was obtained using 3 topics, but they were too broad. The optimal number of topics was set to 10 from the multiple tested LDA instances, while the differences for 20 and 30 topics were quite low.

4. Results

The *ReaderBench* website hosts several experiments grounded in Natural Language Processing, such as text mining and essay writing feedback (Dascalu et al., 2014, Gutu-Robu et al., 2018). The website is developed using novel technologies, such as Angular⁸ for the front-end side and D3.js⁹ and AmCharts¹⁰ for graphical representations. The main advantage of using AmCharts is its ease of use and its wide range of predefined chart types. Also, the framework allows developers to create

⁶ <https://scikit-learn.org/stable/modules/neighbors.html>

⁷ <https://github.com/dumitrescustefan/ronec>

⁸ <https://angular.io/>

⁹ <https://d3js.org/>

¹⁰ <https://www.amcharts.com/>

their own chart shapes based on Scalar Vector Graphic¹¹ (SVG) elements. The quantitative experiments conducted in the current study integrate the visualization in the *ReaderBench* website under the Experiments section. The path to the presented use cases is available in the INTELLIT Experiments Menu¹².

The graphical elements are divided into the following categories based on the type of visualization: bar-charts (display information across a period of time), range-area charts (emphasize minimum, average and maximum values), dumbbell charts (highlight the differences between groups), maps (display geographical data), and timelines (display multi-dimensional features over time). Each visualization is structured as title, description, and chart representation. Beside timelines and maps, all other graph categories contain the following elements: scrollbar for zooming in and out, cursor for identifying the exact value on the chart, tooltip for displaying the value and legend for identifying the chart series.

Even though bar-charts are the most common type of representing information, they summarize large complex data and allow users to identify trends. For smoothing out the spikes in the graphs and for better trend identification we implemented a 5-point moving average for each bar-chart. A bar-chart visualization in the INTELLIT platform is the number of active journals over time, shown in Figure 1. A journal is considered active in the period between the first and the last published volume. The current visualization displays the number of active journals over a period of 180 years, from 1830 to 2010. Data before 1830 was also available in the corpus, but the number of active journals was insignificant.

Several trends can be identified in the visualization. The first half of the 19th century represents the starting point of the Romanian journals. Until 1850, the number of active journals is small, having a maximum peak at 10 in 1845 and 1848. This period included the debut of the first journal written in Romanian, *Curierul românesc* (1829). The second part of the 19th century is more interesting, as we can identify an ascending trend in the number of active publications. This period is also marked by the autochthone literary trend promoted by *Junimea* group which contributed substantially to the growth of the Romanian culture. The impact of the First World War is not significant in the number of active publications, most of them continuing

¹¹ <https://www.w3.org/Graphics/SVG/About.html>

¹² <http://readerbench.com/experiments/intellit>

their activity. In the interwar period, the number of active publications reaches new peaks, 249 in 1935 and 247 in 1938, dropping significantly after the Second World War. During the communist period, the publication trend is constant, having around 50–60 active publications per year, reflecting the total control of the printed press by the Communist Party. After the communist period, the number of publications continues its ascending trend; unfortunately, data for this period is scarce. The map visualization confirms the direct relationship between a particular political regime and the cultural life of a nation. Thus, the sudden drop in the number of publications at the beginning of the communist era has no parallel in other periods of time, not even in the grim years of the two World Wars.

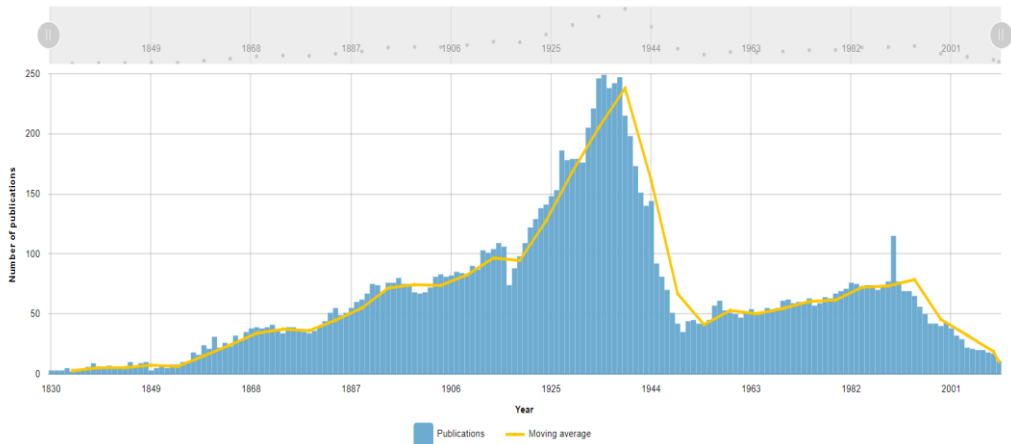


Figure 1. Number of active publications from 1830 to 2010.

The map visualizations contain geographical information, such as the birth locations of writers or the travels of the canonical writers. The birth locations map was firstly introduced in the study performed by (Neagu et al., 2020b) and was evaluated using a questionnaire in a second study (Toma et al., 2020). Users considered the representation cluttered by displaying the number of authors born in each county; thus, the representation was transformed into a heat map (see Figure 2). Higher values are represented in darker tones of blue, while lower values are represented in lighter shades of blue. The legend of the map is displayed in the lower part of the visualization.

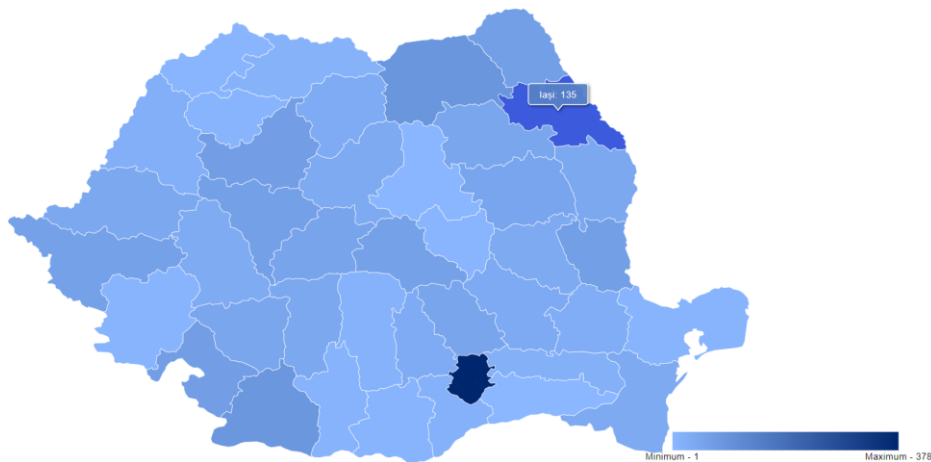


Figure 2. Heatmap with writers' birthplaces.

Timelines are an experimental view, a serpentine that displays the volumes published by a writer (for example, George Coșbuc) across his lifetime (see Figure 3). The timeline is divided into sections colored differently, representing the places where the writer traveled. The publications are overlapped on the same timeline, giving the audience multi-dimensional information: the title, the publishing date of a work and the place the writer was at that time.

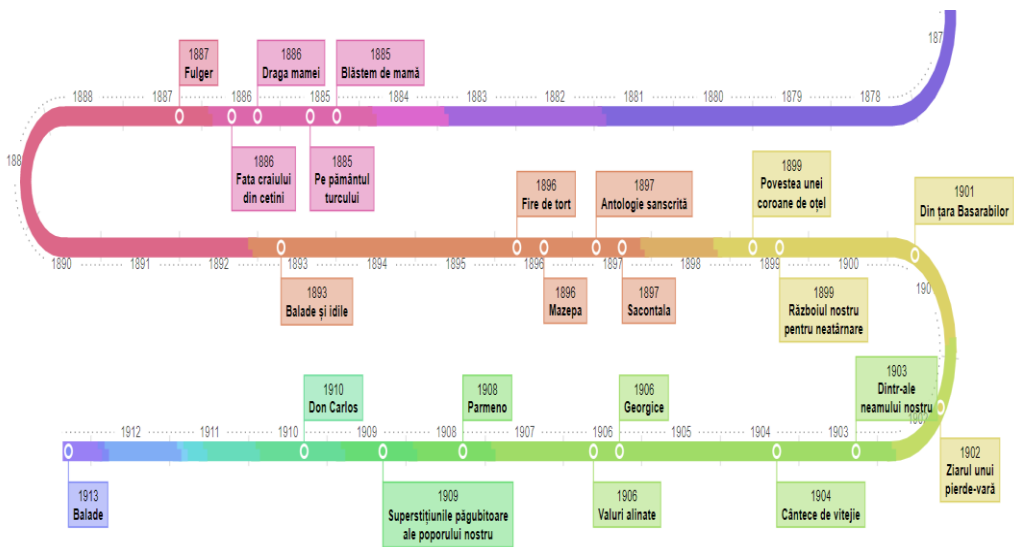


Figure 3. Serpentine timeline for George Coșbuc

4.1 Semantic Distances between Writers

The integration of the NLP experiments in the ReaderBench website is not an easy task, as the three-dimensional rendering of the writers in space based on word embeddings requires custom GPU to run. The text for each writer used to generate the word embeddings was composed of DGLR description and critical quotes. For visualization, the initial representation space has been reduced to three dimensions using PCA and, to plot the results, TensorFlow Projector¹³ was used.

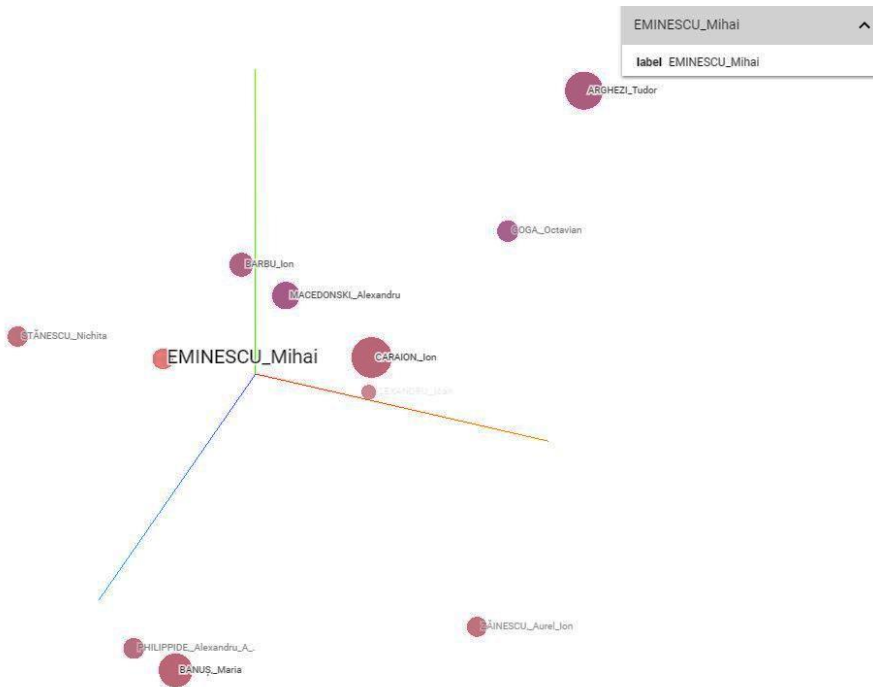


Figure 4. Mihai Eminescu's most similar writers visualization

Figure 4 shows Mihai Eminescu's most similar writers based on the word embeddings. The other writers, more than 4000, were not displayed in this visualization not to alter and interfere with the top 10 neighbors shown. The TensorFlow Project platform is interactive and has the option to search for any writer and filter out based on custom number of neighbors. The inverse cosine similarities or top distances from Mihai Eminescu are shown in Table 1. Afterwards, the results of the visualization and the top distances were manually explored and, most of them, are consistent to the

¹³ <https://projector.tensorflow.org/>

general literary knowledge: some of the writers lived in the same period as Mihai Eminescu and some of them were influenced or inspired by Eminescu in their writings. Another example, the top 5 closest writers for RéthyAndor, a Hungarian writer, based on the provided descriptions in DGLR are also Hungarians, namely: Kakassy Endre (.054), Pálffy Endre (.125), Domokos Sámuel (.128), Faragó József (.241) and Kiss Jenő (.293).

Table 1. Mihai Eminescu's most similar writers distances

Author name	Distance based on description feature
Octavian Goga	.149
Alexandru Macedonski	.162
Ion Barbu	.197
Tudor Arghezi	.202
Philippide Alexandru	.236
Ion Caraion	.248
Nichita Stănescu	.255
Adrian Păunescu	.282

Next, the remaining features (i.e., average publishing year, writers' professions and critical quotes) were weighted and integrated in the final distance, and the nearest neighbors were computed, using the new distances. A sub-list of 9 canonical writers was chosen and the distances between writers are shown in Table 2 using all the similarity features. The closest writers are marked in the table with bold, and, based on the list, Mihai Eminescu is very close to Octavian Goga (.121), Tudor Arghezi (.155) and Nichita Stănescu (.188), which looks consistent to the general literary knowledge. Thus, the greater distance between Eminescu and Vasile Alecsandri (the latter living in the same period of time and being considered, at Eminescu's debut, the most important Romanian poet), compared to the distance between Eminescu and Octavian Goga or Eminescu and Tudor Arghezi (two poets who lived later on) certifies the crucial importance of Eminescu's work for the development of the Romanian literature and culture. In contrast, the distance between Eminescu and other not directly related writers is considerably larger, for example: Adela Xenopol (.675), Fodor Sándor (.740), or Radu Beligan (.760).

Table 2. Canonical writers' distances based on all similarity features

	George Bacovia	Ion Creangă	Marin Sorescu	Mihai Eminescu	Mihail Sadoveanu	Nichita Stănescu	Octavia Goga	Tudor Arghezi	Vasile Alecsandri
George Bacovia	-	.330	.332	.229	.370	.179	.231	.201	.371
Ion Creangă		-	.333	.233	.248	.215	.213	.205	.313
Marin Sorescu			-	.233	.378	.206	.285	.227	.352
Mihai Eminescu				-	.253	.188	.121	.155	.303
Mihail Sadoveanu					-	.314	.237	.290	.261
Nichita Stănescu						-	.202	.197	.357
Octavian Goga							-	.138	.243
Tudor Arghezi								-	.322
Vasile Alecsandri									-

4.2 Modeling of Romanian Literary Trends in History

The initial results of this experiment were presented by (Neagu et al., 2019b) and it was shown that the important topics in the beginning of the 1950s was socialism which, later on, drops significantly starting 1960. After the communism fall, topics of interest were dissidence, politics, and writers in exile. (Neagu et al., 2020a) used the CVLR corpus to perform an analysis of trends over time analysis using 4 narrow periods with historical importance for Romania: 1949-1959, 1964-1967, 1990-1995, 1996-2000. It was discovered that some of the topics were present in all the timeframes, such as: Mass Media / News topic, also Nationalism topic was present in the early communism and in the early post-communism, and some topics were isolated: Political Elections topic present only in the early post-communism, and Literary Censorship topic which was discovered only in the 1996-2000 timeframe. Some of the topics evolved through time, and the experiment showed that the literature in the early communism was strongly influenced by the Communist Party; then, after 1990, writers were much more involved in the public life and in the literary and social debates, defining new directions for literature, talking about the past censorship, exile and writing without constraints.

5. Conclusions

Details on all the writers who contributed to the Romanian literature (DGLR project) and on all the literary events that happened nationwide after

World War II (CVLR project) were digitized through the development of the INTELLIT web platform. The current work presents several analyses based on DGLR and CVLR to better understand the evolution of Romanian literature throughout time. Most analyses are freely available online to the public on the *ReaderBench* website, within the Experiments section.

Future work includes rerunning the current analysis on the full CVLR corpus. The trends over time analysis may show other topics and evolution of topics, referring to the entire period of the communist regime. Another interesting future project consists of linking writers from DGLR with literary events from the Chronology and performing analyses across both corpora.

Acknowledgements

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS – UEFISCDI, project number PN-III 54PCCDI / 2018, INTELLIT – “Prezervarea și valorificarea patrimoniului literar românesc folosind soluții digitale inteligente pentru extragerea și sistematizarea de cunoștințe” and by the Operational Programme Human Capital of the Ministry of European Funds through the Financial Agreement 51675/09.07.2019, SMIS code 125125.

References

- Algee-Hewitt, M. 2021. Stanford Literary Lab [Online]. Available: <https://litlab.stanford.edu/pamphlets/> [Accessed June 1st 2021].
- Allison, S., Heuser, R., Jockers, M., Moretti, F. & Witmore, M. 2011. Quantitative formalism: an experiment, Universitätsbibliothek Johann Christian Senckenberg.
- Athiwaratkun, B., A., W. & Anandkumar, A. 2018. Probabilistic fasttext for multi-sense word embeddings. arXiv preprint arXiv:1806.02901.
- Blei, D. M., Ng, A. Y. & Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S. & Nardy, A. 2014. Mining texts, learner productions and strategies with ReaderBench. In: PEÑA-AYALA, A. (ed.) *Educational Data Mining: Applications and Trends*. Cham, Switzerland: Springer.
- Dumais, S. T. 2004. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38, 188–230.
- Eder, M. 2017. Visualization in stylometry: cluster analysis using networks. *Digital Scholarship in the Humanities*, 32, 50–64.
- Gormley, C. & Tong, Z. 2015. *Elasticsearch: The definitive guide: A distributed real-time search and analytics engine*, O'Reilly Media, Inc.
- Gutu-Robu, G., Sirbu, M.-D., Paraschiv, I. C., Dascalu, M., Dessus, P. & Trausan-Matu, S. 2018. Liftoff - ReaderBench introduces new online functionalities. *Romanian Journal of Human - Computer Interaction*, 11, 76–91.

- Lansdall-Welfare, T., Sudhahar, S., Thompson, J., Lewis, J., Team, F. N. & Cristianini, N. 2017. Content analysis of 150 years of British periodicals. *Proceedings of the National Academy of Sciences*, 114, E457-E465.
- Lee, M. D., Pincombe, B. & Welsh, M. 2005. An Empirical Evaluation of Models of Text Document Similarity. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 27.
- Leydesdorff, L. 2008. On the normalization and visualization of author co-citation data: Salton's cosine versus the Jaccard index. *Journal of the American Society for Information Science and Technology*, 59, 77-85.
- Lilleberg, J., Zhu, Y. & Zhang, Y. Support vector machines and word2vec for text classification with semantic features. 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015. IEEE, 136-140.
- Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representation in Vector Space. Workshop at ICLR, 2013 Scottsdale, AZ.
- Moretti, F. 2000. Conjectures on world literature. *New left review*, 1, 54.
- Moretti, F. 2005. *Graphs, maps, trees: abstract models for a literary history*, Verso.
- Moretti, F. 2016. *Literature, Measured* [Online]. Available: <https://litlab.stanford.edu/LiteraryLabPamphlet12.pdf> [Accessed June 1st 2021].
- Neagu, L.-M., Cotet, T.-M., Dascalu, M., Trausan-Matu, S., Badescu, L. & Simion, E. Semantic Author Recommendations based on their Biography from the General Romanian Dictionary of Literature. 7th Int. Workshop on Semantic and Collaborative Technologies for the Web, in conjunction with the 15th Int. Conf. on eLearning and Software for Education (eLSE 2019), 2019a Bucharest, Romania. "CAROL I" National Defence University Publishing House, 165-172.
- Neagu, L.-M., Cotet, T.-M., Dascalu, M., Trausan-Matu, S., Chisu, L. & Simion, E. Semantic Recommendations and Topic Modeling based on the Chronology of Romanian Literary Life. 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), 2019b Magdeburg, Germany. Springer, 164-174.
- Neagu, L.-M., Dascalu, M., Trausan-Matu, S., Chisu, L. & Simion, E. Automated Modeling of Romanian Literary Trends in History using Topics over Time and Co-Occurrences. 8th Int. Workshop on Semantic and Collaborative Technologies for the Web, in conjunction with the 16th Int. Conf. on eLearning and Software for Education (eLSE 2020), 2020a Online. "CAROL I" National Defence University Publishing House, 151-158.
- Neagu, L.-M., Toma, I., Dascalu, M., Trausan-Matu, S., Hanganu, L. & Simion, E. A Quantitative Analysis of Romanian Writers' Demography Based on the General Dictionary of Romanian Literature. 5th Int. Conf. on Smart Learning Ecosystems and Regional Development (SLERD 2020), 2020b Bucharest, Romania (Online). Springer, 253-261.
- Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M. & Dodds, P. S. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5, 1-12.
- Simion, E. (ed.) 2004-2009. *Dicționarul General al Literaturii Române*, Bucharest, Romania: Editura Univers Enciclopedic.
- Simion, E. (ed.) 2016-2018. *Dicționarul General al Literaturii Române*, Bucharest, Romania: Editura Univers Enciclopedic.

- Simion, E. & Chișu, L. (eds.) 2017-2020. *Cronologia vieții literare românești. Perioada postbelică (1964-1979)*, Bucharest, Romania: Editura Muzeul Literaturii Române.
- Simion, E. & Grigor, A. (eds.) 2010-2012. *Cronologia vieții literare românești. Perioada postbelică (1944-1964)*, Bucharest, Romania: Editura Muzeul Literaturii Române.
- Sinykin, D., So, R. J. & Young, J. 2019. Economics, race, and the postwar US novel: a quantitative literary history. *American Literary History*, 31, 775-804.
- Statista. 2019. Have you accessed or downloaded e-books or digital books free of charge or by paying for them? [Online]. Available: <https://www.statista.com/statistics/1175646/romania-downloading-digital-books-by-type-of-registration/> [Accessed June 1st 2021].
- Statista. 2020a. eBooks in Europe [Online]. Available: <https://www.statista.com/outlook/213/102/ebooks/europe> [Accessed June 1st 2021].
- Statista. 2020b. Percentage of individuals purchasing books, magazines and e-learning materials online in Romania from 2016 to 2019 [Online]. Available: <https://www.statista.com/statistics/1105231/internet-purchases-of-reading-materials-romania/> [Accessed June 1st 2021].
- Thelwall, M. 2017. Book genre and author gender: romance> paranormal-romance to autobiography> memoir. *Journal of the Association for Information Science and Technology*, 68, 1212-1223.
- Toma, I., Neagu, L.-M., Dascalu, M., Trausan-Matu, S., Hanganu, L. & Simion, E. Emerging Patterns in Romanian Literature and Interactive Visualizations based on the General Dictionary of Romanian Literature. *International Conference on Human-Computer Interaction (RoCHI2020)*, 2020 Sibiu, Romania (Online). *MatrixRom*, 91-103.

Advanced Natural Language Processing Techniques for Restoring Old Romanian Documents

Silvia Tomescu¹, Irina Mitocaru², Gabriel Guțu-Robu², Melania Nițu²,
Mihai Dascălu², Ștefan Trăușan-Matu²

(¹ Carol I Central University Library Bucharest)

(² University Politehnica of Bucharest)

Modern society evolves and it is transformed at a rapid pace, whereas the cultural heritage of a nation risks to become less accessible if it is not considered for the constant technology updates. The Lib2Life computerised platform is aimed at enabling four central university libraries in Romania to preserve historically valuable documents, while also providing to the public free online access to documents that are no longer copyright protected. Lib2Life is powered by advanced Natural Language Processing techniques that deal with texts from old documents, index them for immediate retrieval, and provide recommendations of similar documents, based on semantic relatedness. Lib2Life considers three user roles, namely: librarians, readers, and administrators. Librarians can upload documents previously processed using Optical Character Recognition software. Text is automatically extracted, followed by corrections and transformations to reconstruct a coherent digitised representation of the document to be rendered on the web. Using an embedded text editor, librarians can check and edit the extracted information, including tables, images, footnotes or even the generated table of contents based on section headings. Metadata, such as authors, publishing year, or domain can also be edited. Besides inspecting the content of each document, readers can perform advanced semantic searches across a database consisting of textual paragraphs extracted from the digitised documents. Once a document is selected, readers can view a list of similar documents recommended to them. Administrators have full access to the platform, managing users and corresponding roles.

1. Introduction

World cultural heritage has been a long-term concern of many international entities, such as the United Nations Educational, Scientific and Cultural Organization (UNESCO; Labadi, 2013). In 1972, the Convention for the Protection of the World Cultural and Natural Heritage opened a new chapter for heritage preservation (Room, 1972). Additionally, the European Union's (EU) Agenda on Culture, European Framework for Action on Cultural Heritage, together with the Europeana Platform (<https://www.europeana.eu>), defined cohesive strategies, guidelines, and mechanisms for heritage protection throughout the European Union starting from 2018. The need to share a country's history and national identity in order to close the culture gap between countries, alongside with a need to retain material and immaterial heritage beyond borders, emerged in the declaration of cooperation on advancing the digitisation of cultural heritage, signed in 2019 by 29 EU member states. Its content focuses mainly on: a) a pan-European initiative for the 3D digitisation of cultural heritage artefacts, monuments, and sites; b) reusing digitised cultural resources to foster citizen engagement, innovative use, and spill-overs into other sectors; and c) enhancing cross-sector and cross-border cooperation and capacity building in the sector of digitised cultural heritage. Digitisation and preservation require a complex undertaking, involving legislation, institutional cooperation, national and international frameworks, legislative specification, and strategic objectives.

The initiative to provide open access to documentary heritage at a global level was initiated by Michael Hart in 1971 with the start of Project Gutenberg (Lebert, 2008). The project grew with the development of the web in 1990 and a drastic increase was observed after 2000 when more than ten thousand volumes were made publicly available (Stroube, 2003, Lebert, 2008). Considered a long-term solution for heritage conservation, digitisation has many advantages such as: resilience of the medium, its diverse capacity to organize information, accessibility for both research and visualization, creating new opportunities for people with disabilities, all of these benefits raising awareness and preserving national identity through effective dissemination of knowledge. With all the socio-political, economical, and technological considerations, taking into account existing gaps, promotion, and conservation through technologies, digitisation remains a dynamic process, involving leading bodies and research

institutions, as well as the transfer of expertise in the social and economic fields. Recommendations on digitisation issued by the European Commission are gathered in a joint effort of both private and public sectors to boost online access to cultural heritage, while simultaneously preserving it through modern technologies (European Commission, 2011).

Promoting the national documentary heritage through advanced Natural Language Processing (NLP) techniques, the Lib2Life computerised platform approaches the digitisation from the perspective of material heritage by providing open access to a digital deposit of old documents that abide to copyright restrictions, together with semantic search facilities. The aim of this chapter is to highlight the key functionalities of Lib2Life that provides access to a digitised and structured collection of documents, together with semantic search services. The platform is valuable not only in information and documentation practice, but especially for research and accessing historical contents of old books. Lib2Life reflects the orientation towards the principle of open access to science and learning, through the use of open-source semantic tools and libraries, together with knowledge representation techniques.

In terms of structure, this chapter continues with the overview of the Lib2Life platform, details on the ontology used to represent knowledge and structure information, followed by key functionalities, conclusions and future work.

2. Overview of the Lib2Life Platform

The Lib2Life platform (Mitocaru et al., 2020) aims to publish collections of four central university libraries from Romania, to create a shared catalogue of digital document collections, and to support advanced search facilities, together with semantic reading recommendations. This joint effort has additional benefits, including the increase of digitisation skills for librarians, the improvement of skills for ontology design and for structuring information, thus enabling the exchange of good practices. Document processing is achieved through NLP-centred text extraction and indexing mechanisms for immediate retrieval. The most eloquent example of a similar service is the Europeana Platform that aggregates a diverse typology of documents belonging both to libraries and museums, but also news and events (Haslhofer and Isaac, 2011).

The input data consists of PDF documents that have undergone Optical Character Recognition (OCR) processing to extract the text from the old books. An XML file with extracted metadata is attached to the OCR-ed version of the documents based on the information held by libraries in their internal document management systems, namely Vubis (Alewaeters, 1982) in the case of the Carol I Central University Library of Bucharest, and Aleph (Julich et al., 2003) held by the Eugen Todoran Central University Library of Timișoara, the Mihai Eminescu Central University Library of Iași, and the Lucian Blaga Central University Library of Cluj. This information is automatically extracted when uploading a document into the Lib2Life platform; the considered metadata fields include elements such as author names, the year of publication, document language, domain, or a short description. The XML documents were integrated with the scanned file as an additional JSON-encoded (JavaScript Object Notation) metadata attribute.

After performing advanced automated corrections on the extracted texts, alongside with manual adjustments performed by librarians within the dedicated web page, the final curated texts, divided into paragraphs, are then indexed into Elasticsearch, a distributed full-text search engine that provides scalable and real-time search functionalities (Gormley and Tong, 2015). This approach supports the accuracy of the semantic search functionality and the relevance of document recommendations, as a consequence of our target to focus on paragraphs or sections of texts rather than books as a whole. Besides these texts, the Elasticsearch database also integrates data required for the Lib2Life web application, such as user personal information (name and surname), information used for authentication (e-mail, password), and user roles.

Digital services provided by Lib2Life include document pre-processing techniques such as identification of paragraphs, reconstructing words divided into syllables, as well as algorithms to automatically extract text, tables, and images. Lib2Life also incorporates advanced document processing techniques based on semantic relatedness at paragraph level. The Lib2Life platform includes three types of users: readers, librarians, and administrators. Administrators manage account types, assign roles and provide access to the application, while librarians are responsible for uploading, verification of extracted texts, and correction of texts so as to be coherently saved into Elasticsearch. Due to their specific and focused role, administrator functionalities will not be further discussed in this chapter.

The reader refers to a generic user, who can explore the uploaded documents, access the extracted texts, or view the original PDF documents. The reader is also able to explore advanced services, such as retrieving recommended documents, together with search and filtering functionalities.

3. The Lib2Life Ontology

Libraries require instruments that reflect their document collection in a standardised manner to improve information retrieval. Document contents are reflected by logically structured classification schemes, such as thesauri, taxonomies, or ontologies, the latter being formalised knowledge representation models (Gruber, 1993) that can be easily queried (Banu et al., 2013).

The four central university libraries of Romania developed a shared digital repository to provide open access to a consolidated documentary heritage; based on this collection, a dedicated ontology was developed, taking into account a knowledge area structure that includes seventeen domains (Gutu-Robu et al., 2020). The Lib2Life ontology incorporates document attributes and specific features that support data extraction through advanced queries that are exemplified afterwards. The built-in VOWL (Visual Notation for OWL Ontologies) ontology viewer allows users to navigate through the defined classes and subclasses, together with corresponding relationships (Lohmann et al., 2016). Figure 1 shows the integrated ontology viewer zoomed in to display the Biology knowledge domain and its subdomains, as it can be accessed through the Lib2Life platform.

3.1 Design Principles and Methodology

Ontologies are described in the standard Web Ontology Language (OWL), built on top of the Resource Description Framework (RDF), which incorporates specifications for web resources and their metadata modelling, using syntax notations and data serialization formats (Lassila and Swick, 1998). Data are stored in RDF as triplets (subject, predicate, and object) which helps create RDF graphs. In an RDF triplet, the predicate can be interpreted as the relationship between the subject, which is represented by a node in the graph, and the object, which can be either another node, or a specific value. OWL enables ontology-specific definitions, such as classes and subclasses, domains and range of relationships, and also provides support for inference rules (McGuinness and Van Harmelen, 2004).

