Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 204 (2022) 744–752

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Industry Sciences and Computer Science Innovation

Analyzing the benefits of using a document repository to aid decision-making in the field of culture

Bozidar Kovacic[a,*], Vanja Slavuj[a], Martina Asenbrener Katic[a]

[a]*University of Rijeka, Faculty of Informatics and Digital Technologies, Radmile Matejcic 2, 51000 Rijeka, Croatia*

Abstract

The MESOC Toolkit is a free and open-access online service used to aid the measurement of the societal value and impact of culture, currently under development as part of the MESOC (Measuring the Social Dimension of Culture) project. The goal of establishing transition variables that serve as impact measurement indicators is aided by an online document repository system of thematic publications (i.e., a collection of documents on the societal value and impact of cultural policies). It allows the users to input, manage, and search extensive document data based on the relevant criteria.

The paper at hand paper describes the rationale behind the online document repository of the MESOC Toolkit, its implementation, and main functionalities. Furthermore, it offers an empirical evaluation of the benefits that such a system offers over non-automated procedures of document search and analysis. The results of the evaluation give way to the conclusion that there are time-saving and quality benefits for the users of the repository: efficiency in dealing with a large number of documents and quality in searching the documents using multiple criteria.

Keywords: document repository; online toolkit; document search; software evaluation; MESOC

1. Introduction

MESOC, or Measuring the Social Dimension of Culture, is a multinational research and innovation project financed under the Horizon 2020 EU program. The project gathers 10 partner institutions across 7 European countries, namely

* Corresponding author. Tel.: +385-51-584-712; fax: +385-51-584-749.
E-mail address: bkovacic@inf.uniri.hr

10.1016/j.procs.2022.08.090

Belgium, Croatia, France, Greece, Italy, Romania, and Spain, with some crucial activities of the project being carried out in two additional European countries (Finland and Poland). The partner institutions make up a well-balanced, yet diverse and multidisciplinary consortium that includes higher education establishments, non-profit organizations operating within the field of culture, associations for promotional activities in culture, and cities. The consortium is led and coordinated by the University of Valencia's unit Econcult.

The main goal of the MESOC project is to propose, test and validate a novel approach to measuring impact and the so-called societal value of culture and its policies and practices, as suggested within the New European Agenda for Culture [1]. Such an approach adapts and extends the previously developed method for impact assessment based on transition variables [2], taking into account the three crossover themes described in the New European Agenda for Culture (namely Health and Wellbeing, Urban and Territorial Renovation, and People's Engagement and Participation) and ten cultural domains taken from the ESSnet-Culture report (namely Heritage, Archives, Libraries, Book and Press, Visual Arts, Performing Arts, Audiovisual and Multimedia, Architecture, Advertising, and Art Crafts) [3]. Determining these transition variables, for each intersection between a particular theme and domain, offers ample opportunities to carry out the societal impact measurement procedure in a given context (e.g., a city, a region, or a country).

The final output of the MESOC project is an open access online service, tentatively named the MESOC Toolkit, that will allow researchers and practitioners, from the field of culture and beyond, to perform measurements of societal value and impact that various cultural policies and associated practices have in a given social context. One of the principal components of the Toolkit is the online document repository, developed specifically to store and manage large collections of documents – cultural policies, research papers, reports, and other knowledge sources relevant for extracting the transition variables described above. However, before variable extraction is performed, aided by methods such as keyword extraction, each document in the repository needs to be prepared for text analysis. This includes inputting related descriptive data (or metadata) for each document by hand into the repository and uploading the corresponding digital documents in the required file format. Furthermore, repository users need to be able to subsequently access the documents containing full texts already in the repository and view the accompanying document descriptions. If necessary, users should also be able to perform searches based on selected keywords, make further changes to the descriptive data, or receive statistical data regarding the contents of the repository. All of these activities may be used in the process of determining the final subset of transition variables used for measurement.

The paper at hand aims to describe the developed online repository of documents, as part of the MESOC Toolkit, with the following two research questions in focus:

- What are the services offered by the online document repository of the MESOC Toolkit and how are they implemented?
- What is the measurable benefit of using a document repository and its automated services, as opposed to non-automated procedures or less-structured procedures?

