Digital Revival of the Bulgarica Collection of the Central Library of the Bulgarian Academy of Sciences

Maxim Goynov[1], Detelin Luchev[1][0000-0003-0926-5796],
Desislava Paneva-Marinova[1][0000-0001-5998-687X], Silvia Najdenova[2],
Lubomir Zlatkov[1], Lilia Pavlova[3], Evita Pilege[4][0000-0002-6676-0526]

[1] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences,
8, Acad. Georgi Bonchev Str., Sofia, Bulgaria
[2] Central Library of the Bulgarian Academy of Sciences, 1, 15-ti Noemvri Str., Sofia, Bulgaria
[3] Laboratory of Telematics, Bulgarian Academy of Sciences,
8, Acad. Georgi Bonchev Str., Sofia, Bulgaria
[4] Latvian College of Culture at Latvian Academy of Culture, 57, Bruninieku Str., Riga, Latvia
goynov@gmail.com, dml@math.bas.bg, dessi@cc.bas.bg,
silvia_najdenova@abv.bg, lyubcho@gmail.com,
pavlova.lilia@gmail.com, evita.pilege@gmail.com

Abstract. This paper presents the development of a digital library, created for the needs of the Central Library of the Bulgarian Academy of Sciences. Based on previous work by the team from the Institute of Mathematics and Informatics at the Bulgarian Academy of Sciences, this new environment exemplifies the employment of state-of-the-art digital technologies for the presentation of scientific literary heritage, in this case the Bulgarica Collection – the advancements of the researchers in the field of Bulgarian studies in the 1980s. The development is also a component of the research e-infrastructure CLADA-BG.

Keywords: Digital Libraries, Bulgarica Collection, Scientific Literacy Heritage.

1 Introduction

The scientific literary heritage, stored and conserved through the years in libraries, cultural centers, museums, cultural institutions, and the Bulgarian Academy of Sciences, among others, is an exceptionally valuable national treasure. As such, it is of vital importance not only to cherish and protect it, but also to share and disseminate it, so that the rich Bulgarian scientific traditions can serve as an inspiration for advances in education and future research work in particular, and the evolution of Bulgarian culture and society in general. The ongoing developments in the field of information technologies allow for the creation of efficient content management systems, such as digital repositories, electronic libraries, *etc.*, which in turn could assist in achieving those purposes by providing an access to the scientific heritage to a much wider audience than the traditional institutions.

Digital Presentation and Preservation of Cultural and Scientific Heritage. Conference Proceedings. Vol. 13, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS, 2023. ISSN: 1314-4006, eISSN: 2535-0366

This paper presents the development of the digital library (DL) for noteworthy scientific literary heritage, created for the needs of the Central Library of the Bulgarian Academy of Sciences (CL-BAS). For its development, we have employed a state-of-the-art web-based software platform for intelligent digital management and presentation of large databases and knowledge in the fields of culture, humanities and social science. The storage, extraction, and curation of homogenous and heterogenous objects, and the responsive and efficient access and management of the resources are of particular significance. A wide range of documents and formats for digital informational content are supported, allowing for rich interactional functionalities. Functional components such as modules for management and presentation of the objects, modules for metadata management, and administrative services are provided. The platform could be efficiently utilized for efficient digitization in scientific, cultural, and public institutions, e.g., libraries, museums, galleries, collections, archives, *etc.* In cooperation with the National library "Ivan Vazov" – Plovdiv, the Regional library "Peyo Yavorov" – Burgas, and the Central Library of the Bulgarian Academy of Sciences, we have developed experimental implementations of the platform.

2 Materials and Methods

This section presents the primary materials, methods, and tools, used during the development of the digital library for noteworthy scientific literary heritage, created for the needs of the Central Library of the Bulgarian Academy of Sciences.

2.1 Core Environment

The environment for storage, extraction, and curation of data from different fields of the Bulgarian cultural and historical heritage is a web-based platform, generating cultural content management systems such as digital libraries, virtual museums, cultural content repositories and archives. The development is an infrastructure component of the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favour of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH (CLaDA-BG), which ensures quality and adequate curation and management of digitized cultural heritage of the Bulgarian lands from Antiquity to the Modern Age. The collections of specimens are provided with elaborate knowledge of the resources (*i.e.*, metadata), facilitating their digital presentation, storage, and processing (Bulgarian Ministry of Education and Science, 2021).