The Lib2Life ontology was developed using Methontology (Fernández-López et al., 1997), a methodology that provides instructions for the entire cycle of ontology development and use, such as specification, conceptualization, formalization, integration, implementation, and maintenance. The Lib2Life ontology (see Figure 1) was developed by reusing well-known ontologies such the Dublin Core (DC; Weibel et al., 1998) and Friend-of-a-Friend (FOAF; Golbeck and Rothstein, 2008). The DC metadata standard consists of a collection of tags created to describe digital resources. DC is widely used by librarians, archivists, scientists, and software developers, and was accepted as a standard for annotating digital document metadata. The Lib2Life ontology is aligned with this standard, as it uses appropriate DC tags for the classes and properties needed to correctly describe the physical documents of the central university libraries. FOAF is an ontology that can be used to describe people and organizations, their activity, and relationships with other people and objects. Most of its properties and attributes are related to social networks, which were not applicable to the Lib2Life ontology. However, general classes such as “Person” or “Organisation” were used to define authors and publishers.

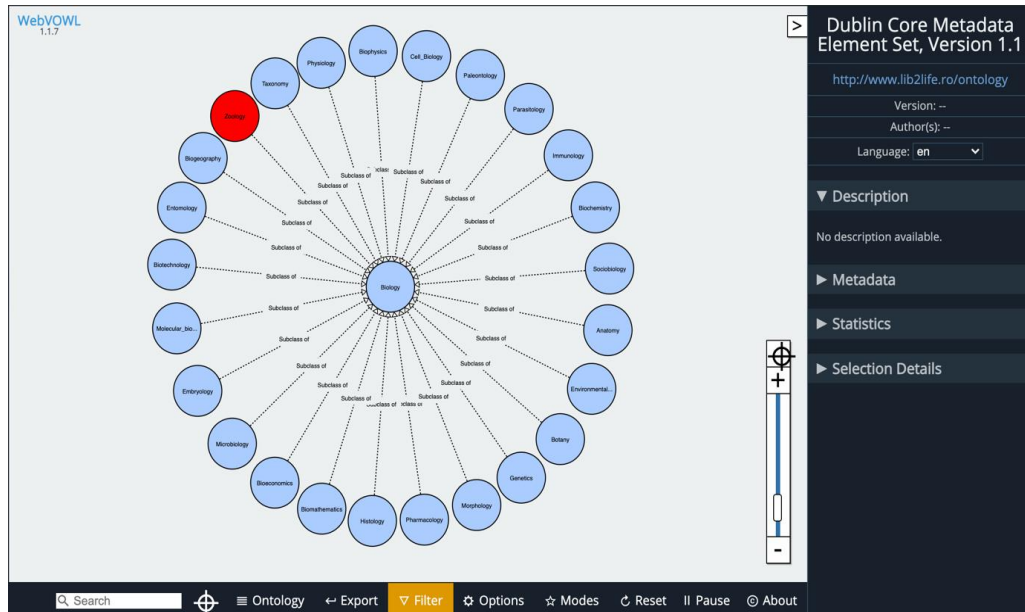


Figure 1. The incorporated Ontology Viewer showing the Biology knowledge domain and its subdomains, as seen in the Lib2Life platform.

3.2 Querying the Ontology

Ontology query languages, such as the SPARQL applied over the RDF graph (W3C, 2013), can be used to extract connected information and make inferences. The Lib2Life platform relies on Apache Jena Fuseki 2 framework (Apache Foundation, 2019) for storing data and executing queries. For exemplification, in the following paragraphs we describe three queries applied on the Lib2Life ontology to extract meaningful data.

The first query discovers all the roles held by an individual. For example, if the Romanian writer named Ion Heliade Radulescu was selected, and the SPARQL query from Figure 2.a is introduced, Figure 2.b shows all the roles held by Ion Heliade Radulescu as they are retrieved by the incorporated Query view of the Protégé ontology editor (<https://protege.stanford.edu>) (Noy et al., 2003).

```
SELECT ?subject ?object
WHERE {
  ?subject rdf:type ?object
  . FILTER regex(str(?subject), "Heliade_Radulescu")
  . FILTER (!regex(str(?object), "owl"))
}
ORDER BY ASC(?subject)
```

a)

subject	object
Heliade_Radulescu_Ion	Philosopher
Heliade_Radulescu_Ion	Philologist
Heliade_Radulescu_Ion	Person
Heliade_Radulescu_Ion	Translator
Heliade_Radulescu_Ion	Writer
Heliade_Radulescu_Ion	Historian

b)

Figure 2. A SPARQL query to find all the roles held by Ion Heliade Radulescu.

The second query is displayed in Figure 3.a and returns how many documents (for example, books) were written by each author from the Lib2Life ontology. As it can be observed in Figure 3.b, our ontology consists not only of Romanian authors, but also of French, German, and other nationalities. At the time of writing this chapter, the development of the ontology is in progress, only several authors and books were included and the author with most added documents is the French writer Anatole France.

```
SELECT ?object (COUNT(*) AS ?c)
WHERE {
  ?subject l2l:writtenBy ?object
}
GROUP BY ?object
ORDER BY DESC(?c)
```

a)

object	c
France_Anatole	"25"^^< http://www.w3.org/2001/XMLSchema#integer >
Javary_Adrien	"3"^^< http://www.w3.org/2001/XMLSchema#integer >
Heliade_Radulescu_Ion	"3"^^< http://www.w3.org/2001/XMLSchema#integer >
Antipa_Grigore	"2"^^< http://www.w3.org/2001/XMLSchema#integer >
Piaget_Jean	"2"^^< http://www.w3.org/2001/XMLSchema#integer >
Alecsandri_Vasile	"2"^^< http://www.w3.org/2001/XMLSchema#integer >
Anghel_Dimitrie	"1"^^< http://www.w3.org/2001/XMLSchema#integer >
Carp_Petre_P	"1"^^< http://www.w3.org/2001/XMLSchema#integer >
Diodorus_Siculus	"1"^^< http://www.w3.org/2001/XMLSchema#integer >
Djuvara_Trandafir_G	"1"^^< http://www.w3.org/2001/XMLSchema#integer >
Kanitz_Felix	"1"^^< http://www.w3.org/2001/XMLSchema#integer >

b)

Figure 3. A SPARQL query to find the number of documents written by each writer.

A third query (see Figure 4.a) extracts all the documents from our ontology, together with their publishing year, in descending order by year. Figure 4.b shows the results, highlighting that our ontology contains both contemporary volumes, and archaic documents as well. It can also be deducted from the title of each displayed document that part of them are written in Romanian, while others are written in English.

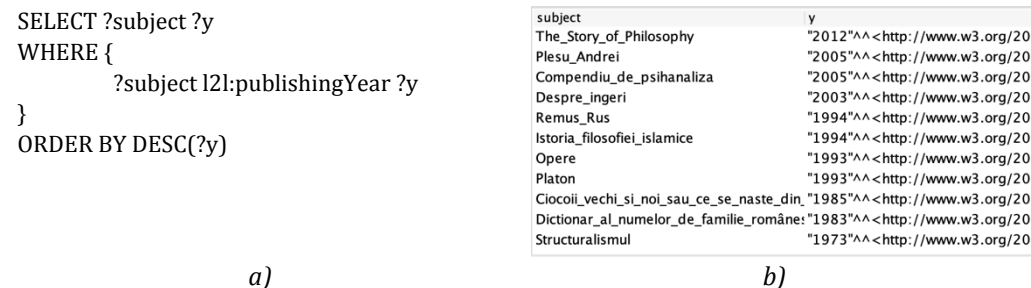


Figure 4. A SPARQL query to display documents existent in the ontology in descending order by their publishing year.

4. The Lib2Life Platform Main Functionalities

This section highlights the workflow of integrating digitized documents into the Lib2Life platform, including uploading, processing, and indexing mechanisms (Nitu et al., 2019). In regard to the digitized documents which serve as input for the platform, several discrepancies between the text contained within the OCR-ed PDFs and the original texts as they appear in the physical documents were reported, such as different font types and various sizes existing in the same section, different headers and footers, paragraph breaks, page breaks, and possible loss of document structure. However, the current collection of imported PDF documents consists of books that were properly processed by the Lib2Life pipeline, described later in detail. Figure 5 introduces the dashboard. The latest added documents are displayed in the middle of the page, together with their metadata, a list of editing options provided for librarians, and a list of recommended documents. On the right side, search fields and filtering options are included. The left side of each page ensures rapid navigation to other services provided by the platform. For example, the Statistics menu displays aggregated information of existing documents in the system. This also enables an analysis of the prevalence of specific publication periods of the catalogued works, while also reflecting the structure of the digitised collection.

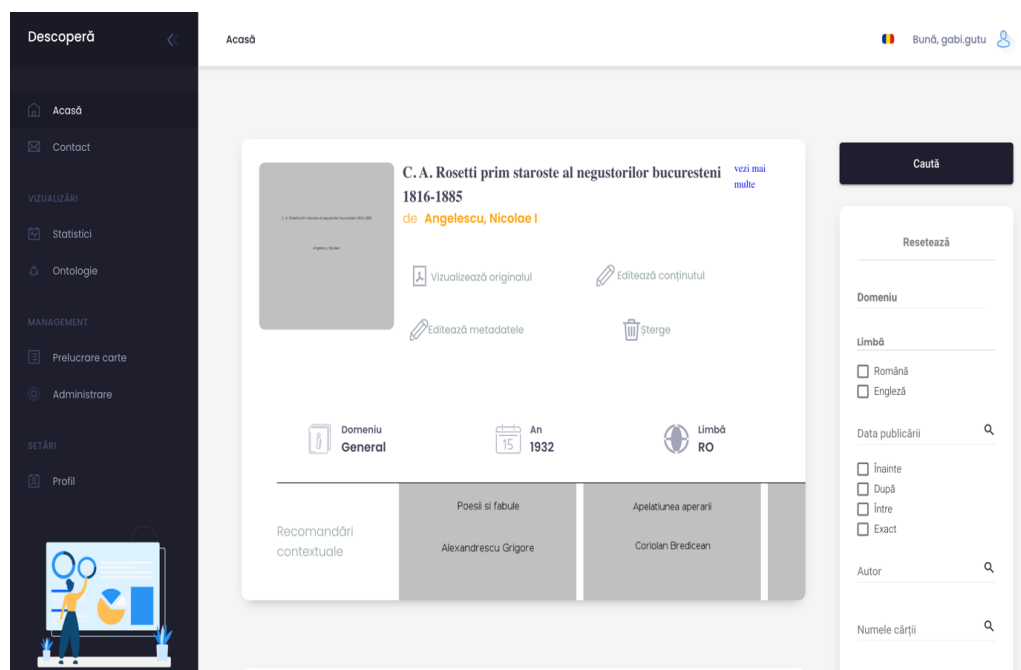


Figure 5. The Lib2Life homepage showing the latest added documents in the middle, together with a search area on the right.

4.1 Document processing

Librarians are provided with additional facilities, rather than exploring the documents and searching. These functionalities include the upload of documents and the automated processing service. This user role also involves books management tasks, like editing the content of either documents being uploaded, or for already registered ones, or even metadata adjustments for existing documents.

Restructuring of extracted text was required to enhance information retrieval. The identified problems in this extraction phase demanded a workflow applied sequentially on input documents that could correlate section titles with their content, recognize and reconstruct paragraph boundaries, combine hyphenated words, and accurately identify and extract images or tables. The title of the document, the author and the publication year are extracted from the first page (if available) and correlated with information from the enclosed XML file.

After defining global metadata, the table of contents (TOC) is automatically extracted based on specific keywords (e.g., “Contents”, “Table of contents”) searched at the beginning of the first few or last few pages of a document. If identified, a text area is pre-filled with the extracted information,

respectively the names of the chapters and the corresponding page where that section begins. Two approaches were considered: a) correlation of section titles with their content, and b) finding the predominant font type. OCR text may contain errors such as whitespaces or symbols, or it may be divided into several lines, which required further validation using regular expressions. The correlation of the section titles with the corresponding page numbers was performed by extracting information from lines in the table of contents ending in digits. TOC entries are analyzed using regular expressions to extract the section title and the associated page range.

For documents with no table of contents found, the second approach is used to recognize sections and paragraphs: the most common font in the document is identified; afterwards, different fonts are detected, which provide potential titles of the sections. Font name, font size, and text positions are stored in a list that is then used to identify the type of text (section title or content), comparing each line of text with the predominant font on that page. The two models were combined in a complex TOC extraction algorithm (Nitu et al., 2019), which can be easily adapted to most PDF document formats. The extracted text is then displayed in an editor, allowing librarians to correct it. If no table of contents is detected, the librarian still has the option to manually enter one. By clicking on a Next button, the librarian is redirected to the third step, not before being asked to confirm whether she wants to take into account the (detected or created) table of contents for text extraction.

In the third step of the document uploading process, the librarian can start several supplementary clean-up and enhancement processes: a) spellcheck and text correction – reconstructs the paragraphs by removing spaces, adds a whitespace to lines that end without punctuation marks, concatenates broken lines, and restores words divided into syllables, and b) image and table extractions – identifies images and tables included in the document; images and tables are then added at the end of the processed text in the corresponding section if a TOC was used, or at the end of the document if a TOC was not taken into account. At least one of the following conditions must be met to detect paragraphs: a) the preceding line marks an ending sentence, and the current line begins with a capital letter; or b) the current line starts with a hyphen and represents a replica of a conversation.

Image extraction is based on two approaches: a) recognizing image captions, and b) searching for shapes on each page and identifying the

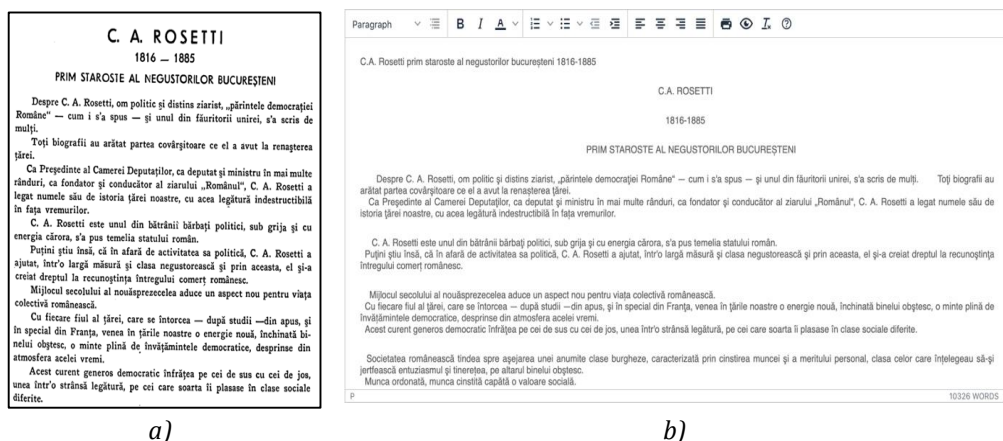
number of colors it contains. For example, if a page contains only text in the middle part of the top area of a page, as well as an irregularly shaped rectangle projected in white, grey, and black, then the rectangle is most likely an image, even if no associated caption is found. The image and text extraction were merged into a single iteration.

During text analysis, three heuristics were applied for finding images. First, if a word was found to describe a figure (for example, the “Figure” or “Fig.” words in their Romanian version were found), the location of that figure was saved. Second, an experiment on the number of characters of a page showed that images were always found on pages of less than 200 characters. Thus, the second heuristic compares the similarity of pixels and is applied to pages without characters; if the pixels on the page are identical, the page is considered empty and thus omitted. The extracted images are converted to Base64 format (Josefsson, 2006), and the title of the figure is inserted into a specific HTML element at the end of the section or at the end of the book, depending on whether the TOC was used or not when extracting the text from the document.

Table extraction is a similar process, in which the algorithm moves tables either at the end of the section, or at the end of the document, depending on the option to consider the TOC or not. The table extraction involves advanced processes due to the state of OCR documents provided in PDF format. An analysis of the document collection revealed that most of the tables contained twisted lines, missing data or had an irregular structure. Some documents contained tables that had even been handwritten. Currently, the existing text extraction software libraries are not able to accurately detect the boundaries of tables or the content itself. Table extraction based on the Nurminen detection algorithm from the tabula API (<https://tabula.technology>) demonstrated an average accuracy of 40%. If a table of contents is used, table extraction is applied to each page separately, storing the coordinates of all tables and mapping the table to that section. If no table of contents was used, the details of the detected tables are stored and associated with page numbers, and then added to the end of the section or document, depending on the identified structure of the document.

In the last phase of the document upload process, the librarian can manually edit the extracted content. The interface includes a visual text editor similar to Microsoft Word, so that text can be corrected, while other elements such as font type, text size, or text color can also be introduced. The

purpose of this manual correction phase is to optimize the document by defining headings to separate the content and to provide a coherent web rendering. A comparison between the original text and the extracted text of a page is presented in Figure 6. The librarian also can delete a document already uploaded and processed by the system. This can be done in case it was incorrectly added or it is a duplicate.



a) b)
Figure 6. A preview of a) a page viewed in the original PDF document, and b) the text of the same page extracted and processed.

4.2 Document Search and Filtering

Recommended documents are provided with the help of the K-Nearest Neighbors algorithm (Keller et al., 1985) that uses semantic distances based on word embeddings provided by the ReaderBench framework (Dascalu et al., 2017). The searching functionality available in the web application relies on the keywords extraction function available in the ReaderBench platform. The set of most important keywords is used to find the closest k entries (which are represented by paragraphs in the context of Lib2Life) from Elasticsearch using a “more like this” query. The query result consists of a list of paragraphs containing at least one of the keywords, sorted by their relevance. The semantic similarity between these paragraphs and the internal representation of the query is then computed to determine the most similar documents to the user’s input text. Starting with those paragraphs, the algorithm identifies appropriate documents for each paragraph and provides a list of the most similar documents (i.e., documents that contain most related paragraphs to the input query).

Document filtering can be performed on multiple criteria, such as: book name, author name, domain, or year of publication. Advanced filtering

is based on the semantic similarity between the search text and the content of existing documents in Elasticsearch, as portrayed in Figure 7, where recommended documents are displayed in the bottom of the section specific to the given document. Each document can be visualized either as the original PDF document, or as a cleaned and coherent web rendering of the processed content.

5. Conclusions and Future Work

The Lib2Life platform aims to store and provide smart access to digitised and curated versions of historical documents held by the four central university libraries in Romania. As of January 2021, the platform stores approximately 300 documents. Readers can easily explore the collection of documents, either by using a search functionality based on keywords or by following the semantic recommendations of documents, when exploring a specific book. Both services use semantic recommendations relying on Natural Language Processing techniques to find relevant documents to user's query, or similar documents. The platform also includes filters for limiting the results of document searches, as well as the capability to read a document in its original PDF format, or in a cleaned format containing processed text extracted from the PDF. Readers can also explore the ontology incorporated within the platform, which emphasize the main classes, their entities, and attributes. This chapter presented a few queries applied on the ontology in order to extract relevant information about the available data, which demonstrate the usefulness of the platform for finding relevant information using SPARQL queries.

In terms of future work, the Lib2Life ontology will be adapted to fully match the content stored in the system. Based on the content of the documents present in the system, a matching algorithm to assign a new document within one of the seventeen domains will be created. The Lib2Life platform will also incorporate a social recommendation component, relying on preferences expressed by other readers, besides the semantic recommendations already available in the platform.

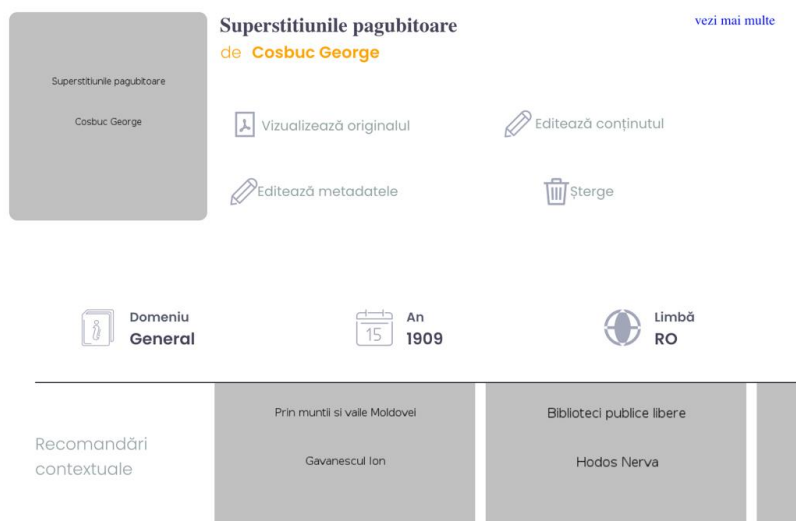


Figure 7. Semantic recommendations of a selected document (shown at the bottom).

Acknowledgment

The work was funded by the Operational Programme Human Capital of the Ministry of European Funds through the Financial Agreement 51675/09.07.2019, SMIS code 125125, by a grant of the Romanian Ministry of Research and Innovation, CCCDI – UEFISCDI project number PN-III-P1-1.2-PCCDI-2017-0689 “Revitalizing Libraries and Cultural Heritage through Advanced Technologies”.

References

- Alewaeters, G. 1982. VUBIS: A user-friendly online system. *Information Technology and Libraries*, 1, 206-21.
- Apache Foundation. 2019. *Apache Jena Fuseki Documentation* [Online]. Available: <https://jena.apache.org/documentation/fuseki2/> [Accessed August 15 2019].
- Banu, A., Fatima, S. S. & Khan, K. U. R. 2013. Building OWL Ontology: LMSO-Library Management System Ontology. *Advances in Computing and Information Technology*. Springer.
- Dascalu, M., Gutu, G., Ruseti, S., Paraschiv, I. C., Dessus, P., Mcnamara, D. S., Crossley, S. & Trausan-Matu, S. ReaderBench: A Multi-Lingual Framework for Analyzing Text Complexity. In: LAVOUÉ, E., DRACHSLER, H., VERBERT, K., BROISIN, J. & PÉREZ-SANAGUSTÍN, M., eds. 12th European Conference on Technology Enhanced Learning (EC-TEL 2017), 2017 Tallinn, Estonia. Springer, 495–499.
- European Comission 2011. Comission Recommendation of 27 October 2011 on the digitisation and online accessibility of cultural material and digital preservation. *Official Journal of the European Union*, 283, 39–45.

- Fernández-López, M., Gómez-Pérez, A. & Juristo, N. 1997. Methontology: from ontological art towards ontological engineering.
- Golbeck, J. & Rothstein, M. Linking Social Networks on the Web with FOAF: A Semantic Web Case Study. 23rd AAAI Conference on Artificial Intelligence, 2008. 1138–1143.
- Gormley, C. & Tong, Z. 2015. *Elasticsearch: The definitive guide: A distributed real-time search and analytics engine*, O'Reilly Media, Inc.
- Gruber, T. R. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5, 199–220.
- Gutu-Robu, G., Ruseti, S., Tomescu, S., Dascalu, M. & Trausan-Matu, S. Designing an Ontology for Knowledge-based Processing in Romanina University Libraries. 8th Int. Workshop on Semantic and Collaborative Technologies for the Web, in conjunction with the 16th Int. Conf. on eLearning and Software for Education (eLSE 2020), 2020 Online. "CAROL I" National Defence University Publishing House, 119–126.
- Haslhofer, B. & Isaac, A. data.europeana.eu: The europeana linked open data pilot. International Conference on Dublin Core and Metadata Applications, 2011. 94–104.
- Josefsson, S. 2006. *The Base16, Base32, and Base64 Data Encodings* [Online]. Online: Internet Engineering Task Force RFC. Available: <https://tools.ietf.org/html/rfc4648> [Accessed September 15th 2020].
- Julich, S., Hirst, D. & Thompson, B. 2003. A case study of ILS migration: Aleph500 at the University of Iowa. *Library hi tech*, 21, 44–55.
- Keller, J. M., Gray, M. R. & Givens, J. A. 1985. A fuzzy k-nearest neighbor algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 4, 580–585.
- Labadi, S. 2013. *UNESCO, cultural heritage, and outstanding universal value: Value-based analyses of the World Heritage and Intangible Cultural Heritage Conventions*.
- Lassila, O. & Swick, R. R. 1998. Resource description framework (RDF) model and syntax specification. World Wide Web Consortium.
- Lebert, M. 2008. *Project Gutenberg (1971-2008)*, NEF, University of Toronto & Project Gutenberg.
- Lohmann, S., Negru, S., Haag, F. & Ertl, T. 2016. Visualizing ontologies with VOWL. *Semantic Web*, 7, 399–419.
- Mcguinness, D. L. & Van Harmelen, F. 2004. OWL Web Ontology Language overview. *W3C recommendation*, 10, 2004.
- Mitocaru, I., Gutu-Robu, G., Nitu, M., Dascalu, M., Trausan-Matu, S., Tomescu, S. & Florescu, G. The Lib2Life Platform – Processing, Indexing and Semantic Search for Old Romanian Documents. International Conference on Human-Computer Interaction (RoCHI2020), 2020 Online. MatrixRom.
- Nitu, M., Dascalu, M., Dascalu, M.-I., Cotet, T.-M. & Tomescu, S. Reconstructing Scanned Documents for Full-text Indexing to Empower Digital Library Services. 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), 2019 Magdeburg, Germany. Springer, 183–190.
- Noy, N. F., Crubezy, M., Ferguson, R. W., Knublauch, H., Tu, S. W., Vendetti, J. & Musen, M. A. Protege-2000: an open-source ontology-development and

- knowledge-acquisition environment. AMIA Annual Symposium, 2003 Washington, DC, USA. American Medical Informatics Association, 953.
- Room, I. X. 1972. Convention concerning the protection of the world cultural and natural heritage.
- Stroube, B. 2003. Literary freedom: Project gutenber. *XRDS: Crossroads, The ACM Magazine for Students*, 10, 3-3.
- W3C. 2013. *SPARQL Query Language for RDF* [Online]. Available: <https://www.w3.org/TR/rdf-sparql-query/> [Accessed August 15 2019].
- Weibel, S., Kunze, J., Lagoze, C. & Wolf, M. 1998. *Dublin Core Metadata for Resource Discovery* [Online]. Internet Engineering Task Force RFC. Available: <https://tools.ietf.org/html/rfc2413> [Accessed September 15th 2020].

Computer-assisted methods in historical linguistics

Alina Cristea, Anca Dinu, Liviu P. Dinu, Simona Georgescu, Ana Uban

(University of Bucharest)

1. Introduction

Natural languages are living ecosystems, they are constantly in contact and, by consequence, they continuously change. The genetic classification of the languages and language change across space and time are two of the main topics in historical linguistics and are significant from multiple points of view (scientific, social, economical, cultural and technological). From a scientific perspective, any advance in historical linguistics is of paramount cultural importance, being inherently connected with human history (Campbell, 1998). Therefore, it is an important source of information for other fields of prehistoric studies, such as archaeology, paleoanthropology or, in recent years, paleogenetics (Haak, 2015), as well as for the evolution of reading, writing, and the human species itself. Along these lines, Longobardi (LanGeLin project¹, 2012) explored the potential correlation of genetic and linguistic distances, starting from what he called Darwin's last challenge: *"If we possessed a perfect pedigree of mankind, a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world; and if all extinct languages, and all intermediate and slowly changing dialects, were to be included, such an arrangement would be the only possible one"* (Darwin, 1859). Language is part of our personal and social identity and any findings in historical linguistics have a meaningful socio-cultural impact, since language evolution faithfully reflects the evolution of society. For example, the relationship between the Romance languages is not limited to their

¹ <https://www.york.ac.uk/language/research/projects/completed/langelin/project/>

genesis, but it has uninterruptedly continued throughout their history; the constant lexical exchange between them shows that the genetic relationship has led to a far stronger contact than the geographic proximity (cf. Jucquois, 2001; Ciobanu and Dinu, 2014a).

Historical linguistics – the original approach to languages that set linguistics as a science three centuries ago – still has to deal with unanswered questions, mainly in the field of etymology. For instance, it is still difficult to decide whether two given similar words in related languages, A and B, are cognates or loanwords.² If they are cognates, the question that must be answered concerns the protoword, and if they are loanwords, the aim is to detect whether A was borrowed from B or *vice versa*.

Traditionally, these problems have been investigated with comparative linguistics instruments (Meillet, 1954; Campbell, 1998), which imply a time-consuming manual process conducted by highly qualified linguists and domain experts, and require a large amount of intensive work. Lately, it has become apparent that experts no longer have enough time and resources to analyze the massive amount of (digital) data that becomes available. For instance, the etymologists' efforts to reconstruct the proto-languages still encounter numerous gaps while using exclusively the classical, manual methods. To take but one example in the Romance etymology, the newest project drawn by an international group of more than 50 linguists for the last 12 years (DÉRom³), is struggling with the bulk work of manually consulting the lexicographic and text corpora, having reached by now only a fifth of what was aimed at, by the end of the first three years (Buchi/Schweickard, 2008; Buchi/Schweickard, 2010).

Under these circumstances, although comparative linguistics has enabled significant advances in the field, only a small part of the over 7,000 languages spoken today have received comparative study attention.

The sheer need to automatically treat the ever increasing volume and complexity of the digitalized data in comparative linguistics has raised the interest in computational tools designed to assist scholars in the tedious aspects of their work, and traditional methods have started to benefit from

² Cognates are words in different languages having the same meaning and a common ancestor. A borrowed word, also called loanword, is defined as a "lexical item (a word) which has been 'borrowed' from another language, a word which originally was not part of the vocabulary of the recipient language but was adopted from some other language and made part of the borrowing language's vocabulary" (Campbell, 1998).

³ <http://www.atilf.fr/DERom>

modern computational approaches (Kondrak et al 2003; Kondrak, 2004, Bouchard-Cote et al 2007, 2013; Ciobanu and Dinu 2014b, 2015, 2018, 2019; List2012; List et al 2017).

As comparative historical linguistics has to deal mainly with linguistic change, either at lexical or at semantic level, computational tools can be designed to serve both, for instance to automatically identify related words and to assess the semantic change. From the 7 steps which form the work-flow (Ross and Durie, 1996) that scholars implicitly follow in historical linguistics, lexicostatistics (e.g. Swadesh 1952) can be seen as an early attempt to give an algorithmic rendering of reconstruction of phylogenies, even though it pre-dates the computer age. Since then, historical linguistics has leveraged technological advances (reluctantly at the beginning), and some pioneering work was initiated on other steps of the work-flow (cognate identification, protoword reconstruction).

The cooperation between computer scientists and traditional linguists, as far as it exists, is still far from being perfect. There is a gap between the backgrounds, methodologies and expertise of the two groups, which often leads them to mistrust and misunderstand each other. On the one hand, the classical linguists have a deep understanding of their field, coming from a long tradition, which computer scientists are not trained in, and often do not understand the subtleties that their models need to encode and process. On the other hand, computer scientists dispose of an arsenal of powerful new digital tools and methods, which traditional linguists are ignorant of, or even fear, because of the common belief that AI will replace traditional research. Computers cannot and should not aim to replace the experience and intuitions of classical linguists. Instead, a fruitful middle ground can be reached when each side contributes with its strengths and leans on the other to compensate for its weaknesses. Computers can be most useful for historical linguistics when computer scientists understand the specificities of the linguistic problems and adapt their methods to cope with often very complex and sometimes small sets of data, avoiding as much as possible the black box models. Similarly, traditional linguists could benefit from relying more on digital tools, which can be used to speed up tedious work and potentially reveal a fresh perspective. In the end, instead of the competing computer-based versus classical approaches, one needs a complementary paradigm of computer-assisted approaches, neither completely computer-driven, nor ignorant of the help computers provide (List, 2017).