The rest of this paper is organized as follows. Section 2 gives a brief overview of related work. Section 3 describes the context in which the repository is developed and offers details of the main functionalities found in the Toolkit's document repository. Section 4 presents the results of the evaluation of the main benefits regarding the use of the document repository and its services. Section 5 concludes the paper and suggests future work.

2. Basic concepts and related work

An online document repository represents a space for storing, organizing, and sharing a vast number of digital documents and their metadata [4] [5], thus eliminating the need to keep the physical (or analog) versions of the same documents [6]. Being further characterized as an online system, such a repository is accessed via the World Wide Web using a web browser, as the document files are not stored on the repository user's device, but on a remote server. In order to ensure that the shared documents are secure, online repositories necessarily employ some type of user-management service [7], which enables its administrators to grant individual users certain types of access rights (e.g., for reading, uploading, deleting, or modifying the documents or their metadata).

The overview of relevant sources reveals that document repositories described in scientific and expert journals vary significantly in their primary aim, domain and field of use, as well as degree of complexity and sophistication (i.e., number and diversity of functions users have at their disposal).

The authors in [8] describe the implementation of a small-scale repository system for the purpose of managing digital resources at higher education institutions. The repository, which acts as an online library of published documents, is implemented as a web application and is based on the MVC (Model-View-Controller) pattern of software design and development. It allows its users to perform the basic CRUD functions (namely create, read, update, and delete) and to search the texts contained within the documents. The system makes a distinction between three groups of users: general users who view only public documents; users that have access to a certain subset of non-public documents; and system administrators who manage the whole system and user accounts.

In [9] the authors describe PRESS – a publication repository semantic system developed for the purpose of disseminating scientific research results in academia and industry. The repository offers its users possibilities to enter (meta)data, perform simple and advanced queries on the data, and integrate it with other similar systems. It uses semantic web technology in order to develop the basic ontology – the basis for efficient knowledge management (e.g., classification of publications) and high-performance query capabilities regarding the metadata (e.g., retrieving information from a full text document). Similar approaches using semantic techniques for searching and synthesizing information from document repositories are found in the SONCA system [10] and Dublin Core system [11].

The approach reported in [12] considers the case of developing the Illinois Data Bank – a web-based publishing platform used to gather, store, and disseminate research data generated in the state of Illinois, USA. Its primary purpose is to manage and share metadata concerning research data as part of the public library system.

Document repository use is reported in the education sector as well. In [13], the authors give details about a document repository used to build a university-level learning materials system for students and academics in the field of IT. The approach organizes the contents based on an ontology of domain knowledge, uses a database for managing documents, and offers semantic representation of the documents stored in the system in order to allow for semantic processing and advanced search techniques to be applied.

3. MESOC document repository

3.1. Context of development and motivation

In this subsection we briefly describe the aspects of the original MESOC approach relevant for the development of the document repository, part of the MECOS Toolkit, and the motivation for its development.

As stated before, the MESOC approach focuses on developing a method for transition-based impact measurement in the field of culture. In order to be able to suggest and validate such cultural indicators, crucial for the decision-making process and the design of new public policies, a complete and comprehensive database or relevant scientific and expert sources is indispensable [14]. Furthermore, the data contained therein should be reliable and relevant if quality decisions are to be made. In that case, the role of the domain expert is crucial and brings in added value to searching and evaluating repository documents.

Keeping a physical repository of relevant documents and searching or analyzing it by hand may prove to be a laborious and highly unreliable task, especially if the previously mentioned completeness criterion is to be considered. Entrusting a piece of software, such as a document repository, to take on the task of managing, searching and analyzing document data should be more efficient, both regarding time and quality. However, the real impact of these benefits should be analyzed and confirmed empirically, which is presented in the later part of this paper.

MESOC's document repository and its functionalities closely follow the MESOC 10x3 Matrix (see Fig. 1). The MESOC Matrix displays the relationships (intersections) between three crossover themes and ten cultural domains (already discussed above) in order to further detail the analysis of documents and help with the process of deciding which transition variables to use in policy-making.

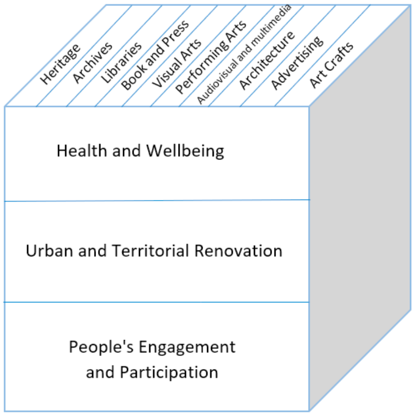


Fig. 1. The MESOC methodology (matrix) as applied to the document repository.