Fundamental for our development methodology (see figure 1) is the separation of the development processes in terms of core and custom functionalities, and back-end and front-end features. Depending on the usability of the implemented feature, it is classified as a core or a custom one, and is included in the respective module of the platform. Thus, more DLs can benefit from a feature implemented for a specific DL.

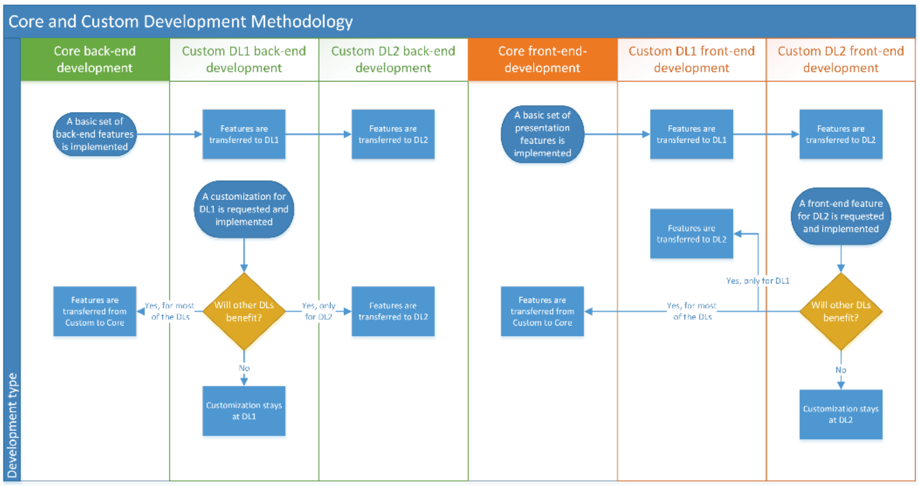

Core and Custom Development Methodology
Development type
Core back-end development
Custom DL1 back-end development
Custom DL2 back-end development
Core front-end-development
Custom DL1 front-end development
Custom DL2 front-end development
A basic set of back-end features is implemented
Features are transferred to DL1
Features are transferred to DL2
A customization for DL1 is requested and implemented
Will other DLs benefit?
Yes, for most of the DLs
Features are transferred from Custom to Core
Yes, only for DL2
Features are transferred to DL2
No
Customization stays at DL1
A basic set of presentation features is implemented
Features are transferred to DL1
Features are transferred to DL2
A front-end feature for DL2 is requested and implemented
Features are transferred to DL2
Yes, only for DL1
Will other DLs benefit?
Yes, for most of the DLs
Features are transferred from Custom to Core
No
Customization stays at DL2

Fig. 1. Core and Custom Development Methodology

The platform provides basic and advanced features in the following directions:

- Definition of data structures (schemes). Complex structures can be implemented easily using lightweight web interface. The structures can be modified and extended after the initial creation. Advanced options for custom validation, presentation, indexing, computations are provided for every element of the defined scheme. Dynamic permission management definition, based on user roles and content data is also available as a part of the data structures definition.
- Data input and management. The platform has a rich set of tools to facilitate data input and reduce the risks of errors during manual data input. Tools for automated translation, transliteration, and suggestions based on previous input are provided.
- Version management. The system keeps a copy of every object edit, thus providing an option to compare and revert changes to a specific version in case of wrong edit or deletion by mistake.
- Bulk import/export of flat/complex structured data. Multiple formats for bulk import/export of data are provided, including XLSX, CSV, JSON. A versatile import mechanism allows importing complex data even when using flat table formats like CSV or XLSX.
- Component based presentation layer. Depending on the requirements for the system, the user interface can be modified using the already implemented features, screens, views, design, *etc.* When a major change is required, a new component can be implemented as a part of the platform's custom presentation layer.
- Flexible search engine options, allowing modifying and overall replacement of search engines (using third party software or other customizations) in order to achieve optimal results for the specific application domain.
- Account management. By default, six roles are defined in the system, depending on the activities needed to support/use it:

- Guest user;
- Registered regular user;
- Editor;
- Model editor;
- Platform administrator;
- Database administrator.

The system allows users to have custom roles, and to classify them in groups in order to have different access for specific types of content.