2. Previous Work

As far as traditional historical linguistics is concerned, the main advances of the classical methods have been made in the field of phonetics, morphology and syntax – the domains considered as the only linguistic fields that would allow a rigorous scientific approach. As the lexicon has been usually perceived as infinite and the semantic evolution as chaotic, the lexico-semantic domain has been constantly left behind in the attempts to find a universally applicable theory (see the works on typology and linguistic universals, almost exclusively oriented towards phonetics and syntax, cf. Croft, 2003; Mairal/Gil, 2006; Good, 2008). Nevertheless, in recent years, new historical linguistics perspectives have been put forth also at the lexico-semantic level.

As previously mentioned, a challenging problem regarding the task of language classification is distinguishing between cognates and borrowings. Borrowings mistakenly taken for cognates bias the genetic classification of the languages, characterizing them as being closer than they actually are (Minett and Wang, 2003; Gray and Atkinson, 2003). False cognates are more harmful than missing valid cognates in language comparison, because they can lead to incorrect conclusions regarding the genetic relationships between languages (List, Greenhill, and Gray, 2017). Thus, the need for discriminating between cognates and borrowings emerges. The challenge and importance of this task is emphasized by Heggarty (2012, page 122) as follows:

“What solution is there, then, if we can neither ignore the problem of distinguishing cognates from loanwords, nor overcome it in the many cases where we do not have the necessarily linguistic knowledge to do so? There is in fact a possibility: to sidestep the question entirely at the data analysis stage, and simply to identify which forms are judged to be somehow correlated with each other, whether by specialists in those languages, or more objectively by computerized approaches.”

Furthermore, Jager (2019) considers the treatment of borrowing (and language contact in general) is an unsolved problem for computational historical linguistics.

Languages borrow words from other languages primarily because of need and prestige. (Campbell 1998, page 59). Thus, this research topic facilitates reconstructing certain aspects related to society and culture for groups of people speaking a given protolanguage, and gaining insights into

their past social interactions and into their social and cultural practices (Epps, 2014). Moreover, establishing the direction and source of borrowing is important to our understanding of the social relations between the groups involved.

Another emerging new lexico-semantic direction of study is computational protoword reconstruction. Complete automation of the reconstruction process is still a desideratum. Oakes (2000) proposed two systems (Jakarta and Prague) that, combined, cover the steps of the comparative method for proto-language reconstruction: discovering regular sound changes, statistically evaluating the identified sound changes, using them to verify real word pairs, and proposing rules to infer the ancestor words from their descendants. The work of computational biologists such as Alexandre Bouchard-Cote (2007), Russell Gray (2003), Robert McMahon (2006), Mark Pagel (2007, 2013) and co-workers took the protoword reconstruction one step further, by applying methods from computational biology to the problem of the reconstruction of language history, often in collaboration with linguists.

Furthermore, the semantic change leading to divergent meanings in pairs of cognates still hasn't been properly studied, mainly for lack of appropriate corpora and preestablished scientific tools. This *aim* could be more easily achieved by using computational tools. Computational models of word meaning based on distributional semantics (Mikolov et al., 2013, Pennington et al., 2014) have shown that word representations generated automatically from large corpora, based on word co-occurrence, can encode word meanings and semantic distances between words. These models allow computational linguists to effectively translate the assessment of semantic distance and semantic divergence from the linguistic domain to the one of vector algebra. Moreover, if the word representations are developed based on diachronic corpora, it is possible to identify and measure shifts in word meanings across time (Hamilton et al., 2016): the geometrical distance travelled by a word representation in the vector spaces corresponding to the different time periods can be used to infer the extent and direction of its semantic drift.

The same approaches can be extended to the multilingual domain by using techniques based on matrix transformations for joining vectorial representations from multiple monolingual domains to one multilingual domain where words from different languages can be represented together

in the same semantic space. These techniques together make it possible to develop algorithms for quantifying the semantic divergence of cognates across languages and across time. Our results from initial computational studies on cognate semantic divergence for Romance languages suggest that the more distant the language pair, the more cognates in the respective languages tend to diverge in meaning (Uban et al., 2019).

3. Case Study - Identifying Related Words and Reconstructing Protowords

In (Ciobanu and Dinu, 2014b, 2015, 2018, 2019) we introduced a computerized approach for two language change-related tasks: identifying related words and reconstructing protowords. We focused on two types of related words: cognates and borrowings.

Investigating pairs of cognates is very useful not only in historical and comparative linguistics (in the study of language relatedness (Ng et al., 2010), phylogenetic inference (Atkinson et al., 2005) and in identifying how and to what extent languages changed over time or influenced each other), but also in other research areas, such as language acquisition, bilingual lexicon induction (Heyman et al., 2017), corpus linguistics (Simard et al., 1993), cross-lingual information retrieval (Buckley et al., 1997) and machine translation (Kondrak et al., 2003).

Our goal was to automatically identify the relationship between words. More specifically, we proposed a methodology for identifying cognates and for discriminating between cognates and borrowings. We employed an orthographic alignment method and we used aligned subsequences as features for machine learning classification algorithms, in order to infer rules for linguistic changes undergone by words when entering new languages and to identify if and how the words are related. Firstly, we addressed the task of identifying cognates. Given a pair of words (u, v), we determined with a good accuracy whether they are cognates or not. Secondly, we investigated the task of discriminating between cognates and borrowings. Given a pair of words (u, v), we determined whether they are cognates or v is the etymon of u .

Methodology

Our methodology for identifying related words (cognates and borrowings) is described by the following work-flow:

- 1) Aligning the pairs of related words using a string alignment algorithm;
- 2) Extracting features from the aligned words;
- 3) Using machine learning classification algorithms to discriminate between the classes (cognates vs. non-cognates, or cognates vs. borrowings).

We employed orthographic alignment for identifying related words, not only to compute similarity scores, as was previously done, but to use aligned subsequences as features for machine learning algorithms. Our intuition was that inferring language-specific rules for aligning words would lead to better performance in the task of identifying related words.

For identifying cognates, we applied our method on a subset of the automatically extracted dataset of cognates that we previously developed (Ciobanu and Dinu, 2014c). We used four pairs of languages: Romanian - French, Romanian - Italian, Romanian - Spanish and Romanian - Portuguese. In Table 1 we report the results for automatic identification of cognates using orthographic alignment.

TABLE 1

Languages	Naive Bayes		SVM	
	accuracy	n-gram size	accuracy	n-gram size
It - Ro	79.0	1	81.5	1
Fr - Ro	82.0	2	87.0	2
Es - Ro	84.0	1	86.5	2
Pt - Ro	73.0	2	86.5	2

The best results are obtained for French and Spanish, while the lowest accuracy is obtained for Portuguese. The SVM produces better results for all considered languages except Portuguese, where the accuracy is equal. For Portuguese, both Naive Bayes and SVM misclassify more non-cognates than cognates. A possible explanation might be the occurrence, in the dataset, of more remotely related words, which are not labeled as cognates. To test this assumption, we analyzed the errors of the system and observed that around

38% of the pairs misclassified as cognates are actually more remotely related words.

For discriminating between cognates and borrowings, we applied our method on four pairs of languages: Italian - Romanian, Portuguese - Romanian, Spanish - Romanian and Turkish - Romanian. For the first three pairs of languages, which include sister languages, most cognate pairs have a Latin common ancestor, while for the fourth pair, which includes languages belonging to different families (Romance and Turkic), most of the cognate pairs have a common French etymology, and date back to the end of the 19th century, when both Romanian and Turkish borrowed massively from French. In Table 2 we provide examples of borrowings and cognates.

TABLE 2

Languages	Borrowings	Cognates
It - Ro	<i>baletto</i> → <i>balet</i> (translation: <i>ballet</i>)	<i>vittoria</i> - <i>victorie</i> (translation: <i>victory</i>) (etymon: Latin <i>victoria</i>)
Pt - Ro	<i>selva</i> → <i>selva</i> (translation: <i>selva</i>)	<i>instinto</i> - <i>instinct</i> (translation: <i>instinct</i>) (etymon: Latin <i>instinctus</i>)
Es - Ro	<i>machete</i> → <i>maceta</i> (translation: <i>machete</i>)	<i>castillo</i> - <i>castel</i> (translation: <i>castle</i>) (etymon: Latin <i>castellum</i>)
Tr - Ro	<i>tütün</i> → <i>tutun</i> (translation: <i>tobacco</i>)	<i>aranjman</i> - <i>aranjament</i> (translation: <i>arrangement</i>) (etymon: French <i>arrangement</i>)

The dataset contains borrowings and cognates that share a common ancestor. Table 3 shows the results for automatic discrimination between cognates and borrowings. SVM obtained, in most cases, better results than Naive Bayes. The best results were obtained for Turkish - Romanian, with an accuracy of 90.1, followed by Portuguese - Romanian with 90.0 and Spanish - Romanian with 85.5 (for Portuguese - Romanian, Naive Bayes obtained a slightly better result than SVM, with an accuracy of 91.6). These results showed that, for these pairs of languages, the orthographic cues are different with regard to the relationship between words. For Italian - Romanian we obtained the lowest accuracy, 69.0.

TABLE 3

Languages	Naive Bayes		SVM	
	accuracy	n-gram size	accuracy	n-gram size
It - Ro	68.1	3	69.0	3
Pt - Ro	91.6	3	90.0	3
Es - Ro	84.4	3	85.5	2
Tr - Ro	89.3	3	90.1	3

In this experiment, we knew beforehand that there is a relationship between words, and our aim was to identify the type of relationship. However, in many situations this kind of a-priori information is not available. In a real scenario, we would have either to add an intermediary classifier for discriminating between related and unrelated words, or to discriminate between three classes: cognates, borrowings, and unrelated. We augmented our dataset with unrelated words (determined based on their etymology), building a stratified dataset annotated with three classes, and we repeated the previous experiment. The performance decreased, but the results were still significantly better than chance (99% confidence level). We obtained the following accuracy values on the test sets, when using the SVM classifier: Italian - Romanian 63.8, (Romanian was always the recipient language in our dataset, i.e., the language that borrowed the words). Portuguese - Romanian 77.6, Spanish - Romanian 74.0, Turkish - Romanian 86.0. For Italian, borrowings turned out to be the most difficult class to identify correctly. For Turkish, the cognates were slightly more difficult to identify correctly compared to other classes, while for Portuguese and Spanish the lowest performance was obtained for unrelated words. In both cases, they were more often mistakenly identified as cognates than as borrowings.

Reconstructing Latin Protowords

For reconstructing protowords, we applied a sequence labeling method (CRF) that, given cognate sets in modern languages, produces the form of their protowords (Ciobanu and Dinu 2018, Ciobanu and Dinu 2019).

From the alignment of related words in the training set, the system learnt orthographic patterns for the changes in spelling between the source and the target language. The method that we employed (Ciobanu and Dinu,

2018) is based on sequence labeling, an approach that has been proven useful in generating transliterations (Ganesh et al., 2008; Ammar et al., 2012). In our case, the words were the sequences, and their characters were the tokens. Our purpose was to obtain, for each input word, a sequence of characters that compose its related word. To this end, we used conditional random fields (CRFs) (Lafferty et al., 2001). As features for the CRF system, we used character n -grams from the input words, extracted from a fixed window w around the current token. To align pairs of words, we employed the Needleman and Wunsch (1970) global alignment algorithm. We also proposed several ensemble methods for combining information from multiple systems, with the purpose of joining the best productions from all modern languages.

To assess how close the productions are to the correct form of their related words, we reported the edit distance between the produced words and the gold standard. We reported both the un-normalized and the normalized edit distance. For normalization in the $[0,1]$ interval, we divided the edit distance by the length of the longest string.

We reported the coverage as well (COV- n , also known as top n accuracy), a relaxed metric which computes the percentage of input words for which the n -best output list contains the correct proto-word (the gold standard). We use n in $\{1, 5, 10\}$. The practical importance of analyzing the top n results is that we offer a filter to narrow down the possible forms of the output words to a low-dimensional list, that linguists can analyze, aiming to identify the correct form of the proto-word. Note that the coverage for $n = 1$ is the well-known measure of accuracy.

In Table 4 we report the results of our individual systems (one for each modern language) and the ensemble results for reconstructing proto-words on two datasets: 1) The dataset proposed by Reinheimer Ripeanu (2001), consisting of 1,102 cognate sets in five Romance languages (Spanish, Italian, Portuguese, French, Romanian) and their common Latin ancestors; some of the cognate sets are incomplete. 2) The dataset proposed by Ciobanu and Dinu (2014c), consisting of 3,218 complete cognate sets in five Romance languages (Spanish, Italian, Portuguese, French, Romanian) and their common Latin ancestors.

TABLE 4

Language	Dataset: Reinheimer Ripeanu (2001)		Dataset: Ciobanu and Dinu (2014c)	
	Edit distance	COV-5	Edit distance	COV-5
Italian	1.57 (0.24)	0.52	1.12 (0.14)	0.62
Spanish	1.78 (0.27)	0.35	1.31 (0.16)	0.59
Portuguese	1.76 (0.28)	0.34	1.30 (0.16)	0.58
Romanian	2.12 (0.32)	0.31	1.36 (0.16)	0.61
French	2.31 (0.35)	0.24	1.52 (0.18)	0.57
Ensemble	1.55 (0.23)	0.49	1.07 (0.13)	0.70

For individual experiments, Italian obtained the lowest average edit distance on all datasets. The best-performing ensemble uses a fusion method which assigns scores based on the rank of the productions in the n -best lists and the training accuracy for each individual system. We also tried applying the ensembles on language subsets (that is, not to take all modern languages into account at once). We investigated all combinations, and in the majority of cases using all modern languages lead to the highest performance among all ensembles. The average edit distance of our best-performing ensemble, on the dataset from Ciobanu and Dinu (2014c) is 1.07, meaning that, on average, the proto-word reconstructions obtained by the system are a little more than one character different from the correct proto-words. Furthermore, the correct proto-word is listed among the 5-best list productions of our system in 70% of the cases.

In Table 5 we show an example of our systems' output n -best lists. This example illustrates how the ensemble can improve over the individual classifiers, by ranking the correct production higher than all the other systems. For all datasets we obtained performance improvements for proto-word reconstruction when we combined individual results using ensembles.

TABLE 5

Language	Word	5-best productions
French	voisin	vosinum, vosnum, vosine, vosinus, voiinum
Italian	vicino	vicinum, vicinus , vicenum, vicenus, vicnum
Portuguese	vizinho	vizinus, vizinum, vicinus , vizinium, vizinnum
Spanish	vecino	vecinum, vecinus, vicinum, vecenum, vicinus
Romanian	vecin	vicenus, vicenum, vicinus , vicinum, vecenus
Ensemble	all words	vicinus , vicinum, vicenus, vicenum, vecinus

Looking at the incorrect productions that have one character different from the correct proto-word, we noticed that sometimes the final consonant is mistaken (5% of the errors). Most commonly, *um* instead of *us* (4.1% of the errors): *serenum* instead of *serenus*, *cantum* instead of *cantus*, *novum* instead of *novus*⁴. Another one-character mistake is, sometimes, failing to double a consonant (4% of the errors). For example, *ll* or *ss*: *colapsus* instead of *collapsus*, *intervalum* instead of *intervallum*, *disociatio* instead of *dissociatio*, *esentia* instead of *essentia*. For productions that have two characters different from the correct proto-word, we noticed the following patterns in the incorrect productions: sometimes the character *f* is mistakenly obtained instead of *ph*: *asfaltus* instead of *asphaltus*, *eufonia* instead of *euphonia*, *diafragma* instead of *diaphragma*. Another interesting pattern is obtaining the desinence *-a* instead of *-us*: *citrina* instead of *citrinus*, *alba* instead of *albus*. When this occurs for adjectives, the productions are not incorrect words in Latin; we obtain the feminine form instead of the masculine.

4. Conclusions

The results obtained so far through applying computational methods for (approaching) problems in historical linguistics are encouraging/promising, and suggest that a close collaboration between classical researchers and

⁴ This difference is only valid if we take as a comparison basis Classical Latin dictionaries that register the *-us* Nominative form as primary. In fact, the Romance languages inherit an Accusative form in *-um*, thus the result we obtained is not a real error.

computational researchers could lead to an improvement in the accuracy of the methods, and moreover to results and insights which would otherwise be difficult to envision/reach by either computational or classical research independently. Thus, the use of computational methods, alongside the classical methodology and rooted in linguistic theory, can help to accelerate the rate of scientific discovery, and can lead to novel solutions, and even to entirely new directions of research.

Acknowledgments. *This research is supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI UEFISCDI, project number 108, COTOHILI, within PNCDI III.*

References:

- Waleed Ammar, Chris Dyer, and Noah A Smith. 2012. Transliteration by sequence labeling with lattice encodings and reranking. In *Proceedings of the 4th Named Entity Workshop*, pages 66–70.
- Quentin D. Atkinson, Russell D. Gray, Geoff K. Nicholls, and David J. Welch. 2005. From Words to Dates: Water into Wine, Mathemagic or Phylogenetic Inference? *Transactions of the Philological Society*, 103:193–219.
- Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2013. Cognate Production using Character-based Machine Translation. In *Proceedings of IJCNLP 2013*, pages 883–891.
- Alexandre Bouchard-Cote, Percy Liang, Thomas Griffiths, and Dan Klein. 2007. A Probabilistic Approach to ´ Diachronic Phonology. In *Proceedings of EMNLP-CoNLL 2007*, pages 887–896.
- Alexandre Bouchard-Cote, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110(11):4224–4229.
- Buchi, Éva &Schweickard, Wolfgang (2008) : « Le *Dictionnaire Étymologique Roman* (DÉRom) : en guise de faire-part de naissance ». *Lexicographica. International Annual for Lexicography* 24, 351-357
- Buchi, Éva &Schweickard, Wolfgang (2010) : « À la recherche du protoroman : objectifs et méthodes du futur *Dictionnaire Étymologique Roman* (DÉRom) ». In : Iliescu, Maria, Siller-Runggaldier, Heidi &Danler, Paul (éd.) : *Actes du XXV^e Congrès International de Linguistique et de Philologie Romanes (Innsbruck 2007)*, Berlin/New York, De Gruyter, vol. 6, 61-68
- Chris Buckley, Mandar Mitra, Janet A. Walz, and Claire Cardie. 1997. Using Clustering and SuperConcepts Within SMART: TREC 6. In *TREC*, pages 107–124.
- Lyle Campbell. 1998. *Historical Linguistics. An Introduction*. MIT Press.
- Alina Maria Ciobanu, Liviu P. Dinu. 2014a. An Etymological Approach to Cross-Language Orthographic Similarity. Application on Romanian. In *Proceedings of EMNLP 2014*, pages 1047–1058.

- Alina Maria Ciobanu, Liviu P. Dinu, 2014b. Automatic Detection of Cognates Using Orthographic Alignment. In Proc. ACL (2) 2014 (the 52nd Annual Meeting of the Association for Computational Linguistics), 99-105, June 22-27, 2014, Baltimore, MD, USA
- Alina Ciobanu, Liviu P. Dinu, 2014c. Building a Dataset of Multilingual Cognates for the Romanian Lexicon. In Proc. LREC 2014 (9th International Conference on Language Resources and Evaluation), Reykjavik, Iceland, 26-31 may 2014, p. 1038-1043.
- Alina Maria Ciobanu, Liviu P. Dinu, 2015. Automatic Discrimination between Cognates and Borrowings. In Proc. ACL (2) 2015 (the 53rd Annual Meeting of the Association for Computational Linguistics), July 26-31, 2015, Beijing, China, p. 431-437.
- Alina Maria Ciobanu, Liviu P. Dinu, 2018. Ab Initio: Latin Proto-word Reconstruction. In Proc. COLING 2018 (27th International Conference on Computational Linguistics), 1604-1614, Santa Fe, New Mexico, USA, August 20-26.
- Alina Maria Ciobanu, Liviu P. Dinu, 2019. Automatic Identification and Production of Related Words for Historical Linguistics. Computational Linguistics, vol. 45, No. 4, 667-704, December 2019.
- Charles Darwin. 1859. On the Origin of Species, London: John Murray.
- Croft, William (2003), Typology and Universals (second edition), Cambridge Univers.
- Epps P. Historical linguistics and socio-cultural reconstruction. In: Bowers C, Evans B, editors. The Routledge Handbook of Historical Linguistics. London: Routledge; 2014. p. 579-97.
- Surya Ganesh, Sree Harsha, Prasad Pingali, and Vasudeva Verma. 2008. Statistical transliteration for cross language information retrieval using hmm alignment model and crf. In Proceedings of the 2nd Workshop on Cross Lingual Information Access.
- Russel Gray and Quentin Atkinson. 2003. Language tree divergences support the Anatolian theory of Indo-European origin. *Nature*, 426:435-439.
- Haak W. et al. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522(7555). 207-211.
- Hamilton, William L., Jure Leskovec, and Dan Jurafsky. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1489-1501. 2016.
- Paul Heggarty. 2012. In Beyond lexicostatistics: How to get more out of "word list" comparisons. In Soren Wichmann and Anthony P. Grant, editors, Quantitative Approaches to Linguistic Diversity: Commemorating the Centenary of the Birth of Morris Swadesh. Benjamins, pages 113-137.
- Geert Heyman, Ivan Vulić, and Marie Francine Moens. 2017. "Bilingual lexicon induction by learning to combine word-level and character-level representations." In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 1085-1095.
- Gerhard, J. 2019. Computational historical linguistics. Theoretical Linguistics | Volume 45: Issue 3-4.

- Guy JUCQUOIS, 2001. Parenté génétique et correspondances phonétiques. A propos du proto-mondial in « Travaux neuchâtelois de linguistique », 2001, 34/35, 59-84.
- Good, Jeff (ed.) (2008), *Linguistic Universals and Language Change*, Oxford University Press.
- Grzegorz Kondrak, Daniel Marcu, and Keven Knight. 2003. Cognates Can Improve Statistical Translation Models. In *Proceedings of HLT-NAACL*, pages 46–48.
- Kondrak, Grzegorz. 2004. Combining evidence in cognate identification. In *Proceedings of the 17th Conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence*, pages 44–59, London.
- John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of ICML 2001*, pages 282–289.
- Mairal, Ricardo/Gil, Juana (eds.) (2006), *Linguistic Universals*, Cambridge University Press.
- List, Johann Mattis. 2012. LexStat: Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS and UNCLH*, pages 117–125, Avignon.
- Johan Mattis List, 2017. Computer-assisted approaches in the humanities. Talk held at the workshop "Research questions in the humanities as challenges to computer science" (Max Planck Institute for the History of Science, Berlin, 2017/12/06-07).
- Johann Mattis List, Simon J. Greenhill, and Russell D. Gray. 2017. The potential of automatic word comparison for historical linguistics. *PLOS ONE*, 12(1):1–18.
- Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. <http://mallet.cs.umass.edu>.
- McMahon, A. & R. McMahon. 2006 Why linguists don't do dates: Evidence from Indo-European and Australian languages. In P. Forster & C. Renfrew (eds.), *Phylogenetic methods and the prehistory of languages*, 153–160. Cambridge, UK: McDonald Institute for Archaeological Research.
- Meillet, A. 1954. *La méthode comparative en linguistique historique*. Paris: Honoré Champion.
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient estimation of word representations in vector space." arXiv preprint arXiv: 1301.3781.
- James W. Minett and William S.-Y. Wang. 2003. On detecting borrowing: Distance-based and character-based approaches. *Diachronica*, 20(2):289–331.
- Saul B. Needleman and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.
- Ee-Lee Ng, Beatrice Chin, Alvin W. Yeo, and Bali Ranaivo-Malancon. 2010. Identification of Closely-Related Indigenous Languages: An Orthographic Approach. *Int. J. of Asian Lang. Proc.*, 20(2):43–62.
- Michael P. Oakes. 2000. Computer estimation of vocabulary in a protolanguage from word lists in four daughter languages. *Journal of Quantitative Linguistics*, 7:233–243.

- Pagel, M., Q. D. Atkinson & A. Meade. 2007 Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449(7163). 717–720.
- Pagel, Mark, Quentin D. Atkinson, Andreea S. Calude, and Andrew Meade. 2013. Ultraconserved words point to deep language ancestry across Eurasia. *Proceedings of the National Academy of Sciences*, 110(21):8471–8476.
- Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. "Glove: Global vectors for word representation." In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532-1543.
- Sanda Reinheimer Ripeanu. 2001. *Lingvistica Romanica: Lexic, Morfologie, Fonetica*. Ed. All. Bucuresti.
- Ross, M. & M. Durie. 1996. Introduction. In Mark Durie & Malcolm Ross (eds.), *The comparative method reviewed. Regularity and irregularity in language change*, 3–38. Oxford: OUP.
- Michel Simard, George F. Foster, and Pierre Isabelle. 1992. Using cognates to align sentences in bilingual corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 67–81.
- Swadesh, M. 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96(4). 452–463
- Uban, Ana, Alina Maria Ciobanu, and Liviu P. Dinu. 2019. "Studying Laws of Semantic Divergence across Languages using Cognate Sets." In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pp. 161-166.

3D Roma – Sarmizegetusa. Turn on the History. From the data collection on the field to the deployment of a Virtual Museum.¹

Antal E.¹, Bota C.¹, Ciongradi E.¹, D'Annibale E.²,
Demetrescu C.², Dima B.¹, Fanini D.², Ferdani²

(¹ MNIT National Museum of Transylvanian History)

(² CNR-ITABC National Research Council – Institute for
technologies applied to Cultural Heritage)

Abstract

This paper presents the first installation produced with the data collected on the ancient Roman city of Colonia Dacica Sarmizegetusa: it can be considered a concrete example of a full workflow, from photogrammetric 3D acquisition to gaming experience, able to contribute to the community of experts in the domain of virtual museum. The visualization of enormous archaeological contexts like a whole ancient city has been a test bed to develop tools and methodologies in order to create and maintain accurate and fully real-time enabled 3D models. In the temporary exhibition, open until 30 September 2016, a multimedia installation based on "natural interaction" solutions was set up: thanks to Kinect and Leap-Motion sensors, visitors can interact with virtual environments and objects, using gestures to experience a more engaging and intuitive experience.

¹ A short version of this paper was presented at the 14th EUROGRAPHICS Conference. Workshop on Graphics and Cultural Heritage, Genoa and published in Eurographics Digital Library. Antal et al. (2016), pp. 75-78.

Introduction

According to the last researches a Virtual Museum is "*...a digital entity that draws to the characteristics of a museum, in order to complement, enhance or argument the museum experience through personalization, interactivity and richness of content. [...] A virtual museum can refer to the site, mobile or World Wide Web offerings of traditional museums or can be born digital content such as net art, virtual reality and digital art* (Hazan et al., 2014, p. 39)."

Given that, a common issue in the production of a Virtual Museum is in the lack of a clean workflow able to connect real archaeological data, collected by experts in antiquity, and the deployment of the final installation. The cause of this situation is most likely the absence of a project that keeps in consideration from the very beginning a structured communication plan of the scientific discoveries and the lack of connection among the different research domains involved. Furthermore, on one side, the content of the archaeological record is often intended to offer only a description of the contexts with small or no attention at all to the visual environment of the site. On the other side, a digital archaeological record (3D point clouds from archaeological excavations, stratigraphic data etc.) requires time-consuming transformations and adaptations in order to make it usable in a virtual museum. A common example is a 3D acquisition where the optimization of the models is a key-point to enable effective virtual museums and real-time experiences.

On April 7th, 2016, the exhibition "3D Rome - Sarmizegetusa. Turn on the History" was inaugurated in the halls of the National Museum of Transylvanian History (MNIT) in Cluj, Romania (see Fig. 2). It was a MNIT and CNR-ITABC (VH-Lab - Virtual Heritage Laboratory) co-production, in



Figure 1: Real-time exploration of the excavation of the Termae (Domus Procuratoris).

collaboration with various partners, including the Trajan Markets - Museo dei Fori Imperiali (Rome) and MCDR. In the exhibition, we tried to overcome the issues mentioned above by connecting museum objects with their original environment together with virtual reconstruction in order to help visitors in understanding the archaeological sites. By means of the use



Figure 2: *Room of the exhibition with the installation running and the masks on the walls.*

of Exposition (defining and describing monument and artifacts to inform visitors and making them aware) and Narrative (providing information and interpretation about monuments and artifacts by arranging them in a sequence) communicative styles (Ferdani, Pagano and Farouk, 2014), we boosted the visitors tour with an immersive experience. In the exhibition, the visitor can explore, using their own bodies to walk through the remains of the grandiose Praetorium Procuratoris, with the termae being excavated (see Fig. 1), and the Great Temple (the largest temple of the entire Dacia province) with a first virtual reconstructive hypothesis in semi-transparent overlay. Next to the Sarmizegetusa virtual scenarios, visitors can explore two contexts of Ancient Rome - the Forum of Augustus and the Temple of

Peace - which ideally represent the formal origin and reference inspiration models of the monuments of the Roman provincial capital. Along the virtual tour, visitors can "pick" archaeological remains up by means of hand gestures and eventually "pass" the virtual object into a separate screen, where another visitor can manipulate it with simple hand gestures, always through natural interaction solutions. In this way, precious objects like the gilded bronze satyr or the Gorgona, the capital of the winged Pegasus and other jewels of the collections of the Museum of Cluj-Napoca, Museum of Deva and the Trajan Markets Museum can be further inspected in their original context of discovery. Also, some original objects that visitor can find in the application were set up in the exhibition room in order to make a strong connection between real and digital artifacts. The original archaeological finds were placed using an innovative concept behind plaster masks that imitate the ones from Pompeii.

The case of Virtual Sarmizegetusa in the last 3 years has been a perfect test bed for a wide spread of acquisition methodologies and techniques. A strong focus has been accorded to going forward the digitalization concept looking for a "digital replica" where geometries, colors, structure of the light, photos and equirectangular panoramas are collected at very high detail in order to make it possible a realistic VR experience. Moreover, the ancient city is located in the Retezat National Park and several samples of vegetation elements were collected in order to recreate the current natural environment (trees, bushes, grass). The Virtual Sarmizegetusa Project is part of a wider series of scientific actions on the ancient site of Sarmizegetusa that includes excavations (MNIT and UBB, MCDR, UNIEXE, UHEI, UVIE) and geophysics analysis (MNIT and CNR-ITABC).

The case of Colonia Dacica Sarmizegetusa

Colonia Dacica Sarmizegetusa was the first and only colony of veterans of the Roman province of Dacia. It was created immediately after the conquest of Dacia by Trajan in the years 108-110 A.D. and until Marcus Aurelius remained the only colony of the province. In the center of the site it has been identified and researched the Forum Vetus. Soldiers of Legio IIII Flavia Felix built the forum wooden construction phase and immediately afterwards the stone phase. They also participated in other assemblies colony construction: the city wall, Horrea near to Praetorium Procuratoris. Colonia Dacica Sarmizegetusa has a quasi-square shape inside the walls to start measuring

22.5 ha, and 33 ha and ha 60-70 Extra Muros. Throughout the first century AD, Sarmizegetusa was the main cultural center of the province, which contributed to the embellishment of numerous *euergetes*. Under Severus Alexander, it got the epithet of metropolis city, which is a testament to his prosperity. It is the only city in the Western Roman Empire receiving this epithet. The reconstruction of the stone amphitheater under Antoninus Pius, Forum Vetus reconstruction in marble after the Marcomanic wars indicates also a flourishing city. Sarmizegetusa was also *Concilium Trium Daciarum* headquarters, the altar of Rome and Augustus (Ciongradi 2007).