3.2. Design of the online document repository

After collecting and analyzing relevant documents – regarding social impacts covering the three crossover themes – they are stored in the document repository. By searching through scientific databases using keywords, domain experts are able to select the relevant documents and transfer them into the repository. Every document part of the repository is tagged using the MESOC Matrix. Furthermore, every document is defined by 24 characteristics (i.e., pieces of metadata). There are two broad categories of these characteristics: (1) data collected from scientific publication databases (e.g., Scopus) and (2) data defined by the domain expert. The former includes year of publication, document title, abstract, keywords, author, institution, document type, language, number of pages, bibliographic reference, search database, open access, DOI, and link to document, while the latter includes research technique, data provider, time period of data, methodology, findings and outcomes, relevance, document category, city, transition variable, and keyword transition variable.

The repository allows the users to input, maintain and search documents using one or more of these characteristics as search criteria. The data that makes up the characteristics may be:

- identical to the data in the scientific database (e.g., title or abstract),
- part of a pre-defined list of values (e.g., document type, language, or city), or
- defined by the domain expert (e.g., methodology, findings and outcomes, or relevance).

3.3. Technical implementation of the online document repository

The following were used in the development of the MESOC document repository: the MySQL relational database, the PHP programming language, the Apache server, and the Phalcon framework for rapid web application development. Since the MySQL relational database, the PHP programming language and the Apache server are widely known and standard in web development, we will briefly focus only on the Phalcon framework.

Phalcon is an open source full-stack framework for PHP, optimized for high performance application development. It is written as a C-extension, so developers do not need to know the C language if they want to use it. The Phalcon interface is provided as a collection of PHP classes under the Phalcon namespace and can be used immediately. It has a unique architecture that allows the framework to always be memory-resident and provide its functionality whenever it is needed, without expensive file statistics and file reading operations common to traditional PHP frameworks [15].

The MVC architecture was used for the development of the MESOC document repository. It is an architectural pattern that divides an application into three main logical components: Model (stores data and associated logic), View (displays information from the Model to the user), and Controller (controls the flow of data into a Model object and updates the View when data changes) [16].

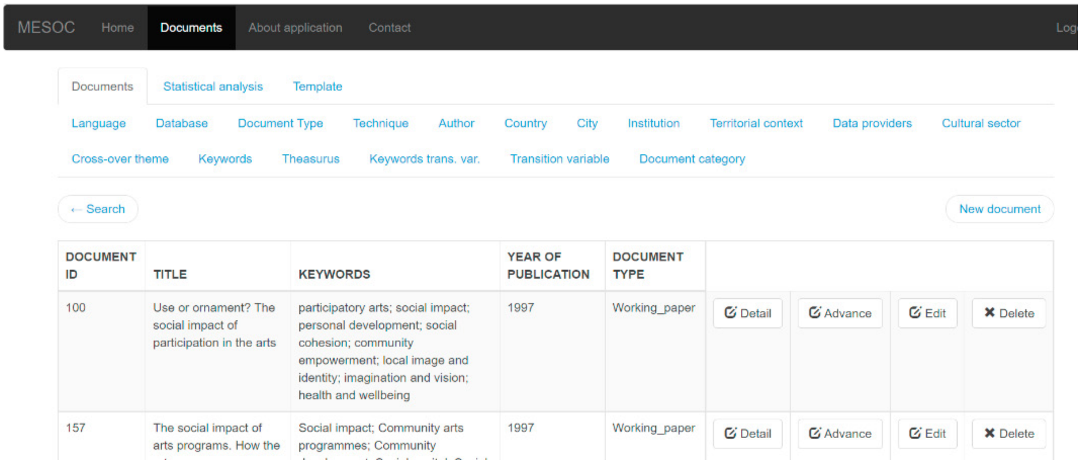
The developed online document repository is based on the MySQL database model described in [17].