- Content management, protection and indexing. The system default setup hides the original content and presents to the end user downsamples (optionally with watermarks). Indexing is available for text documents (like PDFs) allowing full-text search to be performed among all available text objects.
- Search engine optimization (SEO). The architecture we have used is based on the concept of single page application (SPA), which gives great experience to users and performance for every modern browser. Regardless of the fact that SPA is not well indexed and appreciated by search engines, the platform shows excellent results in most of the trending SEO criteria (performance, accessibility, best practices, *etc.*).

At present, there are four incorporations of the platform (*incl.* the current one) – the *Bulgarian Iconographical Digital Library* (n.d.), the *Digital Library of the Public Library "Ivan Vazov" – Plovdiv* (n.d.), the *Digital Library "Peyo Yavorov" - Burgas* (n.d.) and the Digital Library of BAS (n.d.), with their efficiency and usability being constantly monitored (Luchev et al., 2021; Paneva-Marinova, et al., 2022a; Paneva-Marinova, et al., 2022b).

2.2 Dataset

The Central Library of the Bulgarian Academy of Sciences is the first scientific library in Bulgaria, the most significant national center for literary and documentary heritage, and information database for fundamental and applied research. The library holds over two million library documents and has a significant contribution for the development of the national library information resources. The process of digitization of significant collections of the repository has started more than fifteen years ago. During this period, a number of digital datasets with successful application in the digital informational and educational environment have been created.

One of the noteworthy collections of CL-BAS is associated with the international congresses of Bulgarian studies from the 1980s (the Bulgarica Collection). At the First international congress of Bulgarian studies, held in Sofia between 23rd and 31^{s}t of May, 1981, over nine hundred scientific papers and reports were presented, at several panels, seminars, round tables and symposiums. They cover an exceptionally rich set of problems from a wide range of scientific fields, *e.g.*, history, archaeology, history of culture, art studies, linguistics, theory of literature, folklore studies, *etc.* Around five hundred foreign scientists from more than forty countries had participated in the congress. The scientific works are published in more than twenty separate volumes in a variety of languages, including Bulgarian, Russian, English, German and French. Even wider linguistic variety is present in the papers from the twenty-three volumes of the Second international congress of Bulgarian studies, Sofia, 23rd May – 3rd June 1986, including Polish and Spanish, besides the aforementioned languages. The papers of the two congresses are exceptionally valuable, since they represent the achievements and progress in the field of Bulgarian studies – an interdisciplinary science, reaching its zenith during the 1980s, both in and outside Bulgaria. The efforts for digitization, description and recording of the objects is further complicated by the large number of the publications and the variety of languages and thematic fields.

The description of the contents of cultural heritage objects, and in particular the literary scientific heritage, is usually accomplished by employing established standards, such as DublinCore, CIDOC CRM, *etc.* For the Bulgarica Collection, an extension of DublinCore is used, a description scheme reflecting the specifics of the target objects. The platform, which is the basis of the library, provides a module for editing and expansion of the description scheme, and allows for future adjustments, modifications, revisions, and improvements.

3 Results and Discussion

3.1 Used Descriptive Model

The descriptive model aims to provide opportunities for integral presentation of the objects, with regard to the efficient access to the presented knowledge, and is based on a preliminary scientific and bibliographical analysis of the characteristics of the described object. Figure 2 illustrates schematically the descriptive model for an object "Paper" and its primary descriptors. The knowledge level "Paper" includes identification regarding Author, Title, Country, Pages, Language and Issue. The knowledge level "Issue" includes descriptors for identification and description of the respective volume, in which the object is found, such as Type, Author, Title, Subtitle, Extend, Inventory No., *etc.*, and has a special sub-level "Publisher" with descriptors "Publisher", "Publishing Year" and "Publishing Place". The scheme and the metadata of the knowledge level "Issue" are consistent with the library descriptions of the individual volumes of the Bulgarica Collection in the catalogue of CL-BAS.