Virtual Sarmizegetusa Project: the digital acquisition

The creation of the Virtual Sarmizegetusa installation was conducive to the orchestration of different technologies and methodologies simultaneously. Tasks like digital acquisition on the field, inside museum collections and further optimizations of virtual environments and assets stimulated new issues and promoted new solutions, both from a technological point of view (creation of new software) and methodological one (fitting scientific documentation needs with current technologies).

Digital acquisition of the landscape and architectures

The acquisition procedure was carried out using a work-flow structured and validated within the VHLab of ITABC CNR, following several archaeological campaigns in Italy and abroad (Adami *et al.*, 2014). Preparatory work for the photogrammetric campaign was the creation of a topographic network including a series of survey nails and photogrammetric targets distributed evenly on the structures and gathered with a laser Leica total station. Given the large extension and complex articulation of the site, on the field it was necessary to operate within an absolute coordinated system. When necessary, new reference points have been acquired through a dual differential GPS and then added in the monographs of the known points. After the topographical campaign, two types of photographic acquisition were carried out. The former was a terrestrial one, aimed at obtaining high resolution models (before the introduction of the S-1000 UAV). The latter was an aerial survey of the site intended to collect data both for territorial scale (UAV Phantom 2) and high resolution architectonic scale (S-1000 UAV).

Instrument	Specifications	
Total station Leica Flexline TS 06+ 2" R1000	Sensor: EDM laser class 1 Max range: 3500m Angular accuracy: 0.5"	
D-GPS South S82V	Accuracy: Static and FastStatic GNSS surveying Horizontal: 3mm+0.5ppm RMS Vertical: 5mm+0.5ppm RMS	
Reflex Canon EOS650 18MPx	Sensor: Canon APS-C 18MPx Max Resolution: 5184x3456 px Focal lenght: 18-55 Sensor dimension: 22.3 × 14.9 mm	
Reflex Canon EOS6D 18MPx	Sensor: FullFrame CMOS 20 MPx Max Resolution: 5472x3648px Fixed focal lenghts: 14 and 35 mm Sensor dimension: 35.8 x 23.9mm	
Reflex Nikon D3S 18MPx	Sensor: CMOS 12.87MPx Max Resolution: 4256 x 2832 px Focal lenght: 14-35 Sensor dimension: 36.0 x 23.9 mm	
GoProHero4 12MPx	Sensor: CMOS 12.MPx Max Resolution: 4000x3000 px Focal lenght: 15mm Sensor dimension: 5.37 x 4.04 mm	
UAV S-1000	UAV: ITABC-02-S1000 Dji S-1000 Framework Universal Gimbal brushless.com Payload CANON 6D	
UAV Phantom 2	UAV: ITABC-01-Phantom2 Dji Phantom 2 Framework Gimbal brushless Payload GoProHero4	

Figure 3: *Specifications of sensors and tools used in the digital acquisition campaign.*

The photographic sampling is an essential feature and the quality of the post-processed 3D models depends mainly on the photo shooting campaign as well as on the accuracy of the topographic survey. Therefore, the land campaign was conducted using two full frame SLR cameras: Canon D6 with 14mm and 35mm optical lenses and a Nikon D3S with 14 and 24mm (see Fig. 3). The photos were taken at regular intervals, from different angles and with an overlap between pairs of pictures always higher than 70% in order to sample the most of the affected area. The acquisitions were performed in the best possible light condition according to season ie. during diffuse light conditions (cloudy day), or in the absence of cast shadows,

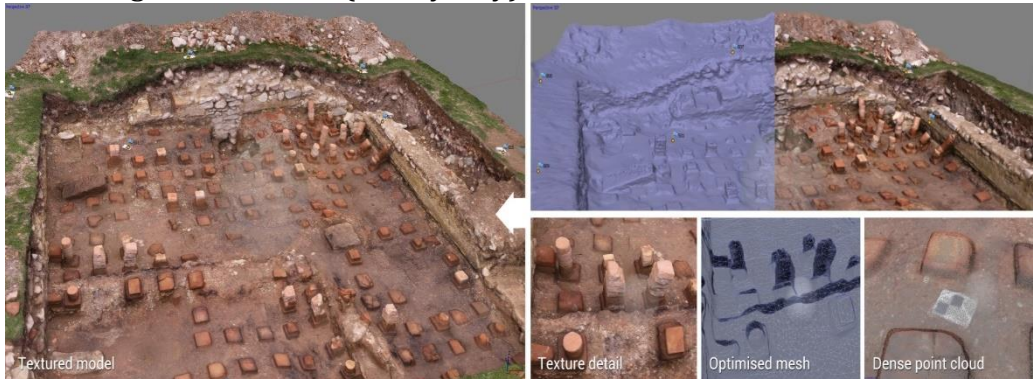


Figure 4: Image processing work-flow to obtain high resolution textured models.



Figure 5: Equirectangular 360 degree high resolution image from the Amphiteater

then concentrating the acquisition in the central hours of day. In the few cases the campaign encountered exposure or white balance issues the use of photographic RAW formats allowed a post production work (semi-automatic) for post-processing adjustment through the open source

software Darktable: this solution has made possible a homogeneous result and good quality for texture building and parameterization of 3D models.

The aerial survey was carried out with two systems, a quadcopter *Dji Phantom 2* with *GoPro Hero4* and an octocopter assembled on top of a *Dji S-1000* framework equipped with a *Canon 6D* on an universal gimbal (brushless.com). The two vehicles, equipped with remotely controlled gimbal, allow to take photos at different angles (from 90 to 45 degree) allowing the archaeological contexts to be easily surveyed. The entire dataset of images has been divided by groups corresponding to the names of individual monuments (see Fig. 10). For each group, photos were imported within *Agisoft PhotoScan*, a dense image-matching software, to perform the alignment of the pictures and create the photogrammetric model. In order to provide a geo-referenced alignment, targets were used as control points. Once obtained the alignment and checked the residual error, dense clouds and then the triangulated polygon mesh were calculated. The dense cloud was treated with special software (*Geomagic* and *CloudCompare*) in order to correct small errors due to matching problems or sub-samples (duplicate vertices, uneven density, etc.). The mesh reconstruction was first performed at the highest level of detail and subsequently has been reshaped (closing holes, elimination of non-manifold edges, etc.). The polygon mesh has finally been optimized with the use of

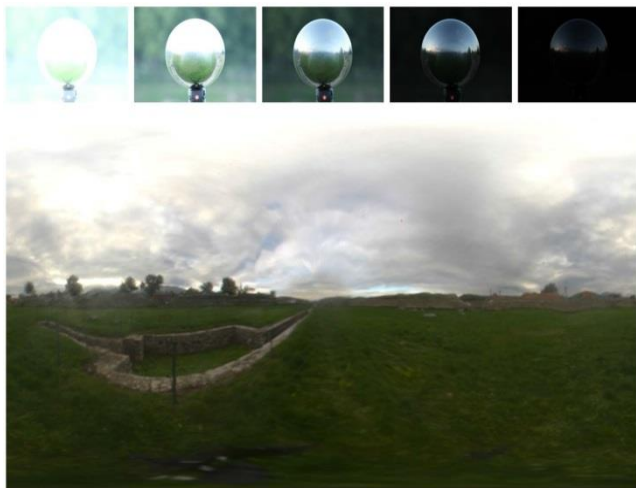


Figure 6: (Up) Photoset of a mirrored sphere used to create the light-probe (8 bit); (Down) 360° spherical high-dynamic range image (32 bit floating point).

special filters which allow to reduce the number of polygons retaining the level of detail of the parts with more complex geometry and architectural details. Once we concluded the optimization tasks, we moved to the parameterization of the mesh and the generation of high-resolution textures using texture building algorithms. The very high detail of the model (geometry and color

information) ensures different uses for documentation and communication purposes (see Fig. 4). Additional optimizations were performed on the textured model in Blender 3D software, where it was possible to fix several texture issues: a new visual tool has been developed as a Blender add-on (Python) in order to make possible a fully flexible pipeline for semi-automatic texture correction (<https://github.com/zalmoxes-laran/BlenderLandscape>).

A series of panoramic photographs (equirectangular projection) have been provided (see Fig. 5): the panoramic photography allows the acquisition of high quality views, both in terms of final image resolution and colorimetric consistency, to be obtained. The individual points of view were placed in a spatial sequence to create virtual paths that ensure an immediate

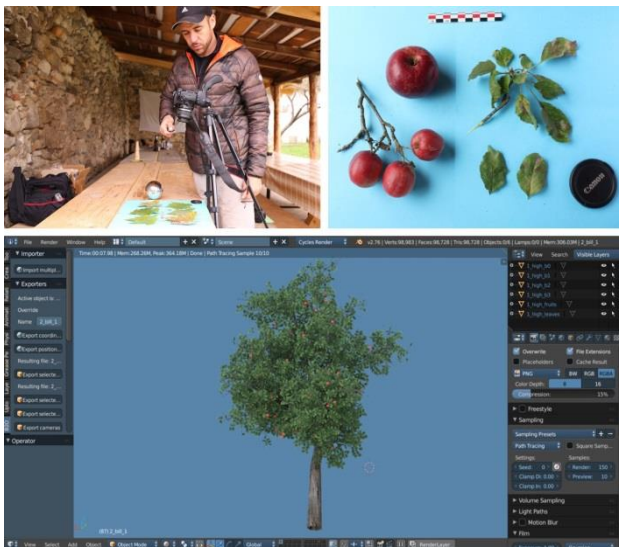


Figure 7: *Samples of vegetation elements to reconstruct the original natural environment.*

vision of the site, both for communication purposes and to prototype successive phases of acquisition.

In order to acquire light-color sample, few panoramic High Dynamic Range Images were generated using low-cost methodology (see Fig. 6). Multiple photos of a mirrored sphere were taken with different intensity levels (Low Dynamic Range Images). The latter were combined

by HDRShop v.1 software (<http://gl.ict.usc.edu/HDRShop/index.php>). The final 360° HDRI can be used in the virtual environment (see Fig. 8) taking advantage of global illumination rendering algorithms (Debevec 2008).

Finally, in order to reconstruct the actual environment and completing the reality based models and to place local plants in their original position, we also modeled the vegetation. Firstly, a library of the local flora has been gathered directly on the field sampling the most prevalent vegetal species: *Acacia*, *mallus Domestica*, *Juglans regia*, *Prunus* and *Vitis Vinifera*. Then, using a procedural generator software, we created the organic models of the plants. The vegetal landscape reconstruction improved significantly the



Figure 8: *Image Based Lighting (IBL) using the HDR image as a light source.*

realism of the virtual world and the sense of presence of the interactive application. (see Fig. 7).

The result was an extensive 3D model able to be used to make derivative archaeological and architectural documentation (orthorectified photos, color corrected images, plans and

sections) and to make 3D content for the virtual representation and simulation of the ancient landscape.

Digitization of museum artifacts

An important criterion in choosing the artifacts that the user has to find in the 3D application was the archaeological discovery context. The artifacts were discovered in the archaeological objective or closely to it, and the digital models of the artifacts were placed in the right context position. The original archaeological objects were taken from the MNIT and MCDR history collection and before it was set up in the exhibition the artifacts were digitized using photogrammetry technique. We tried to present in the exhibition a great variety of objects and contexts, showing prestige objects, architectural pieces but also some common artifacts. At Praetorium Procuratoris monument an inscription dedicated to the financial procurator Marcus Luceius Felix was placed in the Area Sacra which was decorated with statues and altars of a great variety of gods and goddesses, all dedicated by the governors; a brick from the *hypocaustum* installation was set in the place of the researches made in 2015 at the house of procurator and also a ceramic lamp (Alicu, Tarnavschi and Ruscu, 2006) were placed in the *thermae* of the private bath of the procurator. Inside of the so-called "Great Temple" which was in fact a sanctuary of the imperial cult, dedicated initially to Hercules - Commodus, to Diana - Crispina, and to Juno Sospita, the patron of the hometown of emperor Commodus, Lanuvium, were placed one fitting made of gilded bronze belonging to an imperial statue, with the head of Gorgona Medusa (Alicu, Pop and Wollmann, 1979, p.124), a statuette of gilded bronze is depicting the god Dionysos - Bacchus as a child

(Alicu, Pop and Wollmann, 1979, p. 85) and a Corinthian capital (Bota 2003) from a column that was in front of the cult chapel. As a reward for the visitor of the exhibition, we also placed in the exhibition two exquisite artifacts of a gilded bronze fitting of a horse rider (Alicu, Pop and Wollmann, 1979, p.185) and a bronze statuette of vegetation god, Pan (Alicu, Pop and Wollmann, 1979, p. 85), (Floca 1967, p. 47).



Figure 9: Complete set of sensors and tools used in the digital acquisition campaign..

Optimization of models and multi-resolution for virtual environments

When dealing with the representation and interactive visualization of large environments, a common approach is the segmentation of 3D assets and hierarchical out-of-core organizations (Borgeat *et al.*, 2005) for multi-resolution. When properly organized, such structures allow fully explorable and theoretically resolution "limitless" environments. Creating a gaming experience for a Virtual Museum on the subject of a great capital of ancient world (33 ha) fits exactly this scenario. All the models have been optimized (geometries and textures) using multi-resolution approach and according to a given spatial and hierarchical structure, designed on top of the blueprint of the ancient city. The models have been divided into reality based models (unique objects from photogrammetry) and sample-based models (models of type object reused as instanced assets with small parametric-driven differences) like in the case of trees, rocks, sheafs etc. Optimizations have been performed in computer graphic software Blender 3D with manual tools and ad-hoc scripts (see Blender Landscape) and using new dissemination online Front-End, applied to the gaming experience. In order to deploy a smooth workflow and methodology during ingestion phases from 3D modeled objects and environments to the real-time visualization

framework, a set of open-source desktop tools has been developed. Targeting this context, real-time object painting and scene dressing tools, alongside specific optimization tools have been implemented with drag&drop approach. The latter led to small mini-processing units ("*droplets*") that combined with preview tools, allowed fast ingestion pipeline even under strong time constraints. A similar approach was used to process and publish online a set of 3D objects and archaeological sites related to the exhibit: we employed a new Front-End (BLIND reference) with *out-of-core* multi-resolution support, targeting all modern browsers supporting WebGL technology and also mobile devices. The latter is a crucial requirement in this case to connect fruition with the gaming experience (see next section) and the multi-touch features offered to interact with the 3D model. Especially appealing in this context is the capability of the Front-End to present rich and specialized multimedia annotations (images, audio, videos) that well fit presentation requirements, such as multi-language audio and text (English and Romanian) as well as reference photos of original items in the museum (see Figure 11).

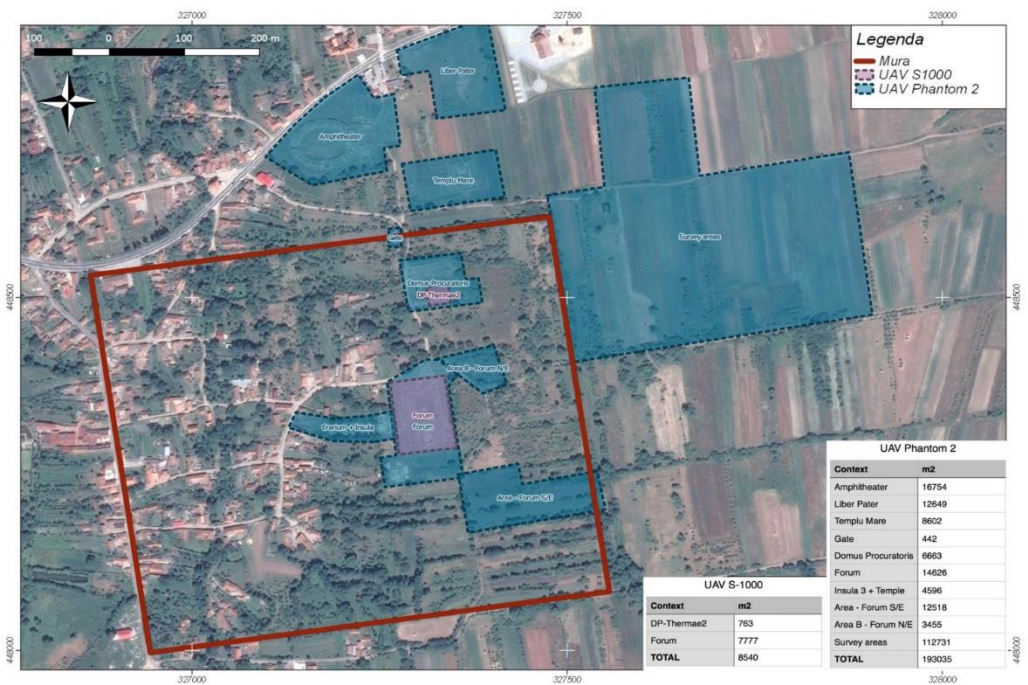


Figure 10: Plan of the areas of the photogrammetric acquisition campaigns.

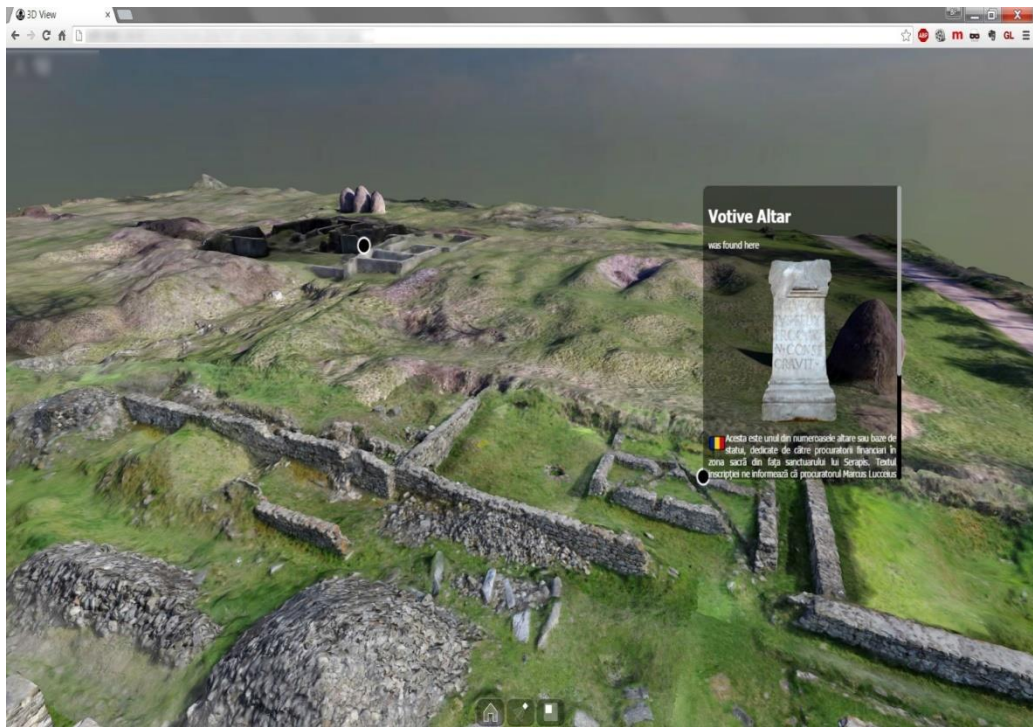


Figure 11: *Interactive online dissemination of two multi-resolution archaeological sites with rich multimedia annotations, including two languages (English and Romanian) used for audio-descriptions of findings.*

Natural interaction in a museum

When we deal with virtual museum applications, the potential of interaction design patterns and deployment of enhanced experiences embodied into a physical installation, can have a great impact in terms of engagement and attraction (Fannini and Pagano 2015).

Recent technologies for gesture-based interaction such as the Kinect sensor and the Leap Motion controller opened great and fascinating possibilities within the Cultural Heritage field. The availability on consumer

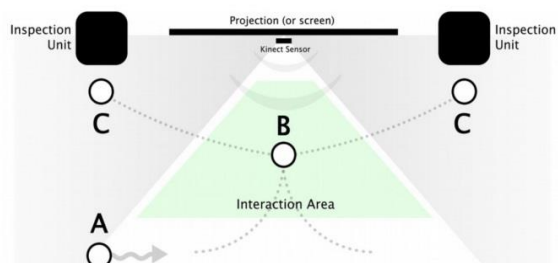


Figure 12: *Setup of physical spaces for dual gesture-based interaction.*

market and the low-cost aspects of such sensors, make them very suitable for a museum engaging space. In order to deploy gesture-based applications, the very basic requirement consists in a physical 3D space where incoming visitors can perform a pre-defined set of gestures. The size of the space may vary, depending on application flexibility and limitations of the sensor: for instance a Kinect requires a minimum of about 2.5 meters, while the Leap Motion controller can sit on a table covering a vertical interaction area of about 60 cm. The software application must be capable of recognizing such physical movements performed in mid-air observed by the sensor, and translate them into specific actions. Clearly, specific sensor limitations must be taken into account at early stages of application design: for instance, accuracy and occlusion play important roles into the creation of efficient 3D user interfaces and interaction models. The main objectives of such gesture-based applications is in fact to increase user engagement, for instance within serious games, through natural mappings from the 3D physical space museum to a 3D virtual space. In this section, we describe the resulting installation for the exhibition "*3D Rome - Sarmizegetusa, turn on the History*", inaugurated in the halls of the National Museum of Transylvanian History (MNIT) in Cluj (Romania). The installation is based on a previous dual natural interaction application (Fanini *et al.*, 2015), presented in Granada expo in occasion of Digital Heritage 2015 event and award-winner of "*Best Exposition - Quality of Content*". The main concept behind the whole application is to create a shared and collaborative gesture-based experience, by taking advantage of two separate natural interaction units. The first one is a serious game allowing a visitor to explore several 3D virtual environments using body gestures to find and collect virtual objects, typically on display in the same museum, like in the case of "Keys2Rome" exhibit (<http://keys2rome.eu/>). The application follows the treasure hunt game mechanics in order to create an engaging educational context for incoming visitors. When an object is collected, a short audio-clip describes the item, finds location and other valuable information connected. Once all the items in a given scenario are collected, the user is rewarded with additional audio descriptions of current environment, historical events and a QR-code that unlocks additional online content using the previously described visualization Front-End. Within the Cluj exhibit, four scenarios have been configured: (A) two large multi-resolution virtual environments of Sarmizegetusa, available from the start of the experience, and (B) two Fora

reconstructions (Forum Pacis and Forum Augusti), unlocked for current player only upon completing the previous row.

The second natural interaction application is an inspection unit, acting conceptually as a "magnifier", where users can manipulate a virtual item by means of basic hands gestures (*see* Fig 13).

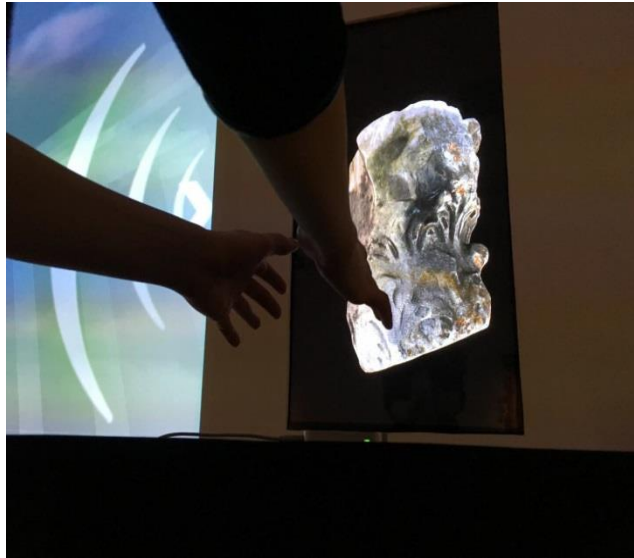


Figure 13: *Leap motion based interaction in the inspection unit application.*

The inspection unit aims at expanding the terms of 3D content fruition by users. The spectator, who passively takes part at the show, now becomes active user thanks to the chance of interacting with the content with his hand gestures. The realization of the whole system involves the design of three interconnected components that aim to produce a modular, portable and adaptable solution for cultural institutions. Specifically, the inspection unit was adapted for minimal hardware requirements: PC, Monitor and Leap Motion. The installation has the hub of innovation in the software, developed using vvvv (<http://vvvv.org/>), a hybrid graphical/textual-programming environment for easy prototyping and development. The environment is dedicated to the management of multimedia content in real time extremely flexible. Of particular interest is the ability to manage 3D contents and multi high resolution video output. Moreover it's compatible with many interaction devices and peripherals (MIDI, Leap Motion, Kinect,). The software includes modules for:

- 3D Model and Texture Ingestion
- Leap Motion Handling
- Communication with main game application
- Real Time Video Effects

The version used in this project provide a mixed input to manage the 3D items, in order to create a sort of "*magnifier*" unit, where users can visualize and manipulate items, eventually incoming from external applications. Virtual items that appear in the screen can be manipulated and inspected by the visitor using the Leap Motion controller. Two gesture interactions with the 3D model are proposed: one-handed manipulation to translate and rotate and two-handed manipulation to translate, rotate and scale.

Integration of virtual and real Museum: the exhibition

The exhibition concept is designed like a 3D video game where the "time traveller" is a treasure hunter. For a better understating of the exhibition, it was set an "intro" room where are presented some reconstructions drawings of the most important monuments from Colonia Dacica Sarmizegetusa archaeological site, some of them being present also in the 3D application. The reconstructions drawings talk about Forum Vetus with its construction phases (wooden, Trajanic stone phase, and Antonin middle and late phase in which the city is marbleized), Forum Novum with its Capitoline Temple, the amphitheater with its phases of construction, the amphitheater baths and also a model of a roman home researched at Sarmizegetusa site. The uniqueness of the exhibition derives from the fact that restored digital 3D monuments and artifacts are combined in the main room with the presentation of real artifacts, discovered during excavations at Colonia Dacica Sarmizegetusa site by Romanian archaeologists. Some of them are located in the exhibition hall inside illuminated niches refined and it can be admired through the mouth of four masks, replicas of the originals from Pompeii. The masks were manually built in plaster using "ronde-bosse" technique after some Greek theatre type masks that were made of marble or terracotta that decorated the painted walls of the roman houses. Behind the masks, the walls are decorated with beautiful posters that imitate the frescoes from Pompeii. The bronze pieces from the archaeological site of Sarmizegetusa are hidden behind the wall and near the masks, there is set up a QR code that offers the visitor the opportunity to "*take*" home the



Figure 14: *Bronze piece of the Gorgona behind the Mask and the visualization of the corresponding 3D online model.*

artifact by visiting a WebGL page with the 3D model and to interact with it using full multi-touch manipulation (pinch to zoom, rotate and pan) using the previously introduced Front-End (see Fig. 14). The interactive visualization also allows to read more information about the real artifact. The rest of the room was populated also with other real artifacts as a marble capital and votive altar, and also a hypocaustum brick. In the corners of the room there were placed two plaster copies of Niobides statutes from Rome in order to illustrate the connection between the two roman capitals from the 3D application. On the front wall, there was built the screen and the monitors for the digital part of the exhibition.

Conclusions and future works

Expected results of this procedure are rich connections between the site, the local museum and artifacts collection. Some typical contexts for the Roman times like *domus*, city gates, furniture that are present in the museum have been provisioned to the visitors in order to enable a deeper understanding of the ancient town. A first result of our research activities is the digitization and documentation of a unique archaeological site and the creation of new ways to communicate and perceive the cultural heritage treasure of an emerging European touristic country like Romania. New

technologies like UAV systems, Dense Image Matching, Natural Interaction, WebGL dissemination and Virtual Reality were used in order to empower at the same time both research and preservation through high resolution 3D models and, on the other hand, the dissemination and provisioning of new tourist attractions. Within two years there will be further exposure with the next step of the project, in which the current city will be fully explorable. We would like to achieve a virtual reconstruction of the city and the most outstanding buildings that have represented the core of the roman life during the 2nd and 3rd century and arrange the digitally restored objects inside them. The final goal of the virtual reconstruction is the development of a series of on-site multimedia connected applications to improve the quality of the visit, disseminate information in a more effective way and allow a better understanding of the archaeological site and art objects to be experienced. Indeed the museums domain is rapidly changing (Pescarin 2014, p. 132-134): Sarmizegetusa will be a perfect cultural district to test and deploy this kind of next generation virtual museums. Museums need in fact to offer constantly new, entertaining and scientifically validated exhibits, providing technological solutions with coherent visual information and above all stories, interaction and comprehensible relations between displayed objects and their contexts.

Bibliography

- Adami *et al.*, 2014) - Adami, A., Cerato, I., D'Annibale, E., Demetrescu, E., Ferdani, D. (2014) *Different Photogrammetric Approaches to 3d Survey of the Mausoleum of Romulus in Rome*. In *Eurographics Workshop on Graphics and Cultural Heritage*, The Eurographics Association, pp. 19–28.
- D., Pop C. and Wollmann V. (1979) *Figured Monuments from Ulpia Traiana Sarmizegetusa*, vol. 55. British Archaeological Reports.
- Alicu, D., Tarnavski, M. and Ruscu, L. C. (2006) *Die römischen Lampen von Sarmizegetusa*. Porolissum.
- Borgeat, L., Godin, G., Blais, F., Massicotte, P. and Lahanier, C. (2005): *Gold: interactive display of huge colored and textured models*. *ACM Transactions on Graphics (TOG)-Proceedings of ACM SIG-GRAPH 2005* 24, 3 (2005), 869–877.
- Bota, E. (2003) *Capitelul corintic în Dacia Intracarpatică* . PhD Thesis, Cluj-Napoca.

- Ciongradi, C. (2007) *Grabmonument und sozialer Status in Ober-dakien*. Cluj-Napoca: Mega Verlag
- Debevec, P. *Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography*. In *ACM SIGGRAPH 2008 classes*, ACM, p. 32.
- Fanini, B., D'Annibale, E., Demetrescu E., Ferdani, D., Pagano, A. *Engaging and shared gesture-based interaction for museums the case study of k2r international expo in Rome*. In *2015 Digital Heritage*, vol. 1, IEEE, pp. 263–270.
- Floca, O. (1967) *Muzeul de arheologie Ulpia Traiana, Sarmizegetusa*. Muzeul Regional Hunedoara.
- Fanini B., Pagano A.: *Interface design for serious game visual strategies the case study of "imago bononiae"*. In *2015 Digital Heritage*, vol. 2, IEEE, pp. 623–626.
- Ferdani, D., Pagano A. and Farouk, M. (2014) *Terminology, Definitions and Types for Virtual Museums*. Deliverable report 2.1c, CNR, CULTNAT.
- Hazan, S., Hermon, S., Turra, R., Pedrazzi, G., Franchi, M. and Wallergard, M. (2014) *Theory Design*. Deliverable Report 3.1, CREF-Cyl, Available at: http://www.v-must.net/sites/default/files/D3.1_update.pdf.
- Pescarin, S. (2014) *Museums and virtual museums in Europe: reaching expectations*. In *SCIRES-IT-SCientific RESearch and Information Technology 4*, 1, pp. 131–140.