3.4. Main functionalities of the online document repository

After logging in to work with the repository, the repository's home page opens to the "Home" item of the menu. It offers the user a welcome message and describes the main characteristics and functionalities of the repository. From there, the user has several options from which to choose vain (navigation menu near the top). For example, the "About application" item provides further information about the repository itself, while the "Contact" item allows users to submit comments about the repository and contact the system's administrators.

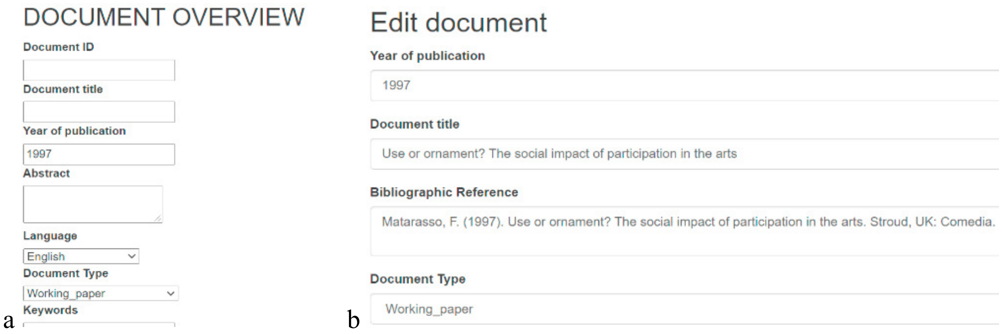
The "Documents" item is where all the main functionalities of the repository reside. These include tools for searching, viewing, modifying, and deleting the existing documents ("Documents" tab), entering (data about) new documents ("New document" button), obtaining various statistics concerning the analysis of documents by various search categories ("Statistical analysis" tab), and editing the search template to change the offered categories as search parameters ("Template" tab). Part of the list of documents currently stored in the document repository is shown in Fig. 2 (obtained by using the *Year of publication = 1997* criterion on the "Documents" tab).

As can be seen from Fig. 2, the list of documents matching the set search criterion offers the most basic data about each document (document ID, title, keywords, year of publication, and document type). Additionally, for each document there are four choices that correspond to the previously mentioned functionalities (represented by buttons). Each search could be further refined by adding more search parameters (keywords) to the search form. Fig. 3 shows how search is performed using three criteria (*Year of publication*, *Language* and *Document type*) and how one document, yielded by the search, is being edited using the editing functionality.



DOCUMENT ID	TITLE	KEYWORDS	YEAR OF PUBLICATION	DOCUMENT TYPE				
100	Use or ornament? The social impact of participation in the arts	participatory arts; social impact; personal development; social cohesion; community empowerment; local image and identity; imagination and vision; health and wellbeing	1997	Working_paper	Detail	Advance	Edit	✖ Delete
157	The social impact of arts programs. How the	Social impact; Community arts programmes; Community	1997	Working_paper	Detail	Advance	Edit	✖ Delete

Fig. 2. List of documents and their associated characteristics stored in the document repository.



DOCUMENT OVERVIEW

Document ID

Document title

Year of publication: 1997

Abstract

Language: English

Document Type: Working_paper

Keywords

Edit document

Year of publication: 1997

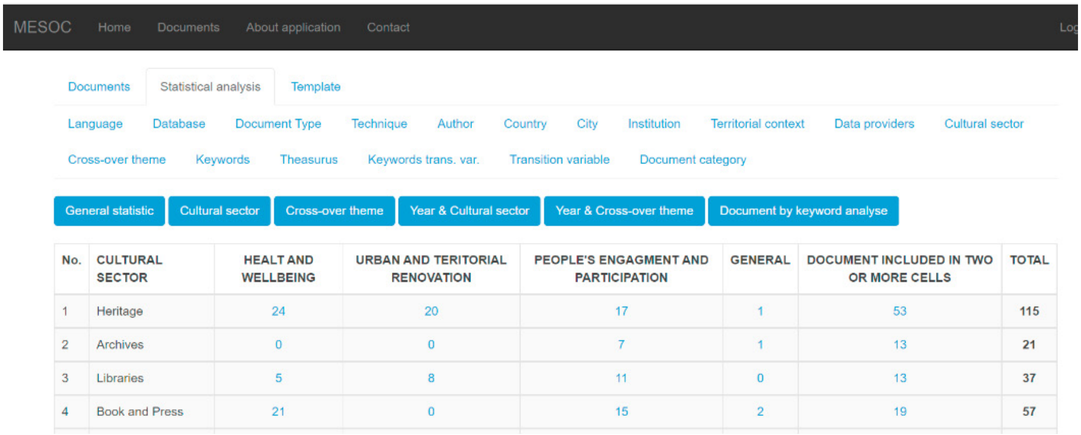
Document title: Use or ornament? The social impact of participation in the arts

Bibliographic Reference: Matarasso, F. (1997). Use or ornament? The social impact of participation in the arts. Stroud, UK: Comedia.