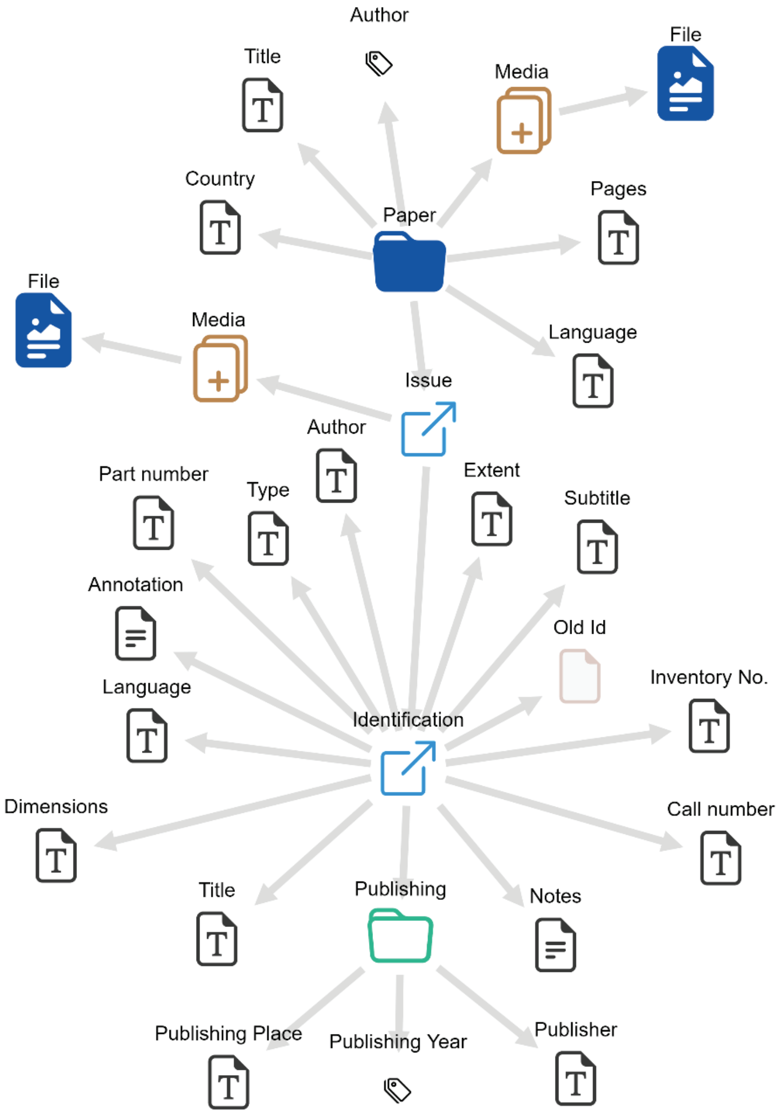


Fig. 2. Descriptive model

3.2 The Resulted Digital Library of CL-BAS

The digital library for noteworthy scientific literary heritage, created for the needs of the Central Library of the Bulgarian Academy of Sciences, uses as a core the already presented environment for storage, extraction, and curation of data from different fields of the Bulgarian cultural and historical heritage. Two screenshots are depicted on figure 3 and 4.

The platform was implemented as a single page application (SPA) using Vue JS and Bootstrap as main front-end technologies. The back-end is powered by a Node JS application behind an Apache web server (a NGINX configuration is also available). MongoDB is used for data storing and querying. The lightweight API based communication between back-end and front-end is one of the main reasons for the notably good performance and fast interaction of the application.

In the standard version of the platform, a MongoDB search engine is used for full text searches (FTS). It is suitable for cases where text indexes are relatively small and full text search is not the main purpose of the application. The version implemented for the Bulgarica Collection of the Central Library of BAS contains huge amounts of text data, and the full text search feature is highly demanded, so the standard FTS was replaced by the SPHINX search FTS engine. It is a high performance, low resource consuming engine and the tests we have performed proved its qualities.