Heritage at the Crossroads. Digitization of Stone Crosses in Prahova County, Romania

Marius Streinu, Oana Borlean, Mihaela Buruiana,
Liviu Mihail Iancu, Bogdan Șandric

(INP – National Heritage Institute)

The results of this study are based on the activities carried out within two distinct administrative projects, but essentially complementary and homogeneous: “Heritage at the crossroads. Digitization of stone crosses and monuments of heroes from Prahova County” and “Heritage at the crossroads. Digitization of stone crosses, hero monuments and historical landscapes in the north and east of Prahova County”. The first project took place in 2017, the second took place in 2018, under the auspices of the European Year of Cultural Heritage and the Romanian Centenary of the Great Union. In the following, we will consider both aforementioned projects as two campaigns of a single development, given the common goal and objectives.¹

The project benefited from the co-financing of the National Cultural Fund Administration and was implemented by three institutions, the Eurocentrica Association as an applicant, the National Institute of Heritage, and the Prahova County Directorate for Culture, as partners. The three institutions have also benefited from the support of a considerable number of county schools², local public administrations, the Prahova County Museum of

¹ Iancu, Șandric 2018, 6.

² Of which it is worth noting as partners: the Theoretical High School of Azuga, the Secondary School of Boldești-Grădiște, Ion Kalinderu High School of Bușteni, the Technological High School of Cerașu, the Secondary School of Gherghița, Grigore Tocilescu Theoretical High School of Mizil, the Secondary School of Olari, Jean Monnet High School of Ploiești.

History and Archaeology and the National Archives - Prahova branch, cultural associations³, but also local historians⁴ and many locals offering insights.

The two projects addressed the need to safeguard special categories of immovable cultural heritage, ignored and considerably degraded: firstly, stone memorial crosses, but also monuments honoring the memory of heroes fallen in World War I and the historical landscapes associated with this conflagration. The causes that have led to the degradation and sometimes destruction of these types of cultural heritage are many, but the most important are the gaps in legislation, the lack of knowledge regarding the meaning and the importance of local heritage protection, and loss of direct connections to the commemorated events and persons. On another hand, the degradation, the destruction, and in some cases even the disappearance of some monuments suggests a lack of interest in monuments and places of national and local memory, as well as of some monuments that by their form and/or significance denote a religious character.

From a legislative point of view, the Law of Historical Monuments no. 422 of 2000 establishes a protection regime only for monuments classified in the List of Historical Monuments (LHM). In order to be registered in this list, the cultural objectives of the immovable cultural heritage (archaeological sites, places of worship, public monuments, cemeteries, crosses, etc.) must go through a strict classification procedure, at the end of which The National Commission of Historical Monuments subordinated to the Ministry of Culture approves favorably the request for classification, and the Minister of Culture issues an order that will enter into effect after publication in the Official Journal. The process is long and tedious, being under the authority of a bureaucratic system that rather discourages the start of any such initiative. At present, the legislation regarding immovable cultural heritage should be amended to provide for the protection of all cultural objectives, including those not yet included in the list of historical monuments. One solution would be to establish criteria, such as age, to prevent modification, relocation, or demolition before carrying out historical studies attesting to the cultural value of the objective and its preservation, even if it will not be classified in the LHM.⁵ Under these

³ The Prahova branch of „Queen Mary” National Association for the Heroes’ Cult, The Cultural Foundation „Traian Tr. Ceptiu”, The Association European Convergencies.

⁴ Gheorghe Bilgă and Traian Ceptiu.

⁵ Șandric, Borlean, Buruiana, Streinu 2018, 16.

conditions, it was necessary for the activities in the aforementioned projects to become a complement to the real image of the cultural heritage in Prahova County and not just the one registered in the List of Historical Monuments.

In this paper, we will tackle the current status of stone crosses, the most affected category of immovable cultural heritage among those that have been the subject of our activities. Therefore, the main objective of the projects was to map, document, and disseminate information in this special category of cultural heritage. The aim of our approach was to make an exhaustive record, as much as possible, of this historical and commemorative, epigraphic, and religious heritage that we could transmit to local authorities and especially to the communities that own it, both through online publication and through promotion by means of events organized in the given region. At the same time, the second goal of our approach of documenting and registering the cultural heritage in question was its inclusion in urban plans, cadastral and topographic plans, and in the agricultural register, aiming that this will ensure greater protection in the future.⁶ A very important aspect pursued in both projects was the use of new information technologies for the conservation and promotion of cultural heritage. But, the most important aim of both projects, in our opinion, was to raise awareness among young people about the importance of preserving these monuments as a repository of local history and local cultural identity.

The activities of the two projects consisted of documenting the stone crosses, creating a database, a website, an interactive map, and organizing educational workshops. Before starting the fieldwork to identify, document, digitize and map monuments, we studied older records, such as monument files from the archives of the National Institute of Heritage, the *Historical Monuments Directorate* fund (Fig. 1 and 2), and the Topographic Map of Romania, 2nd edition, scale 1:25.000 (Fig. 3). This endeavor led us to rediscover a series of monuments inventoried in the second half of the twentieth century, from the above-mentioned archive, as well as the identification of a considerable number graphically represented on the topographic map (Fig. 4). The field surveys resulted in the identification, documentation, digitization, and mapping of 269 stone crosses, from 84 localities, distributed in the central and eastern region of the county. The largest clusters of crosses were identified in the regions of Gherghița - Drăgănești, Cioranii de Jos, Gura Vadului - Vadu Săpat, Călugăreni, Jugureni,

⁶ Șandric, Borlean, Buruiana, Streinu 2018, 16.

Lapoș, and Mireșu, especially along the main roads and rarely in the surrounding villages.⁷

The stone crosses, votive or funerary, are monuments of memory, fulfilling the role of instruments through which the commemoration is realized, especially at a local level.⁸ Their importance for the local communities is indisputable since they are found in the whole southern area of the Carpathians. The stone votive crosses have been used with a second attribute: as the spatial demarcation of rural estates. The commemorative ones have the role of mentioning those who contributed financially to raising it, but also of their extended family, as sometimes happens. They are, in some cases, erected for the purpose of marking a particular deed or event and are located in places with high visibility, such as crossings, along important roads, near fountains and bridges, or in closely related places to the commemorated event. The crosses that have the role of spatial demarcation of rural estates are known as border crosses, being placed on the borders they mark.⁹ The monuments are shaped like a cross aiming at involving the divinity in the act of remembrance or spatial delimitation and at the same time seeking protection against any possible acts of destruction. On the other hand, the existence of simple commemorative pillars, documented in Argeș County, proves that the custom of erecting monuments that serve the commemorative function was deeply rooted and went beyond the simple dimension of Christianity.¹⁰

In the case of Prahova County, the number of crosses dating in the 16th-17th centuries is reduced by comparison to those dating to the 18th-19th centuries, the latter representing the majority of monuments. From the point of view of the spatial distribution of the crosses, they are in close proximity to the areas of exploitation of the raw material, limestone, and sandstone. In Prahova county, crosses are found in large numbers in the area of Istriței Hills – highest concentration, Gura Vadului, Vadu Săpat, Jugureni, and Călugăreni communes, all clusters found in the proximity of stone quarries.¹¹

Beyond their function, the crosses are typologically classified by their shape and structure.¹² Furthermore, their chronology, beginning with the 16th

⁷ Șandric 2018, 35.

⁸ Iancu 2018, 8.

⁹ Iancu 2018, 8.

¹⁰ Iancu 2018, 8-9.

¹¹ Iancu 2018, 9.

¹² Following the classification of Calinic Argeseanul-Constantinescu 1999, 80; Iancu 2018, 9.

century and until the middle of the 20th century, shows a continuity that also entailed a transformation of style and form. In the researched area, most of the crosses are represented as Latin crosses (Fig. 5), but starting with the 19th century, Celtic cross variants appear and co-exist with the Latin ones, by inserting a disc between the arms of the cross (Fig. 6). The typological diversity of the crosses is not limited to the Latin and Celtic ones, another group is known as *crosses with cubs* – an ensemble (Fig. 7). This last group consists of three Latin crosses, one large, placed in the middle and two small but equal in size, which flank the large one. Some crosses, in addition to the body itself, have a chapter and a pedestal. The chapters have different shapes and sizes with a decorative role. Another feature is that they were made to be removable, which led to the disappearance of most of them.¹³

There is a standard in terms of artistic design and decorations of stone crosses from the 17th century to the first half of the 19th century. Thus, in the upper part are represented between four and nine medallions arranged around a large, central one. Within are depictions of the monograms of the evangelists, MT = Matthew, MR = Mark, LUC = Luke, and IO = John, the tetragram of victory of life over death IS-HS-NI-KA, circular inscriptions, rosettes, or stars. In most cases, the body of the crosses is decorated with depictions of various plants (different types of flowers, meanders) and geometric motifs – rhombuses (Fig. 8-9). Starting with the 19th century, the sculptural decorations become scarce, being replaced by the painted decorations that take over most of the surface of the crosses (Fig. 10). The paintings, naïve in style, consist of figurative and anthropomorphic decorations and in the latter case, the depictions include the Mother of God, Jesus, archangels, as well as numerous saints. The inscriptions on the crosses undergo the same transformation. Whereas the lettering of the inscriptions was originally excised, it is replaced by incised letters, in some cases scribbled, and by painted inscriptions. This development is probably related to the erection of crosses becoming a mass phenomenon together with its popularization in the 19th century and early 20th century. This was the time when wealthier rural families could afford the costs of ordering and placing such a monument, although much simpler than earlier variants, which in the past was only accessible for the upper classes.¹⁴ Thus, the study of stone crosses proved to be a useful tool for finding new insights into the

¹³ Aldea 1969, 27; Iancu 2018, 9-10.

¹⁴ Iancu 2018, 10.

local history and daily life of the inhabitants of the region. The historical importance of the stone crosses transcends the traditional directions of research and opens new possibilities in being able to contribute not only to a better understanding of the social event, topographical, genealogical aspects but also to complex studies of linguistics, comparative regional onomastics, etc. For example, a potential future study of crosses in the commune of Călugăreni, where one of the largest clusters was identified with the majority of monuments very well preserved, will most likely lead to establishing interesting local genealogies for the 18th - 20th centuries, at least concerning one of the wealthiest families in the area.¹⁵

The field surveys also contributed to highlighting the phenomenon of *forgetting* old meanings and symbols in the local collective memory. To that end, interviewing the locals about the monuments revealed that most of them were aware of their existence in a particular place and that they were *always there, erected by the elders*.¹⁶ There is only a rather confusing picture of their functionality, that of remembrance and sometimes border marking. More often, the confusion increases among the locals as there is a tendency to attribute to the crosses much greater antiquity than the real one and to exaggerate their historical importance, by associating them with famous characters from Romanian history or with legendary characters.¹⁷

Despite the lack of mnemonic continuity between the time of the raising of the crosses and the present, they can still become important places in the memory of the local community. This can be accomplished through an educated process of recovery, consisting of research, identification of historical and artistic significance, dissemination of results to the locals - including during school classes, followed by the introduction of the monuments in the daily cycle of the life of the local community and including it in the making of a local brand.¹⁸

The project entailed from the outset the use of digital solutions and techniques for recording data on stone crosses and heroes' monuments and later for their dissemination. In fact, one of the needs we were trying to meet is defined in the European Commission's Recommendation on the digitization of cultural material and its preservation (2011)¹⁹, which

¹⁵ Iancu 2018, 10-11.

¹⁶ Crețeanu 1969, 130 n. 7.

¹⁷ Iancu 2018, 11.

¹⁸ Iancu 2018, 12.

¹⁹ EUR-Lex (2020).

explicitly encourages the digitization of books, journals, newspapers, photographs, museum objects, archival documents, audiovisual material, monuments and sites in the idea of safeguarding cultural material and preserving cultural memory. This Recommendation has been published in the Digital Agenda for Europe, an approach of the European Commission that aims to optimize the benefits of information technology for growth, jobs, and quality of life in the European Union, thus being part of the Strategy for Development - Europe 2020. The project enriched Romania's contribution to the digital library of Europe, (www.europeana.eu), which was to add up to the end of 2016 approx. 800,000 objects, and so far, has provided only approx. 147,967 objects.²⁰

A major component of the documentation activity was digital photography. Photography was used for capturing and recording the landscape in which the stone crosses are placed as well as all the details of the monument. For each capture, at least two cameras were used to record at high resolutions. The resulting electronic files were saved in raw, unprocessed formats, which allowed a large number of subsequent interventions to adjust brightness and contrast in order to offer the best rendering of the monument and its surroundings (Fig. 11). Another aspect was the use of photogrammetry, a technique that is based on photographing each component part of the monument twice, and the two photographs overlapping at least 70%. Subsequently, by using a specialized computer program, the photographs recompose the photographed object in three dimensions, in digital format (Fig. 12). The use of this technique is conditioned by the surroundings of the monument, which should be clear for at least 10 meters from any other body so that the camera can move freely around it. The three-dimensional reconstruction of the landscape was done using a drone for photogrammetry, equipped with a high-resolution camera. Certain conditions have to be met in this case as well, such as the location in a clear space and favorable weather.

In order to reveal and record the inscriptions on the body of the stone crosses, we used Reflectance Transformation Imaging (RTI). This technique roughly translates into the transformation of the degree of reflection of an image and is based on computational shooting methods that capture the shape and color of an object's surface, then allowing interactive illumination of the object in any direction. RTI also allows the mathematical

²⁰ Șandric, Borlean, Buruiană, Streinu 2018, 17.

improvement of the shape and color attributes of the object. The amplification functions of the technique reveal surface information that is not observed in the direct empirical examination of the physical object.²¹ Using this method allowed for a better reading of the inscriptions and their preservation in digital form (Fig. 13).

The use of digital techniques was not limited to documentation but also involved the management of the data obtained through documentation, as well as in the process of dissemination. The analytical record for each cross was registered in a database for quick retrieval but also for a viable record system and from where they were transferred to a secondary database that was uploaded online through a website. The technical specifications of the database will also allow in the online environment a quick search of the information searched for by users, either using one of the main criteria, such as administrative location or using the search function in the database by a keyword. The database was also correlated to the spatial data of the location of the stone crosses, data that were obtained by recording geographical coordinates with the GPS. Thus, the content of the analytical records has become attributes of spatial (geographical) positions that can be accessed through a GIS web application (online interactive map). A cartographic tool for disseminating the monuments was also created as part of the desired objectives of our project. This application also allows the reuse of spatial and documentary data by anyone who can upload them in their own map application or in studies or administrative papers such as urban plans or property plans (Fig. 14). Digitization methods and techniques are thus the only viable solution for sustainable documentation and conservation of cultural heritage, as evidenced by saddening events all around the world. Only through high-resolution images, three-dimensional digital models and databases can cultural resources survive to be known and studied by future generations, whether they are at risk of natural extinction, degradation, acts of vandalism, or military conflict. Any effort to conserve immovable heritage cannot have the expected result, regardless of the amount of resources invested by central authorities and relevant non-governmental organizations, without the formation of favorable attitudes and behaviors aimed at its protection as well capitalization for the general public, local communities and of economic agents. For this reason, actions to promote immovable heritage as a resource with significant civic, artistic and

²¹ Cultural Heritage Imaging (CHI) 2020.

economic impact for local communities, to educate the general public to protect and capitalize on heritage, are necessary investments, with low costs, but with significant benefits in the future. From this premise, we directed a significant part of the efforts in the promotion of immovable heritage and education for its protection and judicious use, relying on several strengths: the interest of the young generation in new technologies and their practical applications; the interest of local authorities in building attractive local brands; the novelty of the heritage objectives subject to digitization and of the actions necessary for digitization (fieldwork to locations with special landscapes; drone flights, etc.) and the great extent of dissemination of messages through the Internet and social media.²²

The concrete actions of promotion and education consisted of, firstly, the creation and promotion of the website www.monumentelarascruce.ro which contains the complex database with all the monuments and landscapes discovered during field surveys and the 2D and 3D graphic materials. Its online accessibility and friendly interface for ordinary users and those in academia make this platform easy to use in areas such as education, tourism, research, economics, etc., to the benefit of local communities. The site is a flexible tool, designed to allow a future expansion, but also thematic, to incorporate other elements of heritage, with a commemorative role (eg crucifixes, memorial plaques, etc.).²³ The site consists of several sections:

- a. the catalog of all documented and digitized monuments, through which users can access, based on geographical criteria, the monuments they are interested in, for viewing 2D and 3D models and consulting records;
- b. the interactive map, drawn up on the basis of GPS coordinates collected in the field, showing the territorial distribution of monuments, and also allowing access to information and models for each of the objectives of interest;
- c. the gallery of monuments and landscapes;
- d. aspects from during the inventory, documentation, and digitization campaigns of heritage objects.

Secondly, the promotion of the activities and results of the project through media and social media aroused and actively maintained the

²² Șandric, Borlean, Buruiana, Streinu 2018, 18-19.

²³ Șandric *et alii* 2018, 20.

interest of scholars and, especially, of the general public. The main media channels used in the promotion activity were the publications of the Eurocentrica Association - www.lapunkt.ro and www.europunkt.ro, to which others were added, such as Radio Romania Cultural, Timp Românesc, Radio Deea etc. A special role in the promotion action was assigned to the European Convergence Association, which ensured the promotion of projects in the historical communities of Romanians abroad, so that articles about the project appeared in publications such as TocPress, BucPress, Foaie Națională, etc. The project was promoted on social media through the partners' Facebook pages. Posts about the research team's fieldwork, information gathered from stone crosses and hero monuments, their artistic characteristics, and geographic location, and the use of the hashtag #vanatoriidecruci imprinted a dynamic character on promoting the project with an impact far beyond expectations on the visibility of the project.

Thirdly, workshops and thematic excursions were organized with about 500 students from several schools in Prahova County in order to learn about stone crosses, heroes' monuments, and historical landscapes from the project team and local historians. During these events, the schoolchildren entered competitions on local history, created drawings about the commemorative heritage in a collaborative manner, in teams, participated in the registration of stone crosses and monuments of heroes (by filling in records, taking photos, taking GPS coordinates)(Fig. 15).

Last but not least, conferences presenting the results of the project, recorded following the field campaigns of 2017 and 2018, were designed both as endpoints, where researchers, heritage specialists, teachers, representatives of local authorities, pupils, and students could learn about the project and provided very important feedback to the partners, especially as openers of new directions in the field of conservation and enhancement of stone votive crosses, heroes' monuments and historical landscapes. The conferences were similar to debates about what should be done and what can be done for the heritage at the crossroads (Fig. 16 - 17).²⁴

Following the two campaigns for identification, documentation and digital preservation, we were able to draw some conclusions about this type of special heritage, the stone crosses. The cross-type monuments are in a decent state of preservation, being overwhelmingly affected by the lime covering, which is done periodically, especially around Easter, by the locals.

²⁴ Șandric *et alii* 2018, 21.

Unfortunately, this custom led to deteriorations, such as the covering of the original inscriptions, representations, and paintings and sometimes wears of the stone and produced breakage or cracks. Also, another risk to which stone crosses are subjected is the inclination following the collapse of the pit where the base of the cross was installed, either due to intensive plowing or soil movement. The crosses from the 15th century are especially in danger from this risk. Crosses from the 17th -18th centuries are also exposed to it because they are both massive and tall. Unfortunately, after their collapse, these monuments break, and the fragments spread over a larger area or sink into the ground, becoming unrecoverable. Other threats include agriculture - for those located on agricultural land because they are frequently hit by plow blades or simply demolished and removed from the land as well as human habitation - in the case of those in the yards of locals, who sometimes use the space as a waste deposit.²⁵

The general conclusion on the perception of the locals regarding the stone crosses is more optimistic. The survival of such a large number of crosses, the interest shown towards the project team, materialized in the help given and the general appreciation of our activity, the desire to find out the origin and especially to understand the inscribed text, as well as the punctual attempts to preserve some of the locals, give us reasons to hope for a lasting future for these cultural monuments. Unfortunately, there are worrying aspects such as the lack of unitary reaction of the community when a monument of this type is threatened or destroyed and, on the other hand, the misunderstanding of all the attributes of a votive cross (aesthetic, typological, epigraphic, etc.). This last aspect, together with poorly understood preservation methods, caused irreparable losses especially in the situations in which an attempt was made to save or preserve them by erecting constructions around the crosses, which led either to burying the base inscriptions in a concrete foundation; blocking access to the inscriptions or representations painted on the back or sides; transforming them into shrines by covering them with cult objects (icons, candles) or by transforming the space into a place for storing household waste (Fig. 18). Another problem is the lack of registration of crosses in the administrative records, which, corroborated with the absence from the list of historical monuments, allows for their destruction without any legal consequences. However, some of these situations arose as a result of a general lack of

²⁵ Șandric 2018, 35.

knowledge of minimum principles for the conservation and protection of monuments in local communities, a fault that must be attributed to local and central authorities charged with these responsibilities. In fact, during our campaigns, we aimed not only to document the crosses and stone monuments but also to transfer some of the aforementioned principles to the locals we met and especially to the authorities, each trip beginning with the visit to the town hall of the territorial administrative unit. A consequence of this approach was the intervention of two town halls on two stone crosses that were repositioned vertically after one was intentionally demolished and the other collapsed due to natural causes, as well as the firm request to send the records to local authorities.²⁶

Going beyond the retrieval of information and creation of a systematic record, the biggest gain of the project was raising awareness of young people and encouraging their involvement in taking actions for the conservation and protection of cultural heritage. The planned and finalized activities aimed to create a sustainable interest and care for local markers, which will hopefully perpetuate to the next generation and materialize in efforts for preservation and capitalizations of cultural monuments.

²⁶ Șandric 2018, 35-36.

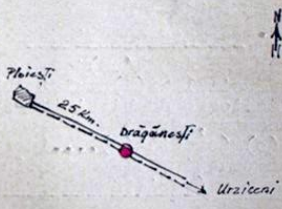
EVIDENȚA MONUMENTELOR ISTORICE DIN R. P. R.		
C. S. C. A. S. DIRECȚIA MONUMENTELOR ISTORICE		FIȘA DE MONUMENT DE ARTĂ PLASTICĂ: Data : 3.X.1965 Indice :
1. Denumirea : CRUCEA DE LA SOPLEA 1.		
2. Adresa : Drăgănești, în curtea locuito- 2. Comuna : rului Toma Mihai (pe drumul ce duce de la gară spre com. Raionul : COMUNA DRAGANESTI Gherghița), Regiunea : RAIONUL PLOIESTI Regiunea PLOIESTI		5. Schiță de amplasament : 
3. Felul monumentului (statuie, cruce, etc.) 3. Cruce		
4. Subiectul : Cruce comemorativă 4.		
6. Anul (sau secolul) edificării : 6. 1655 iulie 13		8. Materialul din care e făcut : Piatră a) Monumentul
7. Autori (sculptor, arhitect, meșter): 7. Ispravnic : Andronic -căpi- tanul din Gherghița		b) Soclul

Fig. 1 - Record sheet for the *Crucea de la Soplea* monument, located in the archives of the National Institute of Heritage, *Historical Monuments Directorate* fund.



Fig. 2 - Photographs of the *Cross from Soplea* from the record sheet of the monument, archive of the National Institute of Heritage, *Historical Monuments Directorate* fund.

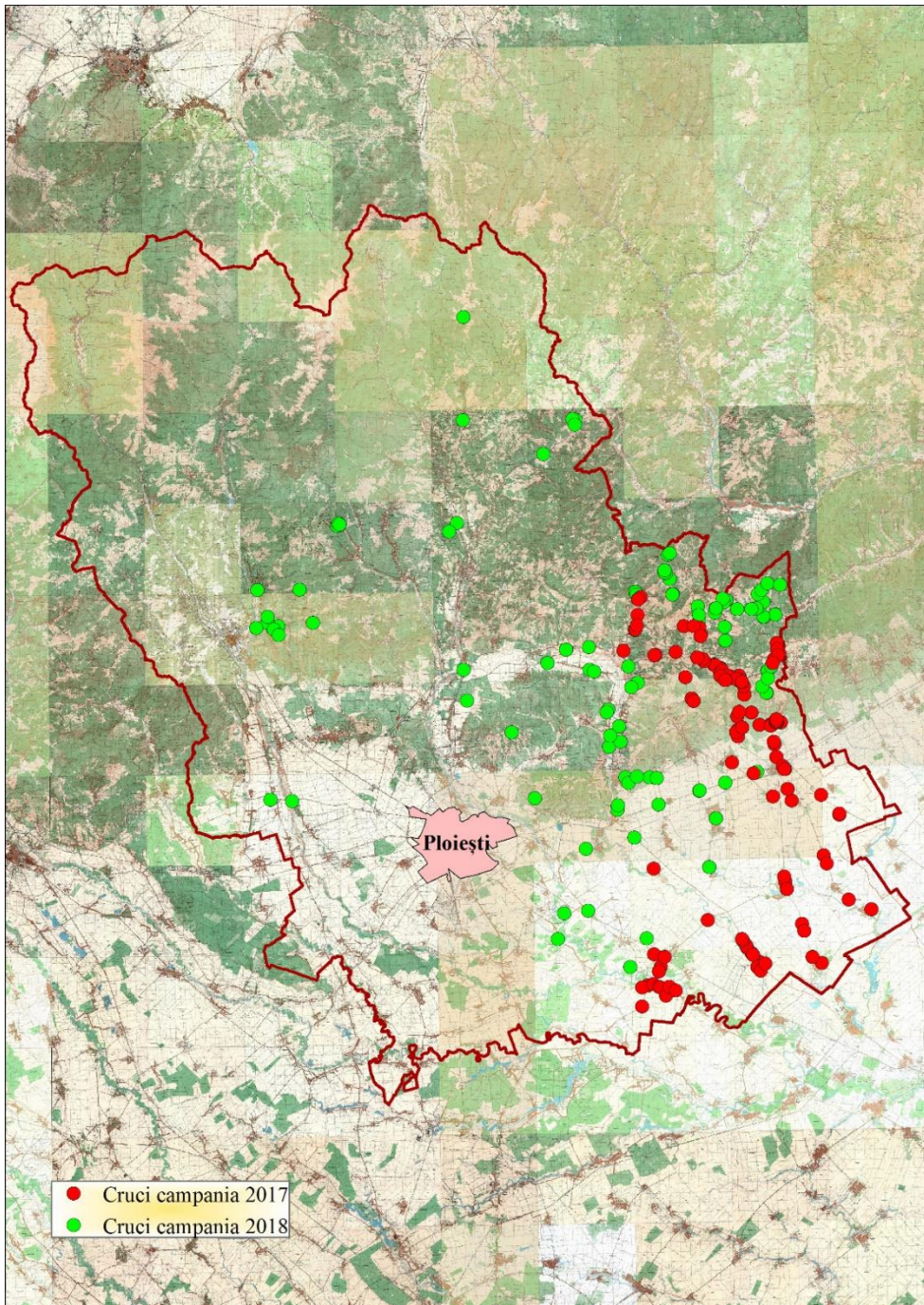


Fig. 3 - The map of the stone crosses identified in the 2017 and 2018 campaigns using as background the Topographic Map of Romania, 2nd edition, scale 1:25.000, made by the Military Topographic Directorate.

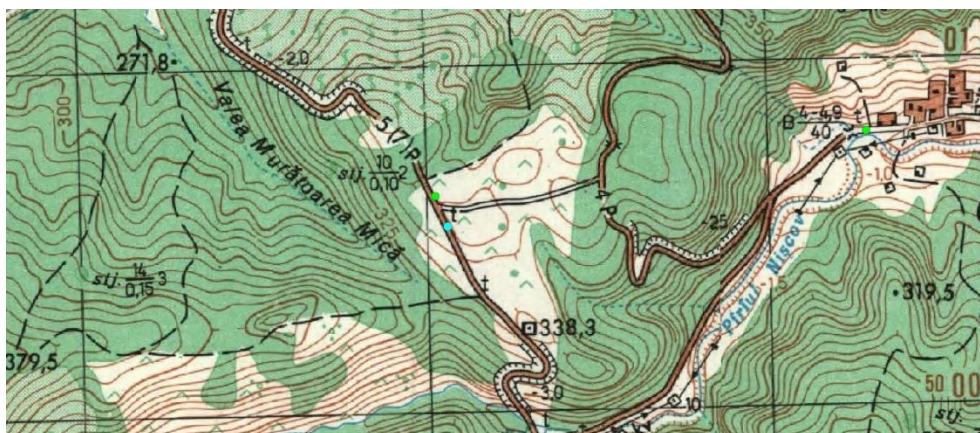


Fig. 4 - The monuments are represented schematically on the topographic map with the Latin cross symbol. The dots in green and blue represent their real position on the field.

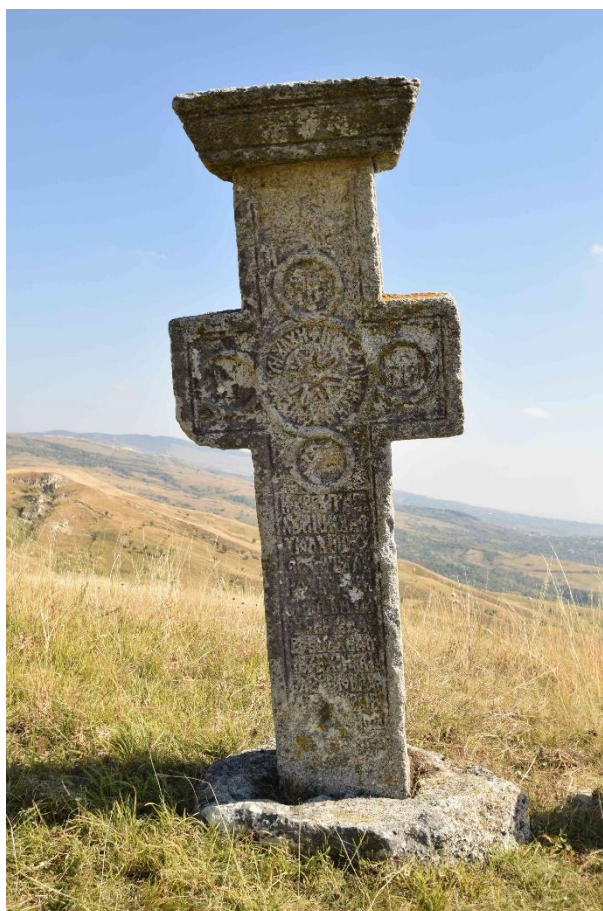


Fig. 5 - Latin cross type located in Jugureni, dated in the 18th century.



Fig. 6 - Celtic cross type located in Urlați, dated in the 19th century.



Fig. 7 - Cross with cubs type located in Mărunțiș, with uncertain dating.



Fig. 8 - Cross with five medallions, the crucifixion formula, the tetragram of the victory of life over death, circular inscription, vegetal decoration and a symbol that could represent the coat of arms of a family, with uncertain dating.



Fig. 9 - Cross with geometric decoration in Ghinoaica, with uncertain dating.



Fig. 10 - Cross with painted decoration in Gura Vadului, dated in the 20th century.



Fig. 11 - Photographic documentation of a cross made during field survey.