Document Type: Working_paper

Fig. 3. (a) repository search form using several search criteria; (b) editing repository entry.

Fig. 4 gives an overview of general statistics about the documents in the repository, given the three crossover themes and ten cultural domains. Additionally, other statistical data is available to users (e.g., by cultural sector or year and crossover theme) that may help in establishing the transition variables necessary for future policy-making.



No.	CULTURAL SECTOR	HEALT AND WELLBEING	URBAN AND TERITORIAL RENOVATION	PEOPLE'S ENGAGMENT AND PARTICIPATION	GENERAL	DOCUMENT INCLUDED IN TWO OR MORE CELLS	TOTAL
1	Heritage	24	20	17	1	53	115
2	Archives	0	0	7	1	13	21
3	Libraries	5	8	11	0	13	37
4	Book and Press	21	0	15	2	19	57

Fig. 4. Statistical data on culture-related documents in the repository.

4. Evaluation of the benefits of using the document repository

The following analysis of repository use aims to evaluate the benefits of its use in the context described above. The analysis relies on the time necessary to search the documents in the repository using a set of keywords, and differentiates between three approaches:

- manual search,
- search using an automated search script (a computer program replacing manual search), and
- repository search.

4.1. Manual search

In manually searching through each document, the user must open each file using the appropriate software (e.g., a PDF reader or a text processor), select the search functionality, enter a keyword, perform the search, and close the document. Documents that are being searched may be original scientific papers (full texts) or just data exported from a scientific database. The time necessary to perform these activities was estimated by observing 3 domain experts perform the described tasks on 6 different documents (2 PDF files, 2 MS Word files, and 2 exported data file). The estimated times (in seconds, minutes and hours) for each activity, performed on a single document, are given in Tab. 1, and are extrapolated to 680 documents (number of documents currently in the MESOC's document repository). As can be observed from the results, a manual search of a single document, using a single keyword, takes approximately 15 seconds, while the search of 680 documents takes approximately 2.83 hours.

Table 1. Estimated time necessary to perform a manual search of documents.

No. of documents	Opening and closing a document			Select the search functionality, enter a keyword, perform the search			Total		
	s	min	h	s	min	h	s	min	h
1	5	0.083	0.00138	10	0.166	0.00277	15	0.249	0.00415
680	3400	56.66	0.944	6800	113.33	1.88	10200	170	2.83

4.2. Search using an automated search script

In order to estimate the time necessary to search a large number of documents using a pre-programmed script, all the document data had to be prepared for export first and then saved as plain-text files. Using the script, the domain expert enters a keyword that will be used for searching the documents. The script then automatically opens each prepared document, searches it for the defined keyword, saves the search results (was the keyword found in the document?), and finally closes the document. This procedure is repeated on every document that needs to be searched (all 680 of them). Times obtained by the experiment are shown in Tab. 2. As can be observed from the table, the time necessary to search all 680 documents is 8.0478 seconds (or 0.0022133 hours).

Table 2. Estimated time necessary to perform a search using a pre-programmed script.

No. of documents	Keyword entry (average time 8 s)			Open a document, search a document, record search results, close a document			Total		
	s	min	h	s	min	h	s	min	h
680	8	0.13	0.0022	0.047893	0.00798	0.0000133	8.0478	0.13798	0.0022133

4.3. Repository search

Performing document search in the document repository is possible only after all the data describing the documents has been entered into the online system. Thus, the following analysis must take into consideration the time necessary for data entry, which may differ depending on the structure of the document (e.g., writing style, number of pages, etc.) and the availability of data in the document. The time necessary for data entry was calculated by observing the amount of time 3 domain experts took to enter data on each of the 6 selected documents (the same documents across all three users) and averaging the obtained times. The result was then rounded to the nearest whole number and equals 90 minutes. If we apply this to 680 documents currently in the repository, the time necessary to enter all that data into the online system amounts to 1080 hours.