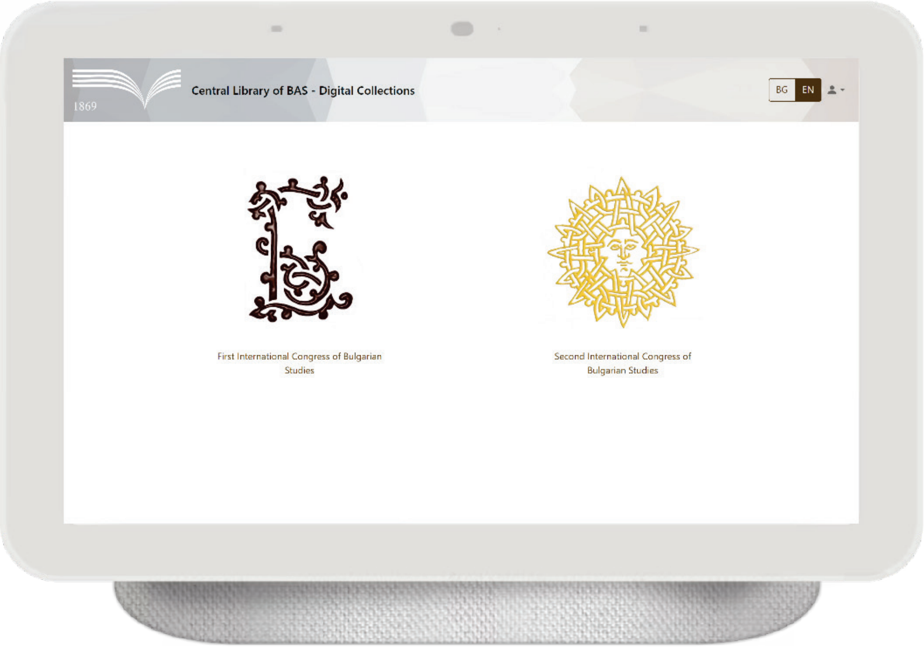
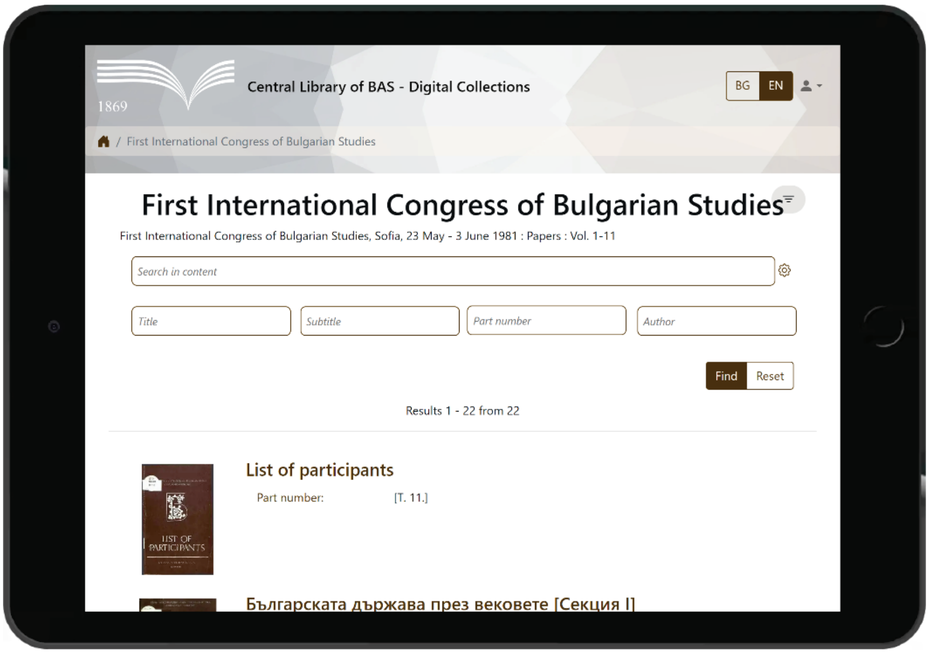

1869
Central Library of BAS - Digital Collections
BG
EN
First International Congress of Bulgarian Studies
Second International Congress of Bulgarian Studies

Fig. 3. Home page of the DL


1869
Central Library of BAS - Digital Collections
BG
EN
First International Congress of Bulgarian Studies
Search in content
Title
Subtitle
Part number
Author
Find
Reset
LIST OF PARTICIPANTS

First International Congress of Bulgarian Studies

First International Congress of Bulgarian Studies, Sofia, 23 May - 3 June 1981 : Papers : Vol. 1-11

Results 1 - 22 from 22

List of participants

Part number: [T. 11.]