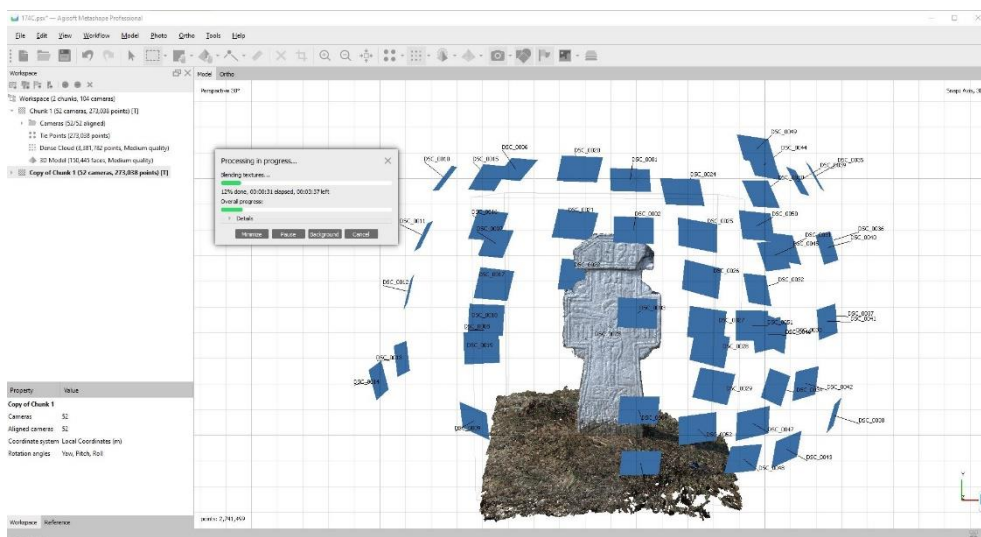


Fig. 12 - Image during the photo processing in the Agisoft Metashape Professional software to generate a 3D digital model of a cross from Loloiasca.

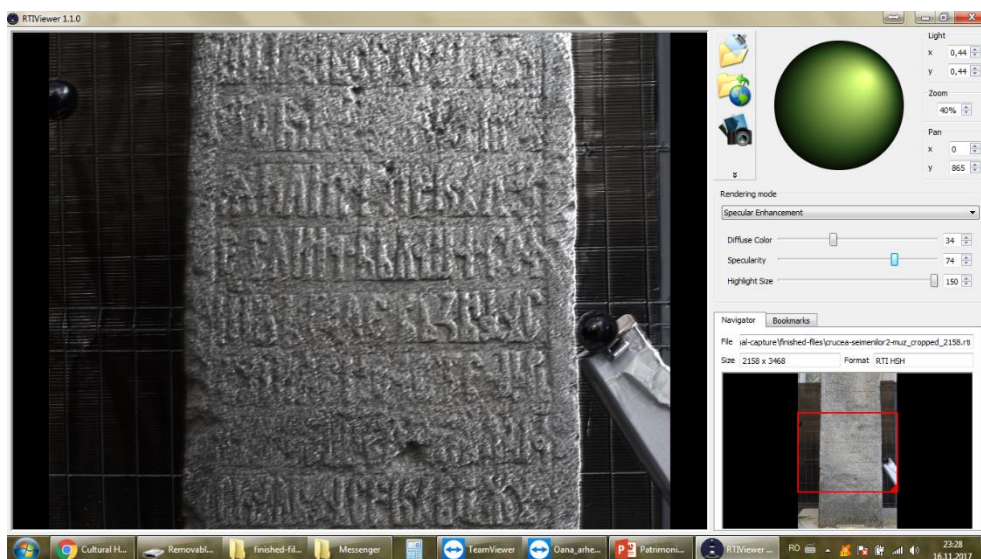


Fig. 13 - RTI image of the inscription from the *Seimeni cross* found in the collection of the Prahova County Museum of History and Archeology.

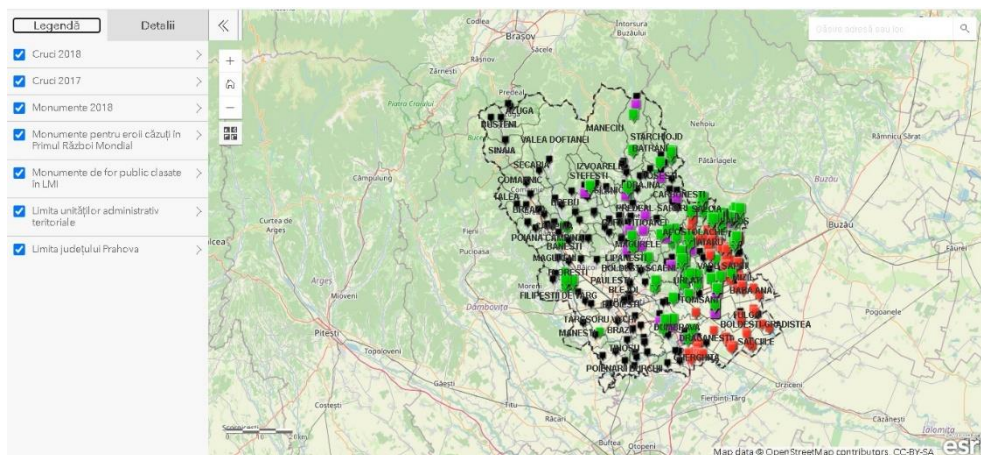


Fig. 14 - The interactive online map can be accessed at <https://monumentelarascruce.ro/harta/>.



Fig. 15 - Image during the documenting of the crosses together with the school children from the Primary School from Boldești-Grădiștea.



Fig. 16-17 - Images from the final project conference held in 2018 at the Teacher's Training House Prahova, Ploiești.



Fig. 18 - Cross transformed into an shrine by erecting around it a brick construction with a shingled roof and decorated with traditional towel.

Bibliography

- Aldea, G. (1969) *Sculptura țărănească în piatră*, București: Meridiane.
- Calinic Argeșeanul, Constantinescu, G. (1999) *Cruci de piatră*, Pitești: Europroduct.
- Crețeanu, R (1969) 'O a doua cruce a lui Constantin Vodă Șerban privind răscoala de la 1655', *Monumente Istorice. Studii și lucrări de restaurare*, pp. 122-130.
- Iancu, L. (2018) 'Cruci de piatră și monumente de eroi, între memorie, artă și societate', in Șandric, B., Iancu, L., Borlean, O., Streinu M., Buruiană, M., Hâncu, G., M., Stan, A., Dinu, G., Nistor, G., Delea, M. (eds.) *Patrimoniu la răscruce. Digitizarea crucilor de piatră și a monumentelor de eroi din Primul Război Mondial din județul Prahova*. București, pp. 8-15.
- Iancu L., Șandric, B. (2018) 'Introducere', in Șandric, B., Iancu, L., Borlean, O., Streinu M., Buruiană, M., Hâncu, G., M., Stan, A., Dinu, G., Nistor, G., Delea, M. (eds.) *Patrimoniu la răscruce. Digitizarea crucilor de piatră și a monumentelor de eroi din Primul Război Mondial din județul Prahova*. București, pp. 6-7.
- Șandric, B. (2018) 'În loc de concluzii', in Șandric, B., Iancu, L., Borlean, O., Streinu M., Buruiană, M., Hâncu, G., M., Stan, A., Dinu, G., Nistor, G., Delea, M. (eds.) *Patrimoniu la răscruce. Digitizarea crucilor de piatră și a monumentelor de eroi din Primul Război Mondial din județul Prahova*. București, pp. 35-36.
- Șandric, B., Borlean, O., Buruiană, M., Streinu, M. (2018) 'Conservarea patrimoniului uitat de la răscruce. Legislație și digitizare.', in Șandric, B., Iancu, L., Borlean, O., Streinu M., Buruiană, M., Hâncu, G., M., Stan, A., Dinu, G., Nistor, G., Delea, M. (eds.) *Patrimoniu la răscruce. Digitizarea crucilor de piatră și a monumentelor de eroi din Primul Război Mondial din județul Prahova*. București, pp. 16-18.
- Șandric, B., Iancu, L., Borlean, O., Streinu M., Buruiană, M., Hâncu, G., M., Stan, A., Dinu, G., Nistor, G., Delea, M. (2018) 'Conservarea patrimoniului uitat de la răscruce. Promovare și educație', in Șandric, B., Iancu, L., Borlean, O., Streinu M., Buruiană, M., Hâncu, G., M., Stan, A., Dinu, G., Nistor, G., Delea, M. (eds.) *Patrimoniu la răscruce. Digitizarea crucilor de piatră și a monumentelor de eroi din Primul Război Mondial din județul Prahova*. București, pp. 19-23.

Legislation

- EUR-Lex (2020) *Recomandarea Comisiei Europene privind digitizarea materialului cultural și conservarea acestuia online (2011)*. Available at : <https://eur-lex.europa.eu/legal-content/RO/TXT/PDF/?uri=CELEX:32011H0711> (Accessed: 10 November 2020).

Technical bibliography

- Cultural Heritage Imaging (CHI) (2020) *Reflectance Transformation Imaging (RTI)*. Available at: <http://culturalheritageimaging.org/Technologies/RTI> (Accessed: 12 November 2020).

Good practices in the ProEuropeana Digital Library **<https://biblioteca-digitala.ro/>**

Vasile Andrei, Bogdan Șandric, Cosmin Miu, Oana Borlean,
Marian Tufaru, Florentina Ghemuț

(INP – National Heritage Institute)

Already for more than a decade now, digital books have become a common thing to the public. Although many of us are still emotionally attached to physical books, we must admit that accessing thousands of books on a single device is an undeniable asset. Regardless of how we choose to read it, first, we need to be able to reach it. Unlike a printed book, whose availability is limited depending on the extent of its circulation, access to a digital book may be virtually unlimited.

In the light of it, the digital library that we present in this article, The ProEuropeana Digital Library, came out to facilitate access to a category of scientific books quite hard to find, that of museum publications. With limited technical possibilities and in a country rather reluctant to freely share books, the project of a digital library dedicated especially to humanistic researchers started in 2011, following some other small projects that aimed to digitize archaeological publications. In ten years, the project has developed; just as in Romania the community of those involved in written culture has become increasingly aware of the added value of freely accessible publications. At first, The ProEuropeana Digital Library emerged from the desire to contribute to historical and archaeological research carried out by the Romanian specialists, but lately, the perspective of the project has been extended outside the field of the humanities.

Thus, ten years after its launch, the project has gained notoriety and covers various fields of cultural publications, while also covering the need to save and unveil older publications, almost unknown to today's public.

Why digital libraries?

What is a digital library? It is essentially a digital collection comprising various books, publications, periodicals, magazines, historical documents, newspapers and/or gray literature (in our case, brochures dedicated to museums or cultural events in Romania). These types of libraries can be dedicated strictly to employees or researchers within an institution (without external access) or may be accessible to the general public. Digital libraries have (or should have in theory) the role of safely preserving and making accessible to the general public various cultural resources in an efficient way, being accessible through appropriate descriptors. Of course, depending on the field we are discussing (such as intelligence or the military), some digital libraries by their very nature should not be open to the general public.

The emergence of such a project in Romania was of course animated by the global trend in digitization, a development that was obviously facilitated by the technological revolution initiated by the internet and the further increasing access of the world population to relatively cheap and mass-produced electronic means of communication and visualization (personal computers, webcams, laptops, smartphones, e-book readers, virtual reality, phablets, and tablets). Digital libraries enable researchers to contribute faster, more securely and more efficiently to the research areas in which they operate. These benefits are not only limited to the efficiency of communication, but can also be tools through which we can lead to the rapid and sustainable development of society, including by protecting the environment (eliminating the need to cut trees and produce paper, but also to recycle paper that has already been produced). Even if a digitized book does not use paper and does not require restoration, care, or special preservation, we must not forget that there is still the problem of recycling electrical components.

To highlight the advantages of a digital online library, we can start by stating that its main asset is its non-physical character (a feature that saves physical space). The second major benefit is security, as these digital warehouses (if we can call them that) are not susceptible to physical destruction. This preservation of data can lead in time to a vast accumulation of experience and can lead to profound changes in society, like innovation, economic development, and technological progress. There is also a highly altruistic component, with some of the creators of such libraries keeping in mind future generations. These libraries can also function as time capsules.

The last few years have unfortunately proved that the cultural heritage of mankind is in constant danger, be it material, immaterial, immovable, or mobile. Some of the unfortunate examples are the destruction caused by the Islamic State in Iraq and Syria (such as Nimrud or Temple of Bel) and the 2018 fire at the National Museum of Brazil. Languages and traditions can disappear, buildings can collapse due to wars, accidents, negligence, or carelessness. Physical libraries are among the most sensitive institutions, mainly due to materials that are particularly susceptible to fires. In Romania, one of the saddest such cases is the destruction of the Central University Library "Carol I" in Bucharest, an institution that suffered invaluable material losses during the Romanian Revolution of 1989.

Although the digital environment can be considered safe and efficient, we must not forget that it is also exposed to change and transformation (as is anything else that exists in this universe). As long as software programs benefit from updates and improvements, employees are constantly familiar with new technologies, and data benefits from backup, institutions that host digital libraries should not face particularly many problems in managing a digital library.

The technological evolution of the last two decades has led to the formation and development of interdisciplinary teams in the field of digital libraries, Romania being positively affected by these changes. The team within the Digital Heritage branch of the National Institute of Heritage is composed of a programmer, staff responsible for scanning, bibliographers, and historians (with various specializations, such as archaeology).

The current team of the institute is constantly updated on the different trends and changes in the field of culture. All these observations show us that the dissemination of scientific articles in the humanities will continue to grow and develop, mainly due to the implementation of new information technologies such as 5G technology. This technology will allow the development of unprecedented download speeds. The current team has developed a coherent and efficient workflow for technical improvement of the library, but also for enriching its content. Employees also ensure quality control within the digital library, noting errors and problems and reporting them to superiors and to the department's programmer.

The European context

Going back in time, we can set the start time of the European Commission's concerns about the creation of digital libraries in the year 2005, when the European Commission launched the i2010 initiative ("i2010 – A European Information Society for growth and employment")¹, aiming to maximize the benefits of new information technologies, in order to accelerate economic growth, making digital libraries one of the flagship initiatives in this process. The Communication "i2010: Digital Libraries" of 30 September 2005² set out the strategy for digitization, online accessibility, and the digital preservation of Europe's collective memory. This collective memory includes various types of prints (books, magazines, and newspapers), photographs, museum objects, archival documents, and audio-visual materials.

The European Union started the project of a digital library in 2006, when the Commission issued a Recommendation on 24 August 2006³, regarding the digitization and online accessibility of cultural material and digital preservation. Member States were to promote a European digital library in the form of a common, multilingual access point to digital cultural material owned by various organizations. Thus, Europeana - Europe's digital library, archive, and museum⁴, was launched five years later, on November 20th, 2011.

Digitizing and preserving Europe's cultural memory is also one of the key areas addressed by the Digital Agenda for Europe in the year 2010. European approaches in this area, including the development of the Europeana, were supported by the European Parliament and the European Council in Parliament resolution of 5 May 2010⁵ and in the Council conclusions of 10 May 2012⁶. A new Commission Recommendation of 27 October 2011⁷ called on the Member States to set clear quantitative targets for the digitization of cultural resources, indicating the expected increase in digitized material that could be part of Europeana. It was also recommended to improve access to public domain material by ensuring that public domain content remains in the public domain once digitised and by promoting the widest possible access to digitised

¹ <https://eur-lex.europa.eu/legal-content/GA/TXT/?uri=CELEX:52005DC0229>

² <https://eur-lex.europa.eu/legal-content/EN/LSU/?uri=celex:52005DC0465>

³ <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32006H0585>

⁴ <https://www.europeana.eu/en>

⁵ https://www.europarl.europa.eu/doceo/document/TA-7-2010-0133_EN.html?redirect

⁶ [https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52012XG0615\(02\)&from=DE](https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52012XG0615(02)&from=DE)

⁷ <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32011H0711>

public domain material as well as the widest possible reuse of the material for non-commercial and commercial purposes.

At Romanian level, by Government Decision, the National Program for digitization of national cultural resources and the creation of the Digital Library of Romania was approved in 2008⁸, appointing CIMEC as the responsible institution, and in 2015 the National Strategy on the Digital Agenda for Romania 2020 was approved, transposing and adapting some of the objectives set by the European Digital Agenda.

Thus, the Digital Library of cultural publications had all the conditions to attract more support and collaboration from all the institutions involved in this field.

Start-up and first achievements

Unlike other similar projects, the Digital Library started as a small project of the former CIMEC - Institute of Cultural Memory (now integrated in the National Heritage Institute under the name of "Digital Heritage Directorate") reflecting its mission to document, digitise and disseminate cultural resources.

The origins of the Digital Library of Cultural Publications project can be found in the collaboration between CIMEC and the "Vasile Pârvan Institute of Archaeology", under the auspices of the European ARENA project, carried out between 2001-2004. Within this project, documents from the historical archive and from the archive-file "Archaeological Repertory of Romania" were digitised, which were then to be preserved but also disseminated online to the general public⁹.

A new project followed, also in collaboration with "Vasile Pârvan Institute of Archaeology" - a process of indexing the publications „Dacia”¹⁰ and „Materiale și Cercetări de Arheologie”¹¹ (Materials and Archaeological Research), whose first series were fully digitized and displayed online. Then, a new project assumed the scanning and online publication of 31 books, studies, and articles on prehistory, epigraphy, and ancient history signed by the preeminent Romanian archaeologist, Vasile Parvan.

⁸ National Strategy on the Digital Agenda for Romania 2020 - Government Decision no. 245 of April 7, 2015, Official Gazette of Romania, Part I, No. 340 bis / 19.V.2015

⁹ www.cimec.ro/Arheologie/Arhiva-Digitala/Sumar.htm and www.cimec.ro/scripts/ARH/RAR-Index/sel.asp

¹⁰ <http://www.cimec.ro/scripts/dacia/selarticole.asp>

¹¹ http://cimec.ro/Arheologie/mca_rom.htm

In the following years (2008-2009) the aim was to digitize various publications in the field of archaeology and history, without establishing an official strategy, but rather according to circumstantial contacts that were established by mutual agreement with various authors. Starting with the year 2011, when CIMEC - Institute of Cultural Memory becomes part of the National Heritage Institute, the development of the digital library focused on the need to exhibit scientific and cultural periodicals published by county or local museums in Romania. We focused on research publications that were difficult to obtain, with limited circulation and limited availability (most of them to be found only in a few specialised libraries). The proposed goal was to digitize as many of the museum publications as possible, starting initially with the periodicals published by the Botosani County Museum and then following a direction that led to the south of the country.

Today, museum publications included in the digital library cover almost every county of Romania. We would like to point out here the project that aimed at cataloguing the studies of history and archaeology of Bucharest, and which involved the digitization of the articles published in the serial publications „București. Materiale de Istorie și Muzeografie” (“Bucharest. Materials of History and Museography”)¹² and „Cercetări Arheologice în București” (“Archaeological Research in Bucharest”)¹³, published by the Museum of Bucharest between 1963 and 2017. A total of 40 volumes and 1,100 articles allow a reconstruction of the history, archaeology, and cultural and social life of Bucharest, from the first prehistoric traces to the present day.

We also mention here another example that we consider representative, a publication dedicated to a distinct field, that of numismatics, *Revista Cercetări Numismatice* (Journal of Numismatic Research)¹⁴ published since 1978 by the National Museum of Romanian History (in 18 volumes), whose digitization made freely searchable a valuable archive of studies in a field of wide interest, that of numismatics and opened new horizons for research, and study in the same field.

¹² <https://biblioteca-digitala.ro/?pub=192-bucuresti-materiale-de-istorie-si-muzeografie-muzeul-municipiului-bucuresti>

¹³ <https://biblioteca-digitala.ro/?pub=311-cercetari-arheologice-in-bucuresti-muzeul-municipiului-bucuresti>

¹⁴ <https://biblioteca-digitala.ro/?pub=310-revista-de-cercetari-arheologice-si-numismatice-muzeul-municipiului-bucuresti>

Workflow and process

For a publication to end up in the Digital Library, a set of steps needs to be carried out. The first step is to obtain the agreement for digital reproduction and online publication from the author or the institution which has published the book. Then follows the stage of scanning and processing the obtained images, which can be semi-automatic or manual, a process that also includes the conversion of images into text, via OCR. The digital file obtained is then cut into parts, on items, which are then attached to the catalogue sheet via a link established by the folder and file names.

Thus, the database of ProEuropeana Digital Library contains several tables with metadata indicating the title, the publisher, type of publication, collection, associated fields, ISSN and ISBN codes, a brief summary, no. of the volume/tom, year of publication, publishing house and the city where it was published, cover photo, keywords. These tables are intended to manage the classification process or cataloguing sheets and allow the reuse and conversion of information (catalogue sheets) already created. We aim to add at least five keywords to every item, being a book, a magazine, a flyer or an article.

The online information search and retrieval application are built on a SQL Server database and an ASP platform. The user of the application can search by a set of criteria such as title, author, topics and domains or string and allows access and editing (cataloguing, editing and attaching files) online.¹⁵

The Digital Library appeared as an idea within CIMEC and it was built around its own concept, so instead of using a purchased format, the institute created a new one. The fact that, for serial publications, each article is catalogued separately, benefiting from a set of metadata that ensures easy identification: title, author, abstract (often in a foreign language), keywords, etc., is a valuable feature of the Digital Library.

Current and future development

Unlike the beginning, the Digital Library has now a much broader purpose, that of increasing Romania's presence in the European digital library – Europeana – hence the name ProEuropeana. Thus, the old catalogue sheets are converted to the EDM (Europeana data model) format which is now used for all the new catalogued publications.

¹⁵ Șandric, B., Matei, D., Vîlcu, A., 2015

As the Culturalia platform (the Romanian counterpart of Europeana.eu) will begin to function, ProEuropeana will be integrated here, which will give it major visibility.

The ProEuropeana Digital Library aims now not only to cover museum publications, but also those published by cultural centres, research institutes, associations as well as all publications of cultural interest regardless of the publishing house. In terms of geographical distribution, Digital Library includes publications from all over Romania.

Currently, the ProEuropeana digital library contains 2,140 publications and 87,068 individually catalogued articles - published all over the country. They can be grouped into 43 domains, covering wide areas of interest, the list is open for new additions: anthropology, archaeology, architecture, art, astronomy, biography, old book, conservation, digitization, documents, education, ethnography, ethnology, cultural events, film, philology, philosophy, geography, history, literature, medals, interdisciplinary methods, historical monuments, museums and collections, museology, music, numismatics, national cultural heritage, politics, cultural policies, religion, restoration, security, science and technology, natural sciences, theatre, tourism.

The ProEuropeana Digital Library will connect with other cultural resources, transforming it into an online platform where institutions or groups will be able to publish their future volumes. It is also worth mentioning that the library aims to add other resources insufficiently known or exploited, such as the catalogue of old and rare Romanian books, the catalogue of incunabula, the Reference Bibliography of Old Books, publications of special relevance in the history of performing arts. In addition to these cultural resources, the National Heritage Institute aims to add to this library publication that include more information on the technical and industrial heritage of Romania.

As a conclusion, we consider that the Digital Library of Cultural Publications - ProEuropeana stands out in a special way among other online indexed journals from Romania –as users can search for publications by author and descriptor. Its main advantage is the dynamic promotion of cultural publications (books, periodicals, gray literature). The project, which, so far, did not benefit from external funding, being a purely internal project carried out with the own resources of the National Heritage Institute, is being constantly improved and developed.

Bibliography

- Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions - "i2010 – A European Information Society for growth and employment" {SEC(2005) 717} COM/2005/0229 final
- Communication from the Commission of 30 September 2005 to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions – i2010: digital libraries [COM (2005) 465 final – Official Journal C 49 of 28.2.2008].
- Commission Recommendation of 24 August 2006 on the digitisation and online accessibility of cultural material and digital preservation, OJ L 236, 31.8.2006, p. 28–30
- Commission Recommendation of 27 October 2011 on the digitisation and online accessibility of cultural material and digital preservation, OJ L 283, 29.10.2011, p. 39–45
- Council Conclusions of 10 May 2012 on the digitization and online accessibility of cultural material and digital preservation, Official Journal of the European Union C 169/5 of 15.06.2012
- Report from the commission to the European Parliament and the Council on the evaluation of Europeana and the way forward, COM/2018/612 final
- National Strategy on the Digital Agenda for Romania 2020 - Government Decision no. 245 of April 7, 2015, Official Gazette of Romania, Part I, No. 340 bis / 19.V.2015
- Government Decision no. 1676/2008 on the approval of the National Program for digitizing national cultural resources and the creation of the Digital Library of Romania, Official Gazette, Part I no. 855 of 19.12.2008
- Matei, D. (2009), *Spre Europeana.eu: O introducere în bibliotecile digitale*. cIMEC – Institutul de Memorie Culturală
- Șandric, B., Matei, D., Vilcu, A.(2015) *ProEuropeana – Biblioteca Digitală a Publicațiilor Muzeale*. Institutul Național al Patrimoniului

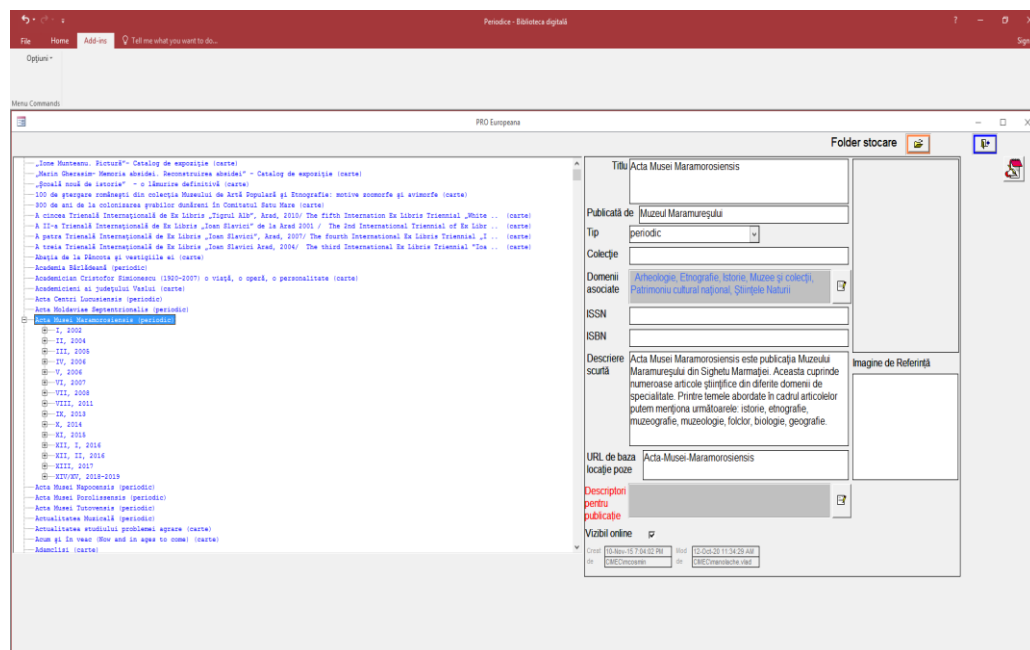


Fig 1. Digital library database (viewing a periodical).

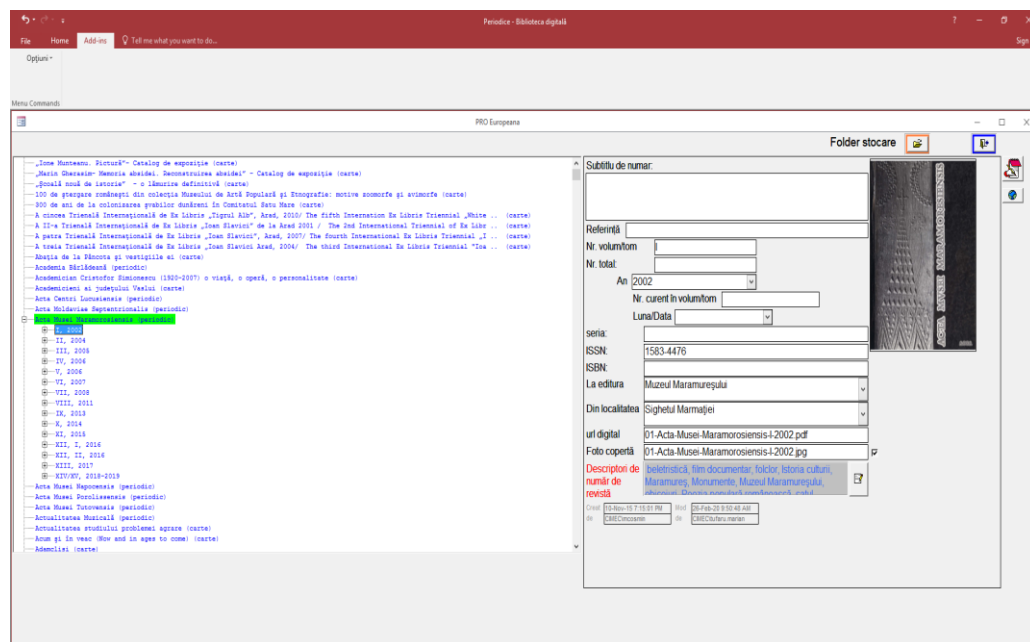


Fig. 2 Digital library database (viewing a volume from a periodical).

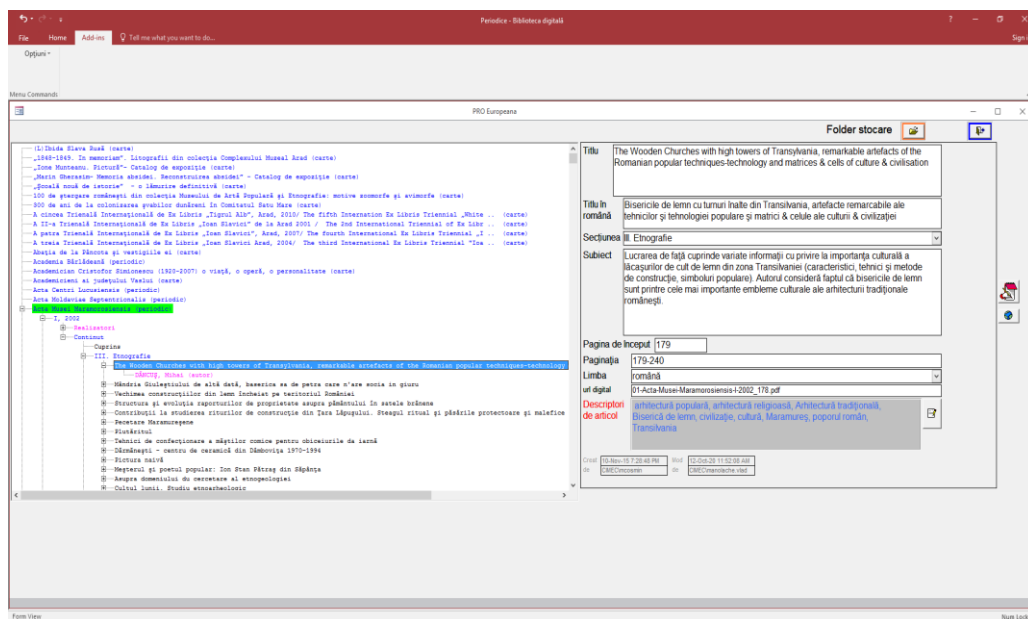


Fig. 3 Digital library database (viewing an article in a periodical).