After data entry, the domain expert is able to enter keywords into the search form and obtain search results on the computer screen. Times obtained by searching the documents using the repository (entering the keyword, starting the search, and getting results) are given in Tab. 3.

Table 3. Estimated time necessary to perform a search using the document repository.

No. of documents	Keyword entry (average time 8 s)			Search a document, display search results			Total		
	s	min	h	s	min	h	s	min	h
680	8	0.13	0.0022	0.025822	0.000043	0.000000712	8.02582	0.130043	0.00220071

4.4. Comparison and interpretation of results

The results presented above reveal that the average time of searching 680 documents using the document repository amounts to 8.02585 seconds (or 0.00220071 hours). If we take into consideration that the estimated time of entering data for 680 documents into the repository (by hand) amounts to 1080 hours (and the fact that cannot be avoided), the overall time for searching all the documents amounts to 1080.00220071 hours. For each subsequent search, this overall time is increased by only 0.00220071 hours.

Given the amount of time necessary to perform manual document search (2.83 hpurs), and the time necessary to input document data into the document repository and perform subsequent searches (1080.00220071 hours), it is after 383 keyword searches using the manual approach that the search time exceeds that of the document repository approach, and the latter approach becomes beneficial.

The analysis conducted above refers to searches that use a single keyword, i.e., a single document characteristic. In order to perform document searches using multiple keywords in the manual approach, it is necessary to repeat the search procedure for each keyword separately. This means that the time necessary for completing a search of a single document also increases significantly. In the same vein, using the automated search script approach, it is necessary to enter multiple keywords into the script, thus starting a search for each of them separately. As the maximum number of keywords is 24, the amount of time necessary to perform a search using all the possible criteria potentially amounts to 192 additional seconds for each document. Using the repository to perform a search means that the user has to enter all the keywords at the same time. Hence, the overall search time is then calculated by adding up the amount of time necessary for data entry (1080 hours), the amount of time necessary to enter keywords (24 x 8 = 192 seconds), and the average time of search (0.047893 seconds), which comes up to a total of 1080.0533 hours.

The added value of performing a search using the document repository is especially noticeable when dealing with the characteristics defined by the domain expert. These are not available for use in the manual search approach as they are not overtly included in the document being searched (i.e., they appear only in the repository, once the domain expert enters them into the system).

The use of a search script offers better results than the manual search approach. However, one needs to take into consideration the amount of time necessary to develop such a script, as well as the time necessary to prepare the data being searched in a suitable format (as it will "be read" by the script). If such a search is performed using the original files containing scientific papers, the data preparation step may be skipped. However, in that case one does not have access to the characteristics defined by the domain expert.

5. Conclusion

The paper at hand gave an overview of the MESOC project and its expected outcomes. One of the deliverables, part of the MESOC Toolkit, is a repository of domain-related documents that is used to manage and search the most important knowledge sources, indispensable for measuring the impact (transitional) variables in the field of culture.

The paper gave answers to two research questions. Firstly, it described the rationale behind the development of the online document repository and presented its main functionalities. Secondly, it presented and discussed the results of an evaluation of the benefits of using a document repository and its automated services, as opposed to non-automated procedures or less-structured procedures. The evaluation was based on the time necessary to perform document data search using the repository, compared to the manual approach and the automatic script approach, and has shown that the use of a document repository gives the most beneficial results. The time required to enter the document data into the repository becomes beneficial after 383 keyword searches using the manual approach.

Future work on the presented document repository system includes a redesign of the user interface in order to increase the overall user experience with the system, and its integration with a system for automatically analyzing text-based documents in order to extract relevant keywords from it.

Acknowledgements

This work was supported by the European Union and funded under H2020-EU.3.6. and H2020-EU.3.6.3.2. (grant agreement ID 870935), and by the University of Rijeka projects uniri-drustv-18-73 and uniri-drustv-18-224.

References

[1] European Commission. (2018) "A New European Agenda for Culture - Background Information." Available at https://ec.europa.eu/culture/sites/default/files/2020-08/swd-2018-167-new-european-agenda-for-culture_en.pdf.