Българската държава през вековете [Секция I]

Fig. 4. Collection review

3.3 Some Issues on the Development of the Digital Library of CL-BAS

The biggest challenges our team faced during the implementation were related to the migration of data from an older platform used as a digital storage. Data was kept in DSpace software, (a version from 2016). Furthermore, the digitization of the Bulgarica Collection (*incl.* International Congresses of Bulgarian Studies) was implemented more than ten years ago by photographing and saving the individual pages from the collections of the congresses and their differentiation into separate archive units (even though in groups, they constitute the individual articles). As a result, a preliminary processing and curation of the objects were required. Moreover, in order to perform the migration, the following modules were designed and implemented:

- **OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) metadata extractor**. This module allows extracting information about objects available in DSpace and their categorization and content location.
- **METS (Metadata Encoding and Transmission Standard) parser**. DSpace is using METS for the description of individual files as a part of an object (in our case, every page as a part of a Bulgarian study issue was described using METS.)
- **MARC 21 metadata processor (machine-readable cataloging, a popular standard for library catalogs and library management)**. This standard was used for management of issues metadata (*e.g.,* title, subtitle, author, language, *etc.*)

- **TOC (table of contents) processor**. After content extraction was executed (all Bulgarian studies issues were extracted as sets of JPEG images), OCR (optical character recognition) was performed. The task of the TOC processor was to identify all issues' tables of contents, parse them and find, locate and describe all individual papers inside every issue.
- **PDF paper builder**. A tool which uses the data already collected from the previous steps and creates a separate PDF document for every paper identified.
- **Metadata summarizer and importer**. The final step of metadata preparation for importing in the new platform.

There are some inconsistencies in the indexing of the content of the collections, which requires collaborative, intensive work and synchronization between the curators of the content and the team, developing the platform.

4 Conclusions

In this paper, we have presented the latest application of our team's digital library environment. Albeit covering a relatively small portion of the huge archives of the Bulgarian Academy of Sciences, we hope it will (re)introduce these works to both the scientific community and the general public and inspire them in their studies in the field. Furthermore, as the importance of conservation, dissemination and proliferation cannot be overstated, we aim to continue reaching out to scientific and cultural institutions and assist them in providing smooth access to their valuable knowledge repositories.

Acknowledgements.

This research work was carried out and is funded partly by CLaDA-BG, the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favour of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH, Grant number DO01-167/28.07.2022, https://clada-bg.eu/bg/.

This research work was carried out and is supported partly by the joint research project "Digital Art and Culture for Educational Applications" (Grant number IC-LV/01/2022-2023) between Latvia Culture College at Latvian Academy of Culture and Institute of Mathematics and Informatics, Bulgarian Academy of Sciences (2022-2023).

References

Bulgarian Iconographical Digital Library. (n.d.). Retrieved March 15, 2023, from https://bidl.math.bas.bg/en

Digital Library "Peyo Yavorov" - Burgas. (n.d.). Retrieved March 18, 2023, from https://plakati.bg73.net

Digital Library of BAS. (n.d.). Retrieved March 19, 2023, from https://clbas.bg73.net/

Digital Library of the Public Library “Ivan Vazov” - Plovdiv. (n.d.). Retrieved March 17, 2023, from https://digital.libplovdiv.com/

Luchev, D., Goynov, M., Paneva-Marinova, D., Stoykov, J., & Pavlova, L. (2021). Synergy of National Cultural Heritage and Technology. *Digital Presentation and Preservation of Cultural and Scientific Heritage*, *11*, 281–286. https://doi.org/10.55630/dipp.2021.11.26

Bulgarian Ministry of Education and Science. (2021). *Natsionalna patna karta za nauchna infrastruktura 2020-2027 g.* [National Science Infrastructure Roadmap 2020-2027]. Retrieved March 16, 2023, from https://web.mon.bg/upload/26649/RoadMapBulgaria_2020-2027_BG_sm_11062021.pdf

Paneva-Marinova, D., Goynov, M., Luchev, D., Zhelev, Y., Monova-Zheleva, M., Pavlov, R., Zlatkov, L., Noev, N., & Pavlova, L. (2022a). Information Day: Research Infrastructure Services in the Humanities and Social Sciences. *Digital Presentation and Preservation of Cultural and Scientific Heritage*, *12*, 319–324. https://doi.org/10.55630/dipp.2022.12.32

Paneva-Marinova, D., Minev, D., Krachanov, I., Luchev, D., Goynov, M., Zlatkov, L., Pavlov, R., & Pavlova, L. (2022b), The Power of Intelligent Content Curation and Context-Based Digital Library Content Usage for Research and E-Learning (The Case of National Library “Ivan Vazov” – Plovdiv, Bulgaria). In *INTED2022 Proceedings. 16th International Technology, Education and Development Conference* (pp. 4013-4021). IATED. https://doi.org/10.21125/inted.2022.1097

Received: March 25, 2023
Reviewed: May 20, 2023
Finally Accepted: June 01, 2023