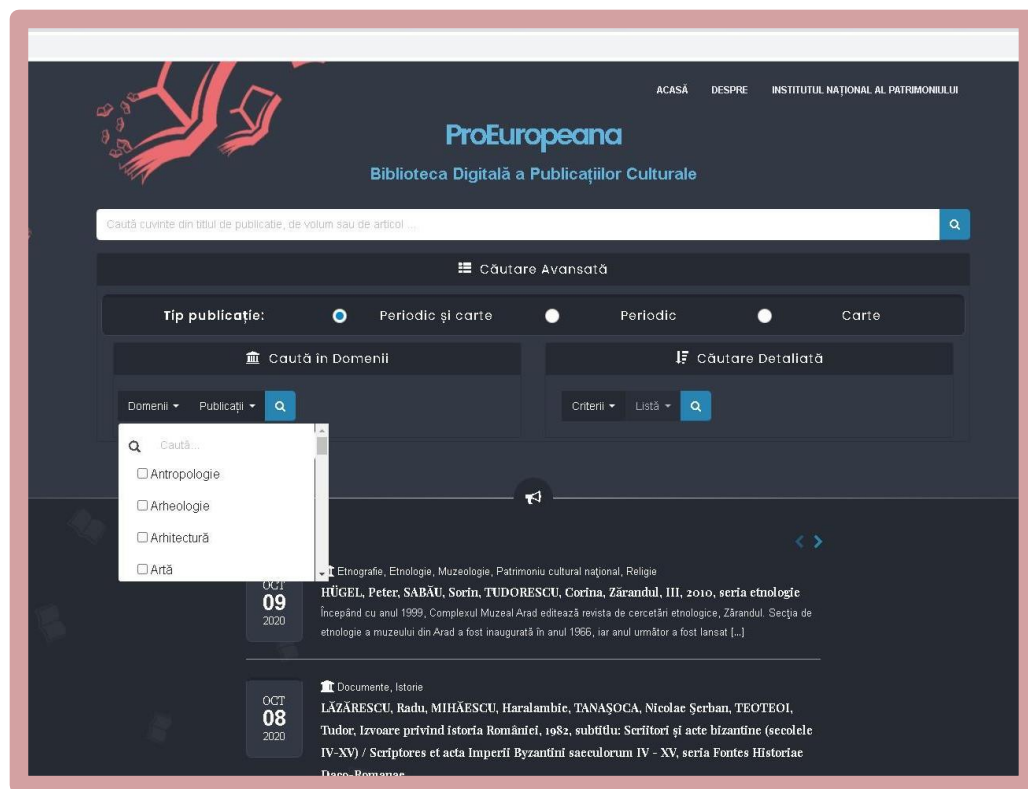


Fig. 4 Publication search interface on the site (search criteria).

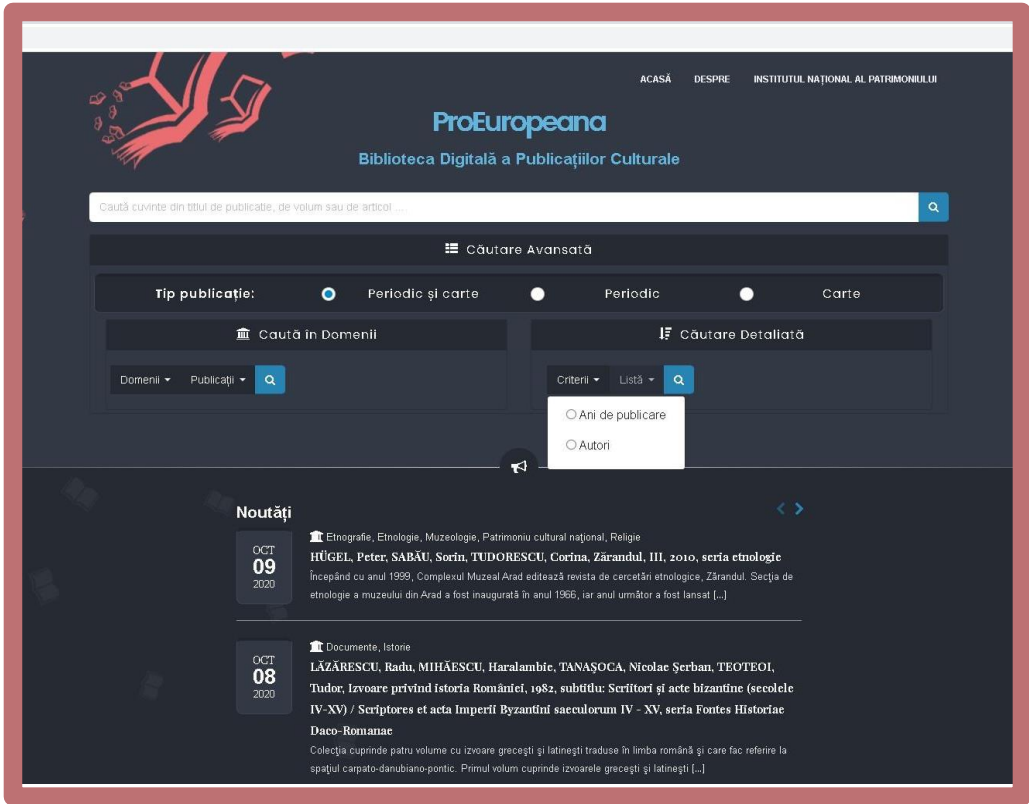


Fig. 5 Publication search interface on the site (search criteria).

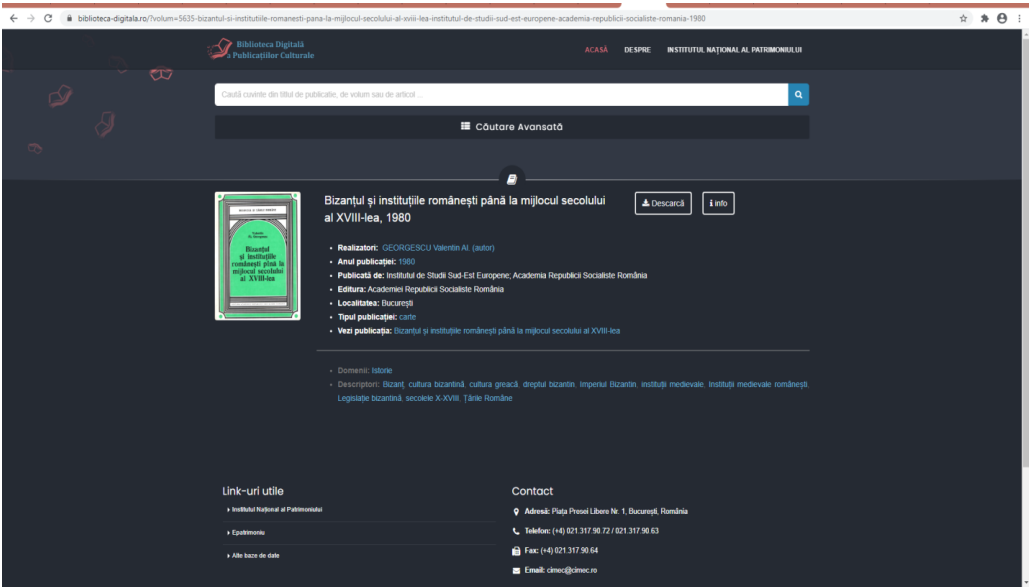


Fig. 6 Publication search interface on the site (a book).

The *E-culture* Project: the *Culturalia* platform and quantitative ambitions

Dan Matei, Bogdan Șandric

(INP – National Heritage Institute)

The *E-Culture* Project

E-Culture is a project carried out (between July 2018 and June 2021) by the Project Management Unit of the Ministry of Culture, in partnership with 29 memory institutions (The National Heritage Institute, museums, libraries, The National Film Archive), plus the Public Television and the Public Radio. It has a funding of about 11 million euros (European funds: 9 million, national co-financing: 2 million).

The declared goal of the project is the creation of *Culturalia*, the Digital Library of Romania, which aspires to be the national pendant of *Europeana*, the European Digital Library¹ (fig. 1).

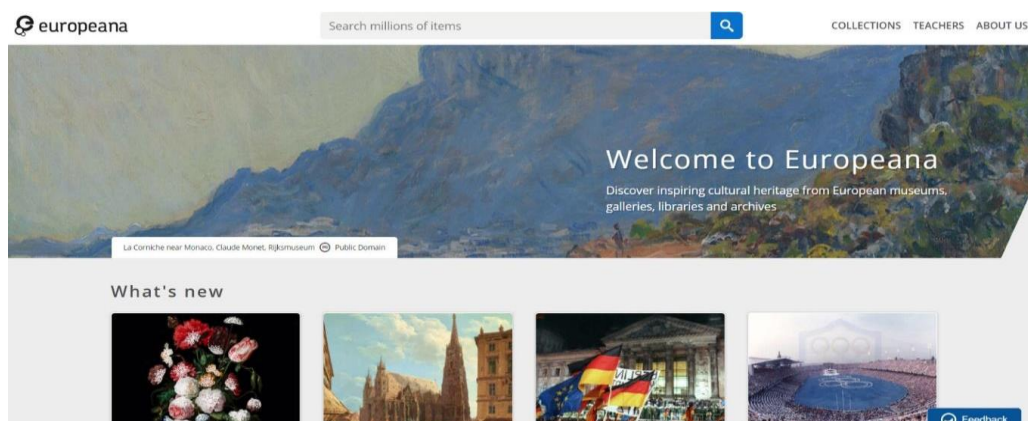


Fig. 1. *Europeana* – the European Digital Library

¹ <https://www.europeana.eu/en>

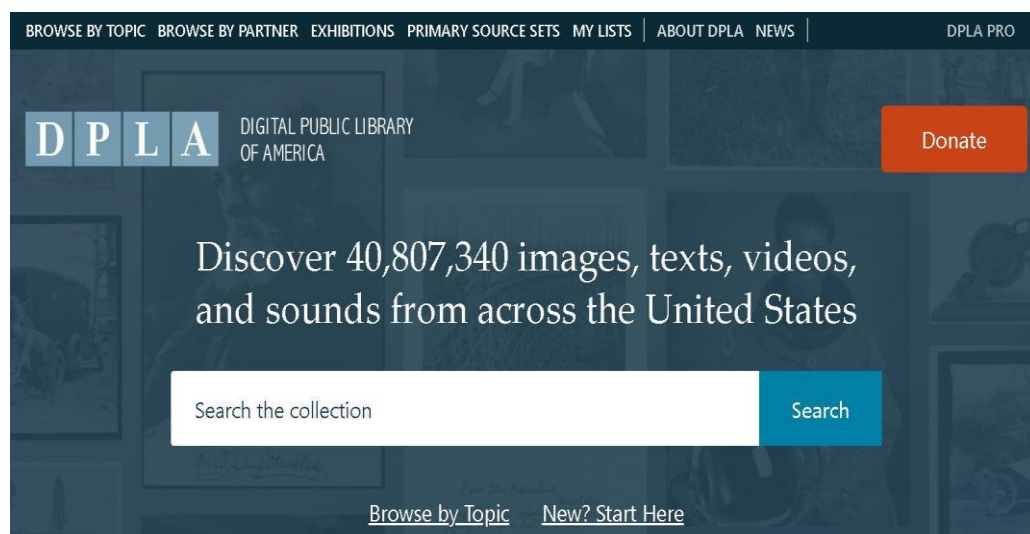


Fig. 2. DPLA –The Digital Public Library of America²

What "club" do we want to join ?

Culturalia will to be a “national” digital library³, similar to those of many European states. For example, DPLA [Digital Public Library of America] (fig. 2) and DDB [Deutsche Digitale Bibliothek] (fig. 3).

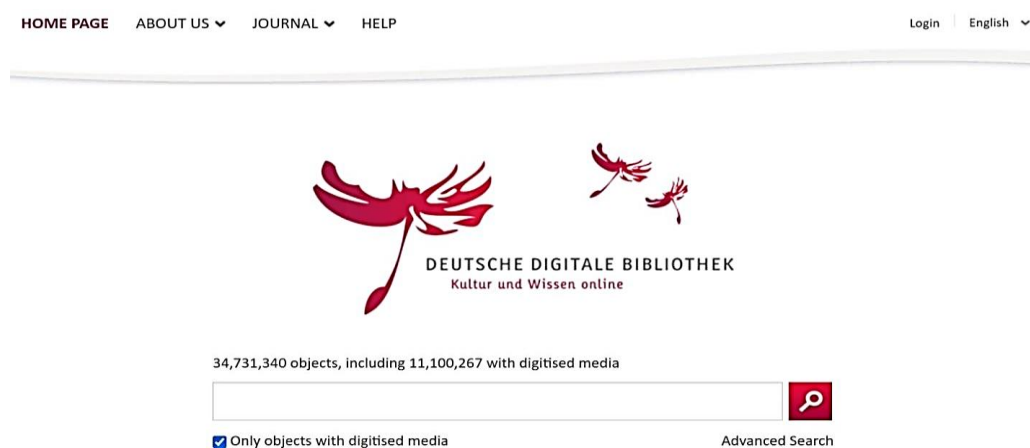


Fig. 3. DDB – Deutsche Digitale Bibliothek [the German Digital Library]

² <https://dp.la/>

³ NB. It is called "library" only because of the lack of a generic term for "library", "museum" and "archive", so it brings together collections of any kind of memory institution, i.e. museums, libraries, archives.

Two major objectives

- A. The development of the IT platform for the shared catalogue and for the digital library portal. It will manage the database containing the metadata (i.e. the catalographic records) of the cultural resources, including resources not exposed online). Thus, it will be similar to the DDB⁴ platform (see Fig. 3), which, of its nearly 35 million resources, exposes online just over 11 million.
- B. The online exposure of over 550,000 cultural resources (texts, images, audiograms, videograms, 3D digital objects), of which about 200,000 are to be exhibited also in *Europeana*.

But the project has also some collateral objectives, such as:

- the computerisation of some of the administrative procedures associated with the mobile cultural heritage (listing/unlisting, the issuance of permits and accreditations, export of cultural objects, the recording of missing goods, etc.);
- assisting the users in the preparation of bibliographies (in standard formats);
- the set-up of a mechanism for identifying and contacting the copyright holders.

Target-users of the project

In order of importance, the targeted users are:

- the general public;
- the teachers/students (as producers/consumers of teaching material);
- the cultural heritage professionals.

Having the general public as the main target implies a new way of cataloguing cultural resources, especially in museums and archives. The librarians are used to producing descriptions of cultural resources for the eyes of the public, but for the museographers and archivists, the

⁴ <https://www.deutsche-digitale-bibliothek.de/>

requirement to catalogue the resources in this way is a real "cultural shock". Proof: do we know many museums and archives that display publicly their full catalogues?

But the focus on the general public is somewhat at odds with the available content. Due to copyright reasons, too few of the cultural resources that will be exposed online during the project are contemporary (or at least created in the 20th century). This may not be a handicap in the case of museum and archival resources, but it clearly is one in the case of textual resources. How many of us are interested in reading an 18th century menaion, in Slavonic?

The current status of the project

The good news: over 400,000 cultural resources are already digitised and catalogued.

The bad news: the *culturalia.ro* platform is not yet (November 2020) operational. And its development is not without technical difficulties.

The unavailability of the IT platform led – inevitably – to nonuniformities in the cataloguing of the 400,000 digitised resources. This will require meticulous deduplication and normalisation activities in the future.

The cataloguing of the textual resources – among the 400,000 digitised so far – has led to an unexpected finding: we are almost running out of Romanian books in the public domain! On the one hand, it's good news, but on the other hand, it's not so gladdening that we have so few books in the public domain.

***Culturalia* as national shared catalogue**

The *Culturalia* platform will be, in fact, an online catalogue. This catalogue will offer a public service, i.e. it will be available *for free* to institutions as well as to individuals. And this is not just for digital resources. In other words, anyone (properly registered with the real name!) will be able not only to consult the catalogue, but also to enrich it. Therefore, on this platform, users will be able to set-up the catalogue of their home libraries (using, for the most part, the bibliographic metadata produced by professional cataloguers), or to produce the bibliographies for their works, for example. And, of course, each contributor will decide how much to expose in public.

But what does "shared" mean (in this context)? The metadata (i.e. the catalographic records) provided by a cataloguer may be reused by others. E.g.:

- a bibliographic record produced by a library (or even by a publishing house) can be (re)used by all libraries (institutional or personal);
- the descriptive record of a contextual entity (person, concept, place, period, event) can be (re)used by anyone.

This sharing has two significant benefits:

- massively reduces duplication of effort, i.e. leads to valuable time savings for the professionals;
- promotes consistency in descriptions, thus leads to improved retrieval.

***Culturalia* as digital library**

Since the *Culturalia* platform also records (and displays) digitised cultural resources, it acts as a digital library. For the general public, the online access to cultural resources is important (even if for the museum objects this takes the form of reproductions, i.e. as "surrogates").

Moreover, the online access:

- diminishes the "geographical discrimination" (i.e. urban-rural, Bucharest-province). After all, a citizen in a mountain village pays taxes like me, in Bucharest. But I have access to a multitude of cultural institutions in Bucharest, while she/he has access at most to a communal library;
- allows museums to display the pieces "hidden" in their storerooms. It is well known that all large museums can exhibit only a small fraction of their collections.

In addition, to search in a national digital library has advantages:

- compared to Google/Bing: allows not only point searches;
- compared to an institutional digital library: allows searches across different collections.

On the other hand, we hope that the platform will also give users reasons to return, just for pure pleasure. For instance, to see:

- the painting/sculpture of the day;
- the archaeological piece of the day;
- the photo of the day;
- the poem of the day,

or to visit the virtual galleries, i.e. curated exhibitions and "guided tours".

Culturalia: a state-of-the-art platform (?)

An important ambition of the project is that the IT platform to be at the highest level of contemporary technology. It is a natural ambition for a European project, but it is also motivated by the low hope of obtaining funds for its evolution in the next decade.

From a technical point of view, this ambition materialized in the following functional requirements:

- The descriptive metadata of the resources has to conform to the linked-data⁵ paradigm, in other words, they are not traditional descriptive records, but subject-predicate-object statements. Specifically, the generic data model of the database is EAV/CR [Entity-Attribute-Value/Classes and Relationships]⁶.
- The conceptual data model has to be based on the CIDOC-CRM ontology [Conceptual Reference Model]⁷ (ISO 21127: 2014), promoted by the International Documentation Committee of ICOM [International Council of Museums], combined with the FRBRoo ontology [Functional Requirements for Bibliographic Records object-oriented]⁸, promoted by IFLA [International Federation of Library Associations and Institutions].
- To implement the Query Expansion⁹ (i.e. a query with a search key retrieves also the resources indexed with its specifics and synonyms). Therefore, a terminological thesaurus has to be used.

⁵ https://en.wikipedia.org/wiki/Linked_data

⁶ https://en.wikipedia.org/wiki/Entity%E2%80%93value_model

⁷ <http://cidoc-crm.org/>

⁸ <https://www.ifla.org/publications/node/11240>

⁹ https://en.wikipedia.org/wiki/Query_expansion

Note that for reasons of intellectual responsibility, each statement will be "signed", i.e. its provenance will be recorded (i.e. who said so? when? on what basis?).

These technical requirements are as of yet unusual in the IT circles in Romania, which explains – at least in part – the development difficulties.

E-Culture: the expectations (but also the difficulties)

The *Culturalia* platform is expected to be operational by mid-2021, i.e. it will be made available to the memory institutions as well as to the general public. On the other hand, the provision of over 200,000 cultural resources to *Europeana* will improve the (currently very modest) presence of Romanian culture in the European digital context.

But after that, a difficult period will follow. On the one hand, the ingestion of the legacy institutional catalogues (which have various formats and sufficient peculiarities) will be tough. On the other hand, it will be very laborious to deduplicate the duplicate statements, which appears – inevitably – in the distributed cataloguing process.

But the social value of this cultural service is worth all the effort.

The *Virtual Genealogical Archive*: a pilot digitization project of the parish and civil status registers in the Bucharest and Braşov county archives (Romania) - <http://arhgenvirt.ro>¹

Rafael-Dorian Chelaru

(Associate Professor at the Faculty of Archival Sciences
at the "Alexandru Ioan Cuza")

In 2013, the Executive Unit for Financing Higher Education, Research, Development and Innovation in Romania - UEFISCDI launched a national competition for collaborative research projects (PCCA), financed from public funds and opened to public institutions and private companies. The applicants were expected to propose research projects resulting in products with practical applicability and market potential both for the public and for the private sector.

In June 2014, a consortium led by the Faculty of Archival Sciences at the "Alexandru Ioan Cuza" Police Academy and including National Archives of Romania and SIVECO Romania S.A. (a software company) was awarded a financial support of 1.250.000 lei (c. 275.000 Euros) for creating an online database called *The Virtual Genealogical Archive – a pilot project created for the National Archives of Romania and third-party users*, through which two archival collections of parish and civil status registers became accessible online to the public. The two collections, namely Bucharest Civil Status Collection and Brasov Civil Status Collection, are currently preserved and

¹ The present text is an abridged version of our article by Rafael Dorian Chelaru (2018), *The Virtual Genealogical Archive (ArhGenVirt). A pilot digitization project of the parish and civil status registers in the Bucharest and Braşov county archives*, *Transylvanian Review*, 3, pp.140-154.

administered by the National Archives of Romania in the repositories in Bucharest and Brasov. The project started in October 2014 and ended in September 30th, 2017.

Civil registers in Transylvania and Wallachia up to 1918 – a brief history

As the two collections contain parish and civil status registers referring to communities and individuals that lived in two different historical areas of nowadays Romania, namely Transylvania and Wallachia, a brief introduction in the history of civil status recordings is needed in order to understand the context in which such documents have been created.

Although in Western Europe the first mentions concerning the necessity of registering baptisms, marriages and burials date back from the second council of Lateran (1139) (Delsalle, 2003, p. 17), in Transylvania, their beginnings can be traced only after the Council of Trent (1542-1563), when the Catholic Church imposed to all parish priests to create and administer parish registers in order to record baptisms, marriages and burials. To create the registers, special blanks for recording the data were produced and, at the end of each year, the parish priests bound them up. All these registers were administered by parishes only, without any civil authority being involved. Moreover, there is no evidence that parish registers were used as tools for state government until the end of 18th c. – therefore, their primary purpose was as confessional instruments used by the churches to control their parishioners.

The few parish registers preserved until nowadays for 17th and 18th century suggest that not all parish priests stick to this rather tedious obligation; on the other hand, enforcing it through sanctions was not quite of primary concern for the ecclesiastic authorities. Therefore, keeping up to date parish registers was far from a regular practice among the recognized confessions (*religiones receptae*) in Transylvania. For example, the oldest extant parish registers produced on the territory of nowadays Braşov county date back from 1607 (Lutheran parish of Hosman), 1651 (the reformed church in Făgăraş), 1687 (the Greek-Catholic parish from Daia), 1784 (the Greek Orthodox parish from Orlat) (Moldovan, 1958, p. 63). The oldest extant register containing recordings of a Jewish community from Braşov was compiled in 1835 for the village of Valea Lungă (Brie, 2010, p. 169). The first printed parish registers in Transylvania were produced in

1784 in a printing house in Sibiu; however manuscript registers continued to be used until the 1850s.

The languages used in these documents were Latin (for Catholics), Hungarian (for Calvinists and Unitarians), German (for Lutherans), Romanian (for Greek-Orthodox and Greek-Catholic communities) Armenian (for Armenian communities). The Jewish religious authorities used either Hungarian, German and, more rarely, Jewish (Brie, 2010, pp. 170-2).

The state became aware of the importance of the civil status records only in the second half of the 18th century. The regulations regarding military conscriptions issued in 1773 and 1784 by the Imperial Chancery in Vienna provided also detailed instructions on how the priests of all Christian confessions should record births, marriages and deaths using special forms provided by the state. Additionally, other regulations issued in 1770 and 1774 decreed that the safe preservation of the registers was a legal obligation for all parish priests, who risked severe sanctions in case of disobedience (Brie, 2010, p. 174).

Moreover, the Habsburg authorities in Transylvania established that every extract and certificate regarding civil status data issued by the parish priests had juridical value in law courts, and every change in the civil status of a person, certified by the law courts, had to be recorded also in the parish registers. Moreover, the periodical compilation of conscription lists and demographic statistics for administrative purposes became also the responsibility of the clergy, who, eventually, was given a juridical status very similar to that of the state officials.

In 1827, the state imposed that every register had to be compiled in two copies, one of which had to be deposited at the archive of the local authority at the beginning of the next year. Also, any change in the civil status data had to be recorded in both copies after its validation in the tribunal. In 1850, new headings were added in the civil data recording forms: two headings for baptisms (used for recording the alive/dead newborn and for legitimate/illegitimate births), one heading for marriages (used for recording the status of the grooms as unmarried/widow) and one heading for deaths (used for recording the cause of death).

In 1894, the Austro-Hungarian monarchy established the civil status registers, which replaced the parish registers as documents of legal value. The authorities created within the existing counties (e.g. Brasov and Fagaras, now parts of actual Brasov county) new administrative units called

circles (Popovici, 2005, p. 68) for registering the civil status data, each unit being provided with a civil status office which had the task to fill and keep the registers in good order. Each year, the subprefect of the county had the responsibility to check and seal every register, which were to be preserved in the communal archives.

The legal effects of the 1894 law lasted until 1918, when Transylvania became part of Romania.

In the case of Wallachia, there is no evidence of civil status registers before 1831, when the Organic Regulation (the first constitutional act of Wallachia) was adopted. Thus, following the dispositions of the Organic Regulations, the Orthodox parish priests from Wallachia began to constitute and administer parish registers under the supervision of their superiors. Each year, the Metropolitan see in Bucharest eventually collected all the registers with the help of archpriests and stored them in special rooms located in its palace near Dambovită river (Ungureanu, 1960, p. 32). The state administration did not interfere in this process until 1863, when the Office for Statistical Data of the United Principalities (created in 1859 within the Ministry of the Interior) was given the task to centralize the parish registers, thus replacing the Metropolitan see. Each year, the Office for Statistical Data (OSD) distributed to all communes a set of sealed and certified registers, three for recording births, marriages and deaths and one for recording wedding agreements, each in two copies (Ungureanu, 1960, p. 34). The city houses had to re-collect them until January 15th next year at the latest, so the parishes could no longer keep registers. The obligation to record the civil status data still remained with the priests from similar reasons as in Transylvania – lack of sufficient administrative personnel able to handle this task. Moreover, the priests were obliged to issue civil status certificates in standard format, such as birth certificates, which were countersigned by the local civil authority and the OSD officer. Marriages could not be consecrated without the special attestation of the church (in Romanian *peciuri*), which proved that there were no legal impediments to prevent their conclusion. Finally, each priest had to compile and deliver to the OSD a statistical report for his parish every quarter of the year.

In addition, the Law of Communes (1864) established that the mayors carried full legal responsibility regarding the administration of the civil status registers; the law was enforced by several instructions which detailed the procedures to be followed by the priests when filling the registers and

certificates. However, as the significant number of complaints preserved in archives suggests, not all the priests did fulfill these tasks with due rigor – there have been reported many cases when the civil status data were either not recorded correctly or many entries were corrected without proper legal caution. Finally, the Civil Code of the Romanian principalities, promulgated in December 1864, transferred the full responsibility of filling and administering the civil status registers to the civil authorities.

The provisions of the Civil Code were enforced and detailed in the instructions issued in 1866, which introduced new procedures regarding the civil status registers. There were instituted four types of registers: births, marriages, deaths and agreements of marriages. The civil status acts were compiled on separate sheets, which were then bound up at the end of each year, in the form of register, each in 2 copies: one for the communal archive, the other for the county law court. Moreover, while handwriting a civil status act the officer had to leave a blank margin on each page for quick reference, where the name(s) of the person(s) had to be written. This margin was also used for further legal annotations such as remarks on marriage, death (recorded on the birth certificates), divorces or the Romanian citizenship (recorded on the marriage certificates). These annotations were made by the civil status officers, who informed also the law courts to record them in their copy.

In 1911, a new instruction on the civil status registers was issued (in effect starting from January 1912) – it introduced a new type of register for births with pre-printed forms to be filled by the officers. However, this type of register was used only to record legitimate births; the foundlings continued to be recorded in the regular registers as before 1911.

Characteristics and significance of the collections

The *Bucharest civil status* collection includes 334 parish registers dating from the period 1832-1865 and 2095 civil status registers covering the period 1866-1912. Basically, there are 4 types of registers: those recording births, marriages, deaths and marriage agreements. From 1866 until 1884, the civil records of Bucharest follow the administrative organization of the city into 5 sectors designated with colors (red, blue, yellow, black and green). Therefore, in the Bucharest archives for each year in this interval there are at least five registers for every type of event and for each sector the civil status records are numbered starting from number 1 to the infinite. After 1881, the civil status data are recorded only under each year.

The civil registers from Bucharest collection, created especially since 1866, are rich in information not only on the vital data of the individuals, but also on their social and contextual milieu. If parish priests usually recorded data in a tabular form (see Figure II), introducing only the basic data i.e. name, surname, place of the event, name of the parents/godparents, witnesses etc., the civil officers recorded also additional data such as social status, profession, age, religion, address etc. of the parents of the newborn child or of the newlywed grooms, not to mention the witnesses of the event. The narrative style of most civil records before 1912, even if it uses a typical administrative language, adds more local flavor, turning the civil status certificates into veritable microstories, of particular interest not only for the genealogists, but also for those interested in the social history of the epoch (Pavel, 2009, p. 547-560).

The Brasov county civil status collection holds 669 parish registers dating from 1639-1895, and 1570 civil registers from 1895-1968, all organized by communes (Popovici, 2005, p. 66). In the case of parish registers (*libri parochiales, felekezeti anyakönyvek, Kirchenbücher*) (Moldovan, 1968, p. 162), there were in use 6 types, as following:

1. matricula (recording births, marriages, deaths);
2. marriage agreement registers;
3. confirmation registers;
4. registers containing data referring to the number and status of the parishioners;
5. family registers;
6. ordination registers.

Archived in the Brasov civil status collection, one can also find the minutes of the parish meetings concerning the status of parish population, the material contributions of the parishioners for their churches, ecclesiastic ordinances, lists with children enrolled in the parish or local schools, names of their teachers, data concerning vaccinations etc. Parish registers are sometimes filled with marginal annotations on various topics such as local weather, political and social events, epidemics and other news (Boar, 1988, p. 154-8).

The importance of the civil status collections for researches in various disciplines is beyond any doubt. Demography, genealogy, sociology, toponymy, onomatology, epidemiology, local history or family history etc.

are among the most privileged (Brie, 2010, p. 164-8; Bodale, 2008, p. 52-5; Bolovan S., Bolovan I., 1995, pp. 47-51). As in the case of the archives in Western Europe and elsewhere, an increasing number of amateur historians and genealogists arrive in the reading rooms of the Romanian archives in search for the history of their ancestors and the civil status collection represents their first choice. As historians often said, civil registers are in most cases the only documents which can shed a light on the intimacy of an ordinary individual that lived in the past (Brie, 2010, p. 193).

There are several differences between parish registers and civil registers especially regarding the scope of vital data recording. The first difference lies in *the balance between individual, community and territory*. Parish registers usually cover very imperfectly a given territory – this is due to their scope focused mostly on the community of a parish (and parishes in Transylvania or Wallachia did not have precise limits on the ground). Therefore, in some cases, civil status events were recorded with great delay or even escaped recording due to the great distances between the parochial office and certain areas of a parish, especially in the rural areas. On the contrary, the scope of the civil registers is mainly *territorial*: the registers include recordings referring to individuals living in a certain territory. Provided that all civil registers from a given region would have been preserved, one can reconstruct the main stages of life for all individuals living in that particular region. However, almost no sense of community can be retraced in these documents simply because the individuals recorded in the civil registers usually came from different social and professional milieus and shared different religious affiliations.

A second difference lies in the *religious vs. secular dichotomy*. The parish registers bear a clear confessional scope, recording only those vital events that involved individuals sharing the same confession (or religion) and which were consecrated by the church and marked by the customary religious rituals. Given the fact that during the 19th century, before civil registers were established, the number of individuals who abandoned any religious affiliation becomes significant and their life events were hardly recorded in the parish registers. On the contrary, in the civil registers there are recorded persons no matter their religious affiliation.