[2] Barata, Felipe Themudo, Francesco Molinari, Jesse Marsh, and Sónia Moreira Cabeça. (2017). "Creative Innovation and Related Living Lab Experiences: A Mediterranean Model." Évora: Universidade de Évora.

[3] Bína, Vladimír, Philippe Chantepie, Valérie Deroin, Guy Frank, Kutt Kommel, Josef Kotýnek, and Philippe Robin. (2012) "ESSnet-Culture Final Report." Available at https://ec.europa.eu/assets/eac/culture/library/reports/ess-net-report_en.pdf.

[4] Takayshvili, Liudmila (2010) "Concept Document Repository to Support Research of the Coal Industry Development Forecasting" in J.-S. Pan, S.-M. Chen, and N.T. Nguyen (eds) *Computational Collective Intelligence. Technologies and Applications. ICCCI 2010. Lecture Notes in Computer Science*, Berlin, Springer.

[5] Hadzhikolev, Emil, Stanka Hadzhikoleva, and Daniela Orozova. (2018) "Digital Model of a Document in a University Document Repository." *Proceedings of the 20th International Symposium on Electrical Apparatus and Technologies (SIELA),* 2018: 1–4.

[6] Manoj, G., Deep Ishan, V. Kalyani, K. C. Sahana, and R. P. Madhavi. (2015) “Online Document Repository System.” *International Journal of Advance Research in Computer Science and Management Studies* **3** (**3**): 74–80.

[7] Stern, Warren, Yonggang Cui, Jose Gomera, Maia Gemmill, and Katherine Bachner. (2018) “Science and Engineering Team Repository Selection and Validation.”, Upton, Brookhaven National Laboratory.

[8] Kajitori, Kazuaki, and Kunimasa Aoki. (2016) “Implementation of a simple document repository system.” *International Journal of Modern Education and Computer Science* **9**: 12–19.

[9] Chrysakis, Ioannis, Emmanouil Dermitzakis, Giorgos Flouris, Theodore Patkos, and Dimitris Plexousakis. (2017) “PRESS: A publication repository semantic system.” *Proceedings of the 13th International Conference on Semantic Systems - SEMANTiCS 2017*. Available at https://publications.ics.forth.gr/_publications/PRESS_-_A_Publication_REpository_Semantic_System.pdf.

[10] Nguyen, Hung Son, Dominik Ślęzak, Andrzej Skowron, and Jan G. Bazan (2011) “Semantic search and analytics over large repository of scientific articles” in Robert Bembenik et al. (eds) *Intelligent tools for building a scientific information platform: Advanced architectures and solutions*, Heidelberg, Springer.

[11] Rodríguez, Alejandro, Ricardo Colomo, Juan Miguel Gómez, Giner Alor-Hernandez, Ruben Posada-Gomez, Ulises Juarez-Martinez, Jose Emilio Labra Gayo, and Krishnamurthy Vidyasankar. (2009) “A proposal for a semantic intelligent document repository architecture” *Proceedings of the 2009 Electronics, Robotics and Automotive Mechanics Conference*, 69–75.

[12] Fallaw, Colleen, Elise Dunham, Elizabeth Wickes, Dena Strong, Ayla Stein, Qian Zhang, Kyle Rimkus, Bill Ingram, and Heidi J. Imker. (2016) “Overly Honest Data Repository Development.” *Code 4 Lib Journal* **34** [online]. Available at https://journal.code4lib.org/articles/11980.

[13] Do, Nhon, Thuong Huynh, and An Pham. (2011) “Organization model of semantic document repository and search techniques for studying information technology” *World Academy of Science, Engineering and Technology* **59**: 549–554.

[14] Lladó, Anna Planas, and Pere Soler Masó. (2011) “Design and application of a system of evaluation indicators for municipal cultural policies.” *Evaluation Journal* **17** (**3**): 277–292.

[15] Phalcon.io, “About Phalcon Framework”, 2022. [Online]. Available: https://phalcon.io/en-us/about. [Accessed 15-Jan-2022].

[16] Bucanek, James. (2009) “*Learn Objective-C for Java Developers*” New York, NY: Apress.

[17] Jaksic, Danijela, Sanja Candrlic, and Patrizia Poscic. (2022) “From User Requirements to Document Repository Enriched with Metadata – a Case Study”. *iSCSi – International Conference on Industry Science and Computer Sciences Innovation*, Porto, Portugal