A third difference is given by the *thoroughness of civil data recording*, more likely to be met by the civil registers than the parish registers. This is explicable by the fact that, unlike in the case of civil status offices, whose

main task was precisely the recording of civil data, the parish priests had many ecclesiastic duties among which keeping parish registers was one of the least important. Moreover, the parish priests were controlled not by state officials, but only by their ecclesiastic superiors, who proved to be more reluctant in chastising for errors or negligence in keeping parish registers than for other transgressions.

Lastly, in the two types of registers the information is organized differently – while the parish registers record births, marriages and deaths in a tabular format with a minimum amount of data, the civil registers contain the full transcription of the certificates issued for each type of event.

In the reading room of the Bucharest Municipal Archives, there are also available finding aids (in register format, handwritten), which offer a basic by-name indexing for births and marriages corresponding to each year from 1866 until 1912 (see Figure III). In the case of death certificates, the available finding aids are incomplete and contain many errors, thus making them practically useless. For Brasov collection, there are no available indexing aids with the exception of a basic inventory containing a short description for each register.

Project summary

The project team realized the complete digitization and basic indexation of Brasov collection, the complete digitization of the finding aids, the complete digitization of the registers from the years 1866-1912 and the complete by-name indexation of births and marriages for Bucharest collection, and finally the configuration of the database and the project website (<http://www.arhgenvirt.ro>). Since September 2017, more than 500.000 images in total are available on-line for free consultation.

The digitization of the registers has been carried using compact cameras with 18 Mpx resolution mounted on photo stands (due to the limited financial resources, the acquisition of large format book scanners was not possible). However, the quality of the images realized meets the standard requirements for public access, the photographic reproductions of all registers being easily readable. In the case of Brasov collection, the registers were photographed as such, with two pages per frame, while in the case of Bucharest collection the registers were photographed piece by piece. In some cases, the digitization process was extremely difficult due to the poor state of conservation of several registers; in very few cases, the registers could not be photographed.

The indexation of the metadata was completed by manually introducing the relevant data from the available finding aids into Excel files. For Bucharest collection, the completion of indexation by name required not only introducing the data, but also double checking them as in the case of the names (i.e. Romanian, German, Jewish etc. spelled in various orthographies). In the case of Brasov collection, in the absence of finding aids, a similar indexation was not possible within the project as it required considerable human resources i.e. several dozens of employees able to read various languages and scripts and to collect all the metadata directly from registers in a record time. In order to accomplish this, a far more generous funding was needed than it was granted.

The database was structured using the MySQL 5.7 programming language, while the web application uses a Symfony framework 3.4 (with PHP 7.0) suitable for long term projects. The indexation of metadata was carried using Elasticsearch 4, which significantly reduced the time for returning the results of queries. The files and all the software package were installed on an Apache 2.2x server.

Among the main functionalities and options provided for all users, we can list:

- a) search by name option and the visualization of the civil status record associated for each name (available for Bucharest collection);
- b) page by page visualization of each register (available for Brasov collection) with search options by actual locality, parish and year(s);
- c) personal working space available for each user, where "favorites" (records, links to images, personal comments etc.) can be listed and stored for further use;
- d) an assistant application for creating one's own genealogical tree.

In the web application there has been implemented the Boolean search for narrowing the search results.

In order to use the database, each user has to create a personal account in order to register and to use the available working space, whose utility is of primary importance as the images cannot be downloaded or printed due to copyright infringement. As the National Archives are the legal owner of the copyright, it is only them which can grant the permission to download or print images taken from archival documents which are part of the National Archival Fund.

Difficulties and challenges in implementing a digitization project in Romania

In pursuing the necessary activities in order to accomplish the expected results, the project team faced several difficulties mainly related to limited financial resources and precarious research and administrative infrastructure. The limitations of the project budget became clearer when the initial estimations regarding workload had to be drastically adjusted and a proper balance between resources and expected outcomes had to be continuously and carefully checked. For example, acquisition of scanners had been abandoned due to prohibitive costs for much cheaper compact cameras. The metadata indexing had to be limited only to those which can be collected from the available finding aids. Finding a common language with IT specialists was not an easy task and many discussions and exchange of ideas were needed to harmonize the different approaches. Last but not least, poor administrative support (lack of sufficient personnel with competences in research projects working in support services) overburdened the management with additional tasks.

The main challenge for such projects is the long-term maintenance of the database which has to be supported from resources to be identified in the future.

In conclusion

Hopefully, the digitizing of the civil status collections preserved in the Romanian archives will be continued in the future years, as it represents the only solution to safeguard an important part of the national archival patrimony. The *Virtual Genealogical Archive* project intended to open the path for similar endeavors and to show a possible way to reach the expected results.

Acknowledgments. *The results of our research have been made possible through the project named "The Virtual Genealogical Archive - a pilot project destined for the National Archives of Romania and third party users" and co-financed through the research programme "Partnerships - Collaborative Projects for Applied Research - PCCA 2013" by UEFISCDI.*

Bibliography

- Boar Liviu (1988), 'Colecția registrelor parohiale de Stare Civilă', in *Îndrumător în Arhivele Statului. Județul Harghita*, București, pp. 154-158.
- Bodale Arcadie M. (2008), 'Colecția de stare civilă – între realizări și deziderate. Actele comunale de stare civilă de la D.J.A.N. Iași', *Revista Arhivelor-Archives Review*, 1, pp. 51-83.
- Bolovan Sorina, Bolovan Ioan (1995), 'Registrele parohiale de stare civilă din Transilvania - izvoare de demografie istorică', *Revista arhivelor*, 1-2, pp. 47-51.
- Brie Mircea (2010), 'Registrele de stare civilă din Transilvania în a doua jumătate a secolului al XIX-lea. Semnificație documentară', in Dan Octavian Cepraga, Sorin Șipoș (eds.), *Interpretazioni del documento storico: valore documentario e dimensioni letterarie*. Oradea: Ed. Universității din Oradea, pp. 159-182.
- Delsalle Paul (2009), *Histoires de familles, les registres paroissiaux et d'état civil, du Moyen Âge à nos jours, démographie et généalogie*. Besançon: Presses universitaires de Franche-Comté.
- Moldovan Liviu (1958), 'Registrele confesionale de stare civilă din Transilvania', *Revista Arhivelor*, 1, 163 -174.
- Pavel Mihai Alin (2009), 'Colecția stare civilă din București – izvor de informații genealogice. Registrele de căsătorii', *Revista istorică*, 5-6, pp. 547-560.
- Popovici Bogdan-Florin (2005), 'Considerații pe marginea prelucrării colecției de registre de stare civilă', *Buletin de informare și documentare arhivistică*, pp. 65-70.
- Ungureanu Gh. (1960), 'Actele de stare civilă sub regimul Codului civil. III', *Revista Arhivelor*, 1, pp. 31-64.

Figure I – A death certificate recorded in 1867 from a civil register (Bucharest collection)

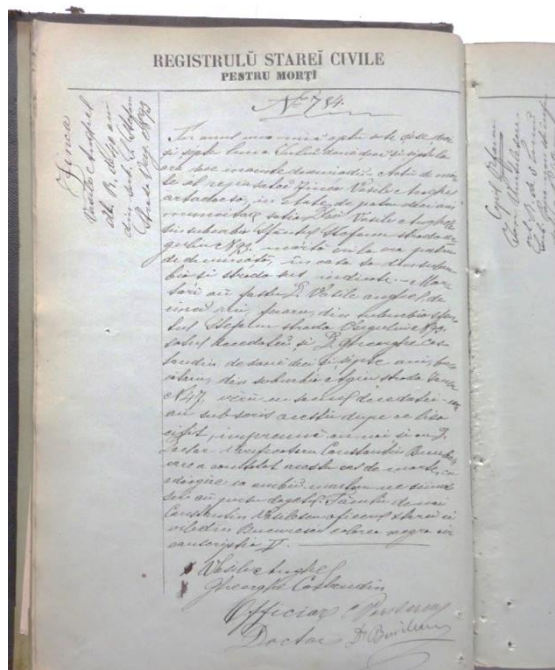


Figure II – Page from a parish register
of the Orthodox church in Braşov (1819)

[illegible]

Figure III – page from a finding aid with data regarding marriages from 1910 (Bucharest collection)

LISTA ALFABETICA
de actele înscrise în registrele de CĂSĂTORII ale Statului Popular și Războiului
Oficiul de Stat civil, în cursul anului 1910

Nr.	Actului	Vol.	Numele și prenumele căsătoritilor		Data actului		Observații
					Lucru	Zis	
930	6		Babu Traian	Anghel Iovășcu			
932	4		Bănel Avram	Imel Buta			
950	4		Băsan Avram	Iurcan Sima			
959	5		Bălcăș Nicol	Iușcă Iovășcu			
962	6		Bărbulescu Gheorghe	Păcuru Sofia			
981	5		Băran Gheorghe	Schneidmiller Iosif			
987	4		Bănuș Vasile	Thier Bertha			
1025	4		Bănuș Ilie	Tufescu Iovășcu			
1041	4		Bănuș Mihail	Mihail Iovășcu			
1050	6		Bănușescu Gheorghe	Vădu Iovășcu			
1091	5		Bănuș Gheorghe	Wroblewski Iovășcu			
1104	6		Bănuș Iosif	Bănuș Gheorghe			
1113	5		Bănuș Pașcu	Bănuș Iovășcu			
1154	6		Bănuș Iosif	Bănuș Iovășcu			
1158	4		Bănuș Nicol	Bănuș Iovășcu			
1162	4		Bănuș Nicol	Bănuș Iovășcu			
1174	1		Bănuș Nicol	Bănuș Iovășcu			
1176	4		Bănuș Nicol	Bănuș Iovășcu			
1181	5		Bănuș Nicol	Bănuș Iovășcu			
1201	4		Bănuș Nicol	Bănuș Iovășcu			
1203	4		Bănuș Nicol	Bănuș Iovășcu			
1205	4		Bănuș Nicol	Bănuș Iovășcu			
1254	6		Bănuș Nicol	Bănuș Iovășcu			
1267	5		Bănuș Nicol	Bănuș Iovășcu			
1278	6		Bănuș Nicol	Bănuș Iovășcu			
1309	5		Bănuș Nicol	Bănuș Iovășcu			
1316	6		Bănuș Nicol	Bănuș Iovășcu			
1318	4		Bănuș Nicol	Bănuș Iovășcu			
1331	7		Bănuș Nicol	Bănuș Iovășcu			
1369	4		Bănuș Nicol	Bănuș Iovășcu			
1390	8		Bănuș Nicol	Bănuș Iovășcu			
1403	7		Bănuș Nicol	Bănuș Iovășcu			
1421	8		Bănuș Nicol	Bănuș Iovășcu			
1481	7		Bănuș Nicol	Bănuș Iovășcu			

The Special Fund deposit of the National History Museum of Romania - a history of the Ceausescu era in gifts received from abroad

Cristina Păiușan-Nuică

(researcher at the National Museum of History of Romania)

How the History Museum of the Socialist Republic of Romania appeared?

The History Museum of the Socialist Republic of Romania was inaugurated on May 8, 1972, with a one year delay. Nicolae Ceausescu, the leader of the Romanian Communist Party, had wanted its inauguration to take place on May 8, 1971, on the 50th anniversary of the founding of the Romanian Communist Party (Ilie, 2012, p.188-200).

The most important pieces were collected from museums all over the country, especially prehistoric and ancient treasures, thus forming the patrimony of the largest Romanian museum.

For the communist regime, the national history and the museums were simple means of propaganda, made programmatically by applying the documents adopted by the leadership of the Romanian Communist Party, acts that outlined the ideological and educational orientation of museums in Romania.

The museum themes were related to communist propaganda true history and art, so the permanent and temporary exhibitions, which were often traveling through the country's cities, had to reflect important moments in national history, the past of the R.C.P. and the revolutionary actions of party and state leaders.

The grid of communist propaganda and the "struggle of the Romanian people against external enemies and internal exploiters" was applied to all eras and exhibitions, from prehistory to the revolution of 1848, from the

founding of the Romanian Communist Party (May 8, 1921) to its role in the daily life of Romanians.

Shortly after Nicolae Ceausescu came to power (March 1965), museums began to reflect especially decisions taken at the IX Congress of the R.C.P. (July 1965), as well as speeches made by Comrade Nicolae Ceausescu on the 45th anniversary of the R.C.P. (May 8 1966) (Ilie, 2013, p. 198).

The Agitation and Propaganda Section of the Central Committee of the Romanian Communist Party, together with the Council of Socialist Culture and Education, also controlled the ideological correctness of museum exhibitions until the "July theses" 1971, inspired by the Romanian leader's visit to North Korea.

Apparently, the theses naturally continued the ideological line of the party, but Ceausescu stressed that the party should have from that moment an essential role in all cultural activities carried out in the country, implicitly in museums. A cultural mini-revolution followed which gradually removed the opening to the West, which began timidly at the end of the Dej era and in the early years of the Ceausescu era, from 1960, until 1971.

Analyzed after more than half a century, the mini-revolution brought by the "theses of July" represented a regression, a return to the ideologisation on the Stalinist pattern of the '50s. But the Stalinist template was transposed into Romanian content and so an ideology emerged –ceaușismul.

"Ceaușismul" was a type of Stalinist adapted to the Romanian state, focusing on Nicolae Ceausescu and his ideas on the development of the "new man" and "the multilaterally developed socialist society".

Institutionally, in September 1971, a new ideological control institution appeared called the Council of Socialist Culture and Education (CSCE), which coordinated and guided the cultural activity in Romania. The president of the institution was a member of the government.

The Museums Directorate, which dealt with the transposition of party ideology into history and museum themes, was coordinated and controlled directly by the CSCE.

In this historical context, the History Museum of the Socialist Republic of Romania appeared, after many failed attempts started by the popular democratic regime in 1954 (Ilie, C 2013 174-288). Since 1968, the thematic structure of the new museum has been built, then, starting in the spring of 1970, the museum's collection was assembled by transferring the most important pieces from museums across the country.

The most important historians participated in structuring the theme of the Socialist Republic of Romania Museum.

The museum was to be inaugurated on May 8, 1971, when the Romanian Communist Party celebrated its 50th anniversary, which was not possible because the transformation of the historic building - the ancient building of Central Post Office (built at the end of XIX century) into a museum became difficult, and the collection of the museum treasure also. The museum was opened one year later.

A grand ceremony was organized on May 8, 1972. The inaugural ribbon was cut by N. Ceausescu and then entered the museum collection, along with the scissors used, being part of the museum exhibits today.

After 6 years from its inauguration, the most important museum of the republic and the main instrument of historical propaganda became an important center for the propagation of the personality cult of Nicolae and Elena Ceaușescu.

The museum was the result of propaganda work, as stated by Mihnea Gheorghiu, president of the Academy of Social and Political Sciences of Romania, at the opening of the Museographers' Lectures Session in 1976. "The museum is also a propaganda work, this is too often seen, a propaganda of course differentiated, suggestive, intelligent, full of initiatives, in a word - passionate" (Gheorghiu, M., 1976, p.4).

One of the themes of the 1976 session was "Museums and the man of the multilaterally developed socialist society" which aimed to promote and strengthen the role of Romanian museography which "museography contributes to the development of socialist consciousness as an essential factor in building society and, above all, to educating the young generations in the spirit of socialist humanism and love of country " (Gheorghiu, M., 1976, p. 4-6).

Florian Georgescu¹, the director of the museum of those years, presented the lecture "Museum of History in the light of the documents of the XI Congress of the Romanian Communist Party" (Georgescu, F., 1976, 7-11) in which he presents the way in which the historical truth was established. The fundamental theoretical, ideological and political charter of the party, achieves a masterful synthesis of the life and millennial struggle

¹ Florian Georgescu (1924-1997) professor, director of the History Museum of the Socialist Republic of Romania (1971-1984).

of the Romanian people, of the activity of the popular masses, of the progressive forces, of the revolutionary movement.

These pages of high and innovative theoretical, scientifically grounded and clarified, basic issues of our history are approached in these pages. The formulated theses are also for the history museums a brilliant guide of their theoretical and practical activity. For the history museum has the high ideological and cultural-educational mission to present the entire history of the Romanian people, which must be presented as a fresco of the incessant class struggles, of the battles fought by the popular masses for freedom and social justice, for the defense of the national being and independence, for progress and civilization" (Georgescu, F., 1976 7).

How the Special Fund deposit appeared?

From the moment of taking power, in March 1965, N. Ceaușescu, Secretary General of the Romanian Communist Party, received, according to the official protocol, symbolic gifts on the occasion of foreign visits, offering the same type of souvenirs from Romania to the officials visited.

Working visits to various areas of the country, urban or rural, were another source of gifts and souvenirs received by the Romanian leader. The anniversary of the birthday of the "Comrade" (January 26), of the Communist Party (May 8), the day of "liberation from fascist yoke" August 23, Republic Day (December 30), and many other events, are just as many occasions to offer gifts to the leader and then to the presidential couple. Over time, these occasions multiplied, being accompanied by the anniversary of "Comrade Academician, doctor, engineer Elena Ceausescu" (January 7).

Most of the gifts received were taken over by the Party Household and used at various temporary thematic exhibitions dedicated to "Comrade Secretary General of the R.C.P." There were also exceptions if the object or objects received were to the liking of the Ceausescu couple, then they were directed to one of the protocol houses where they lived permanently or temporarily.

Starting with the second half of 1977, the party and the state, as well as "the entire Romanian people" were preparing for the grand anniversary of the comrade's 60th birthday.

Among the anniversary events, a major one, carefully coordinated and supervised by senior party officials, was the exhibition "Evidence of love, high esteem and deep appreciation enjoyed by Comrade Nicolae Ceausescu and

Comrade Elena Ceausescu, of extensive friendships and collaboration between the Romanian people and the peoples of other countries" within the National History Museum in Bucharest.(Oanță-Marghitu, S., 2018 329-354)

The exhibition was opened on January 23, 1978, on the occasion of "Comrade Nicolae Ceausescu's 60th birthday and over 45 years of uninterrupted revolutionary activity."

The objects received by the Ceausescu couple were handed over by the Party House to the museum, being reinvented, sorted, divided into two main categories: external gifts and internal gifts and then on many other subcategories and displayed in a festive way.

Numerous articles have been written about the exhibition, including that of Nicolae Petrescu published in 1980, citing previously published articles, entitled "Honoring the Personality of Comrade Nicolae Ceausescu and Comrade Elena Ceausescu by working people in Vaslui County illustrated in the homage exhibitions by at the National History Museum of Romania" in which it was said: "By the large number of valuable objects from all areas of the country and parts of the world, donated by central party and state institutions, collectives or workers in industry or agriculture, science and art, of great political personalities from different states or international organizations, the exhibition represents, as it was so beautifully characterized, a "comprehensive geography of feelings of warmth and natural gratitude" for President Nicolae Ceausescu and Comrade Elena Ceausescu" (Petrescu, N., II/1980 11-18).

Thus, these objects received as a gift by the couple Nicolae and Elena Ceaușescu quickly became museum pieces, being exhibited until December 1989, in the permanent exhibition of the History Museum of the Socialist Republic of Romania, briefly called "Homage".

Gradually the History Museum of the S.R.R. it was transformed from an instrument of communist and party propaganda, into an important mechanism in the evolution of the personality cult of Nicolae Ceaușescu and his wife.

The "Homage" increased from year to year, the original space was expanded, reaching in the '80s to 2,500 square meters divided into ten rooms. The pieces, which arrived at the museum and then in this exhibition, narrate the way the Romanian leader was viewed in the country and the way in which the propaganda apparatus created an international allure that far exceeded the real role that the Romanian leader had internationally.

The digitization of the Special Fund of the National History Museum of Romania, consisting of the former exhibition "Homage" and the pieces intended for the exhibition, but which for various reasons were not part of it, represents one of the MHMR projects started in 2020, continued in 2021 and which will be available online in a future MHMR online project.

The cultural goods went through the process of identification, inventory, research of provenance, components, content, material, manufacture, the year in which they were made and offered to the Ceausescu couple and were divided into categories.

Researching the pieces of the fund, I realized that with their help I can write a chronological history of Romania's international relations and that each piece received from abroad has a significance and points to a bilateral diplomatic event.

From the first exhibition in January 1978 until November-December 1989, at the 14th Congress of the Romanian Communist Party, the number of exhibited pieces increased rapidly, with new ones being added every year. The transfers from the Party Household, a section of the Central Committee of the R.C.P., which collected all the gifts received by the Ceaușescu couple, took place several times in the years 1978-1989 - the years in which "Omagiala" - became the basic exhibition of the museum, it has grown and diversified its types of objects.

At the end of the 80's, the Ceaușescu couple called the National Museum of History "our museum", 10 rooms in the exhibition space being occupied with the glorification of "The hero among heroes Nicolae Ceaușescu" (Ilie, C., 2013 216-221).

The exposure of these objects became more diverse mainly due to the increase in the number of objects that had to be displayed, but also to the diversity of types.

With the help of these objects, gifts received from various countries of the world, the official propaganda created an overwhelming impression on the visitor - convinced of the world-wide allure of Nicolae Ceaușescu's personality.

30 years after the end of the Ceausescu regime - the study of the Special Fund is a necessity for understanding: communist and Ceausescu symbolism, the evolution of the personality cult, the way the Romanian leader was viewed by the leaders of other states, but also the ways of building the myth of Nicolae Ceaușescu.

In 1978 the foreign pieces entered in the exhibition "Evidence of love, high esteem and deep appreciation enjoyed by President Nicolae Ceausescu and Comrade Elena Ceausescu, of the broad relations of friendship and collaboration between the Romanian people and the peoples of other countries" were the ones received on the occasion external visits or on the occasion of receiving foreign delegations visiting Bucharest.

The objects received from abroad benefited from an extended space, being displayed at the beginning simply, respecting the geographical areas from which the Romanian leader received them, but gradually an elaborate context was built, and "deep appreciation" and "wide relations of friendship and collaboration between the people Romanian and the peoples of other countries" were transformed into tributes to "the fighter for world peace" and "to the most known and appreciated diplomat of Romania".

In those years, Romanians were regularly brought to the museum, these visits being part of the ideological program of each state institution, schools, factories and enterprises.

They did not know that, within the internationally agreed diplomatic protocol, the exchange of gifts and souvenirs from that country, is a well-regulated custom, these gifts accompanying all official visits and sometimes even informal ones.

The State Protocol deals with the purchase of objects representative of the country, without great value, well-crafted and finished, beautifully packaged and with specific insignia: the colors of the national flag, the coat of arms of that country, inscribed with the country's name, sometimes personalized offered. These protocol gifts generally had texts in an international language.

Traditional consumer products for the country (handicrafts made of shells, cigarettes) were often offered in a special package. In the case of the Romanian state were offered: bottles of wine and spicy drinks: brandy, țuică, palincă.

The customs of protocol made every external visit, every reception of officials, delegations or even people to be accompanied by this exchange of gifts.

These gifts received according to diplomatic customs by the Ceausescu couple received a propaganda load after their first major exhibition, on January 23, 1978, at the History Museum of the Romania.

Most of these goods were in the diplomatic custom, others not, being some precious objects, others out of the ordinary such as: elephant fangs,

furs of wild animals, weapons captured from enemies and offered as an offering to the Romanian communist leader, but also ritual objects which brought health and good luck to the wearer.

A typology of the pieces from the Special Fund is difficult to make, their variety being large, but we can divide the princes from abroad into several categories:

- I. **Common items** given as gifts by heads of state and government around the world, official delegations: office sets, pens, wine bottles, cigarette boxes, traditional table sets (tablecloths and napkins), sets of glasses (crystal, glass, plastic, wood, ceramics inscribed with the flag of that country, possibly with special dedications), plaques and medals marking historical events, flags of that country, souvenirs with images of that country, records and sets of records with traditional music, jewelry boxes, etc.
- II. **Art pieces:** paintings, sculptures with an ideological message or representing monuments from the respective country. Most of the gifts given by the neighbors in the communist countries are statues of some heroes of the communist movement from Marx, Engels, Lenin, Stalin, to local heroes.
- III. **Personalized gifts** offered by various businessmen, factory managers, who wanted to promote their products and conclude contracts with the Romanian state, capturing through these gifts the benevolence of the Ceausescu couple.
- IV. **Clothing or leather goods offered to Elena Ceaușescu** representing the fashion of the respective countries: numerous coupons made of natural silk, cotton, linen, wool, tercot (China, North Korea, African countries, India), sun hats, slippers beach (especially from Guinea and Liberia), clogs; traditional handbags made of straw, rushes, fabrics or elegant skins of rare wild animal skins (snake, leopard, crocodile)
- V. **Jewelry offered to Elena Ceaușescu**, mostly traditional from their countries of origin, other valuable: brooches, earrings, bracelets, bead necklaces, shells or various other materials
- VI. **Clothing gifts offered to Nicolae Ceaușescu:** hats, slippers, ties, scarves, cigarette boxes with dedications, shoes traditional or snakeskin shoes, crocodile, etc.

- VII. **Models of machines, planes, installations or equipment** - part of a marketing policy or achievements of communist industries (mainly USSR, but also from Poland, China, Yugoslavia, Hungary, Bulgaria).
- VIII. **Impactful gifts** offered by countries that wanted to obtain food, sanitary, oil aid, preferential loans with low interest rates or non-reimbursable material aid from Nicolae Ceaușescu, involving him in mediating conflicts.
- The most important in this category are the two American-made weapons captured by Vietnamese soldiers and offered to the Romanian leader, but also a piece of the remains of the 1000th American F-105 D aircraft, shot down at 29 April 1966, in the province of Bac Thai - RD Vietnam, offered to Nicolae Ceausescu by Ho Si Min on May 10, 1966.
- IX. **Orders, medals** according to the international protocol established for visits at the level of state president offered by states around the world
- X. **Doctor Honoris Causa diplomas** from universities around the world, offered to both Nicolae Ceaușescu and especially Elena Ceaușescu, "world-renowned scientist", as well as other academic and university degrees
- XI. **Photos of the Ceausescu couple** from previous visits in that country
- XII. **Plush toys** (Australia, Austria), children's or board games, miniature costumes of some states.

Most of the 5,000 pieces in the Special Fund can be integrated into one of the above categories. A study of the ten categories will present a complete history of how propaganda has transformed diplomatic customs into "evidence of high international price" and the marketing policy of foreign firms, in "recognizing the international role of Comrade President of the Socialist Republic of Romania".

The official propaganda of the communist regime gradually began to overlap with that of the person Nicolae Ceaușescu, who had gradually come to the fore, building a kind of synonymy between Ceaușescu, the Romanian Communist Party and the romanian state.

The two common slogans of the 80s were: "Communist Party-Ceausescu-Romania" and "Ceausescu-Peace", were the leitmotifs of any events and publications.

In the period 1978-1989, the image of Nicolae Ceaușescu became the most exported product of Romania abroad, in the 80's the official propaganda practically reducing to his person, the external image of the Romanian state.

The cultural assets from the Special Fund of MHMR were inventoried chronologically, keeping the date of their receipt and the country of origin.

The exhibition of the pieces was initially made in an amalgam of exhibits from the country and those from abroad. At the opening in 1978, the amalgam was deliberately created to create the impression of grandiosity.

A multitude of objects, some bright, others exotic and unheard of in Romania, inscribed in different languages of the world, giving the impression of an altar dedicated to "the most beloved son of the people."

The purpose of this article is to highlight a project started at the National Museum of History of Romania, a research that is based on the analysis of the museum heritage from the communist period.

The Special Fund, the pieces received from various countries, would be the basis of a paper on Romania's foreign policy in gifts and medals, and the pieces received from all over the country would be the basis of a paper on the staging of the Ceausescu personality cult with the help of anniversary and other gifts types.

Acknowledgement

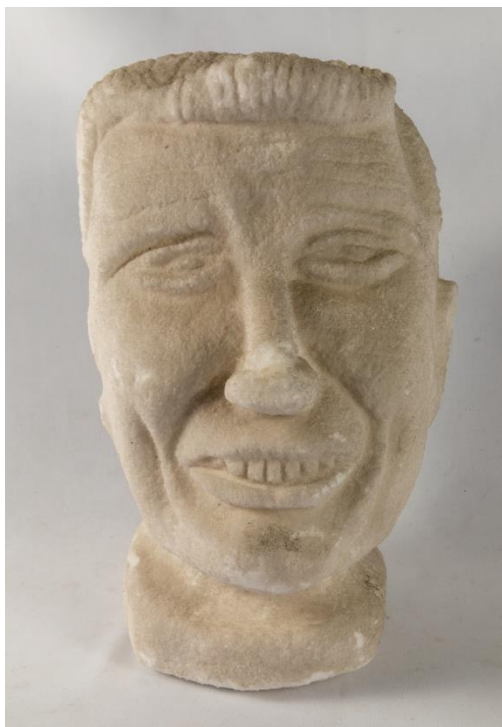
This text is part of a larger study being published in Romanian that deals with the entire Special Fund, both objects received from all over Romania and those received from abroad.

Bibliography

- Georgescu, F., 2/1976 – Florin Georgescu, "History Museum in the light of the documents of the XI Congress of the Romanian Communist Party", Muzeul Național (National Museum), 2/1976, p. 7-12.
- Gheorghiu, M. 2/1976 – Mihnea Gheorghiu, 1976, "Obiectivele actuale ale activității muzeale" (*The current objectives of the museum activity*), Muzeul Național (National Museum), 2/1976, p. 3-6.
- Ilie, C 2013 – Cornel-Constantin Ilie, 2013, *Regimul comunist și muzeele de istorie din România* (*The communist regime and the Romanian history museums*), Dobrogea Publishing House, Constanța
- Oanță-Marghitu, S., 2018 – Sorin Oanță-Marghitu, "Un regim al schimburilor de daruri" ("A regime of gift exchanges"), Muzeul Național (National Museum), 30/2018, p. 329-354.
- Petrescu, N., II/1980 – Nicolae Petrecu, „Cinstirea personalității tovarășului Nicolae Ceaușescu și a tovarăsei Elena Ceaușescu de către oamenii muncii din județul Vaslui ilustrată în expozițiile omagiale de la Muzeul Național de Istorie a României” ("Honoring the personality of comrade Nicolae Ceaușescu and comrade Elena Ceaușescu by the working people from Vaslui county illustrated in the Homage exhibitions at the National History Museum of Romania"), *Acta Moldaviae Meridionalis. Anuarul Muzeului Județean Vaslui*, nr. II/1980, p. 11-18.



Bronze statuette offered by the President of Mexico, Luis Echeveria, 9.VI. 1975



Statuette made of salt received from the Netherlands



Model of the Moon Capsule offered by Eugen Cernan, commander of the spacecraft "Apollo 17" on the occasion of the visit to the Romania, November 1974



Model of a red sail ship received from the Soviet Union, 1978



Model "Chevrolet" received by Nicolae Ceausescu from the only Romanian Chevrolet dealer, USA, 1973



Two traditional napkins offered by President Marcos to Elena Ceaușescu during his visit to the Philippines in April 1975



Traditional statuette offered to Nicolae Ceausescu by the President of the Republic of Benin, Mathieu Kerekou, Bucharest, 22. VII.1976



Model of a traditional weapon received from Zimbabwe, July 1983



This publication is part of the project “Heritage in the Digital Age. Studies and Best Practices in Romania.”, coordinated by the Romanian National Commission for UNESCO

