Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 204 (2022) 760–767

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Industry Sciences and Computer Science Innovation

From User Requirements to Document Repository Enriched with Metadata – a Case Study

Danijela Jaksic*, Sanja Candrlic, Patrizia Poscic

University of Rijeka, Faculty of Informatics and Digital Technologies, Radmile Matejcic 2, Rijeka 51000, Croatia

Abstract

The paper describes a research on collecting user requirements from international users in order to develop a document repository based on metadata, for collecting, storing, searching and analysing relevant cultural and socioeconomic documents. This research is a part of wider research within the project called MESOC that aims to propose, test and validate an innovative and original approach to measure the societal value and impact of culture, cultural policies and cultural practices. The main contributions of this paper are: a) a set of user requirements in the cultural domain collected from international users, b) a data model and database for the document repository, and c) a metadata model used to satisfy the user requirements.

Keywords: user requirements; data model; document repository; metadata

1. Introduction

Measuring the Social Dimension of Culture (MESOC) is a Horizon 2020 research and innovation project that aims to propose, test, and validate an innovative and original approach to measuring the societal value and impact of culture, cultural policies, and cultural practices that addresses three crossover themes of the new European Agenda for Culture: 1) health and well-being, 2) urban and territorial renovation and 3) people's engagement and participation [1].

In order to analyze cultural and socio-economic data, gain knowledge from it, and find a way to measure the societal value and impact of culture and cultural policies and practices, a document repository must first be developed. This

* Corresponding author. Tel.: +385 51 584 725.
E-mail address: danijela.jaksic@inf.uniri.hr

1877-0509
10.1016/j.procs.2022.08.092

document repository would allow for the collection, preservation, and retrieval of documents in the cultural and socio-economic fields. Examples of such documents (all of which are from the public domain) include: research papers, case studies, project reports, policy briefs, initiatives, etc. Documents are stored in their original language but described using repository attributes in English to facilitate retrieval and interpretation.

The first step in the development is to collect user requirements. Then, user requirements must be translated into a database model, based on which a database is created to store documents and all their metadata. After that, a web-based document repository system must be developed. This system will collect, store, search and retrieve documents and all relevant information. In addition to general database searching, the system must have the functionality of text and keyword-based searching (queries). The purpose of this system is to analyze, extract knowledge, and measure the societal value and impact of culture and cultural policies and practices.

This paper presents the first phase of this project: a) user requirements gathering, b) database modeling based on user requirements, and c) designing a database for the document repository. Since the paper only covers the deliverables planned for the early phase of this project, the paper does not cover the development of the web application.

2. Related work

Research on digital transformation has signaled a shift from technology to organization and people [2]. This means that digitization, digitalization and datafication must be considered in the development of digital document repositories. Based on [3] digitization refers to the encoding of actions or representations of actions in a digital format (zeros and ones) that can be read, processed, transmitted, and stored by computational technologies. Digitalization refers to the way in which social life is organized by and around digital technologies. Datafication refers to the practice of turning an activity, behavior, or process into meaningful data.

Digitization is the basis for collecting, preserving, and promoting cultural heritage and a new method for broader access to cultural heritage [4] The general legal framework of the EU digitization policy aims to give a new digital life to Europe's museums, libraries, and archives, among other things with the intention of improving the preservation of and access to cultural heritage [5].

The digital document storage system must be able to adapt to changing technologies or user requirements and it must be compatible with all currently available storage devices to ensure the long-term retention and archiving of documents [6], [7].

A well-designed data model, built in the spirit of the FAIR principles (Findability, Accessibility, Interoperability, and Reusability), can help homogenize the heterogeneity of different types of digital repositories and objects, as well as the diversity of designs, standards, and formats being considered. DRSI initiative [8] aims to share, archive, and disseminate public sector information and provide research datasets for scientific institutions. In [9] a semantic data model for a digital repository, also based on the FAIR data modelling principles, was described. It emphasized the need for conceptual designs and provided a list of requirements for document metadata. The value of models and metamodels in the context of digital repositories was also recognized in [10], where the design and implementation of the repository that supports storing and managing of artifacts such as metamodels, models, constraints, meta-data, specifications, transformation rules, code, templates, configuration or documentation, and their metadata was presented.

In [11], a project based on creating a website that facilitates the use and exchange access to a corpus of resources from various disciplines for scientific and educational purposes was described. The main objective of this research was to create a website that would help maintain and preserve Ecuador's linguistic and cultural heritage, focusing on document storage rather than knowledge extraction.

As can be seen from the literature review, the need for such repositories is constantly growing, especially in the cultural and socioeconomic fields. Moreover, this area of research has the potential for great social impact. Given this, the focus needs to be on all three relevant processes: the creation, collection, and storage of digital documents (digitization), the use of digital document repositories for this purpose (digitalization), and the use of analytics and knowledge extraction procedures to transform the data from digital repositories into meaningful information (datafication).

3. Phases of development of the MESOC repository

In this section the first step of repository development is described in more details. The phases include user requirements gathering, database modeling based on user requirements, and designing a database for the document repository. Fig. 1 shows a diagram of all project phases.

Fig. 1. Diagram of project phases.

3.1. User requirements

When developing a usable system, it is important to address the needs of the users. To understand user needs, one must know the user's tasks, goals, context of use, and skills [12]. Finally, user requirements emerge from this.

User requirements can be heterogeneous, gathered from different users with different linguistic and cultural backgrounds, and even gathered informally using different methods [13], as was the case in our study. An international project with users scattered in several European countries required a systematic approach. The goal of a database developer is to select the most appropriate requirements elicitation methods, specify a set of requirements, validate them with users, and negotiate their acceptance if necessary [14].

In designing our data repository, we found one of the recommendations from [15] very important to choose the most appropriate method for elicitation: Provide a range of query interfaces to accommodate various data search behaviors. To minimize the impact of heterogeneous tasks leading to heterogeneous requirements, users in our case study expressed their needs through descriptions of their intended different searches.

In the development of a Data Repository, such as a database or a data warehouse, the requirements identified from the informal specification are mainly used to define the conceptual model. [16].

The structure of the repository began to form after the systematic classification of user queries. Enriched with other data, such as metadata, we created a model that provided the needed information.

Some examples of user requests in the form of queries can be found in Table 1.

Table 1. Examples of user requirements.

User requirement
I want to access documents published in the selected year.
I need access to documents that use a selected technique (one or more) in their research, such as survey, focus groups, interview, ...
I need to access documents that pertain to the selected area of input, e.g., individual, community, social, ...
I need to access documents based on the specified keyword, which is one of the words from an open thesaurus.

In addition to these search queries, users defined criteria and rules for grouping documents from the repository and displaying them in a counted manner. Each time a user runs the procedure to count documents that meet a particular criterion, the procedure runs through the entire dataset for evaluation. In this way, users - policy makers - can evaluate trends and content of the collected documents.

3.2. Entity-relationship model for document storage

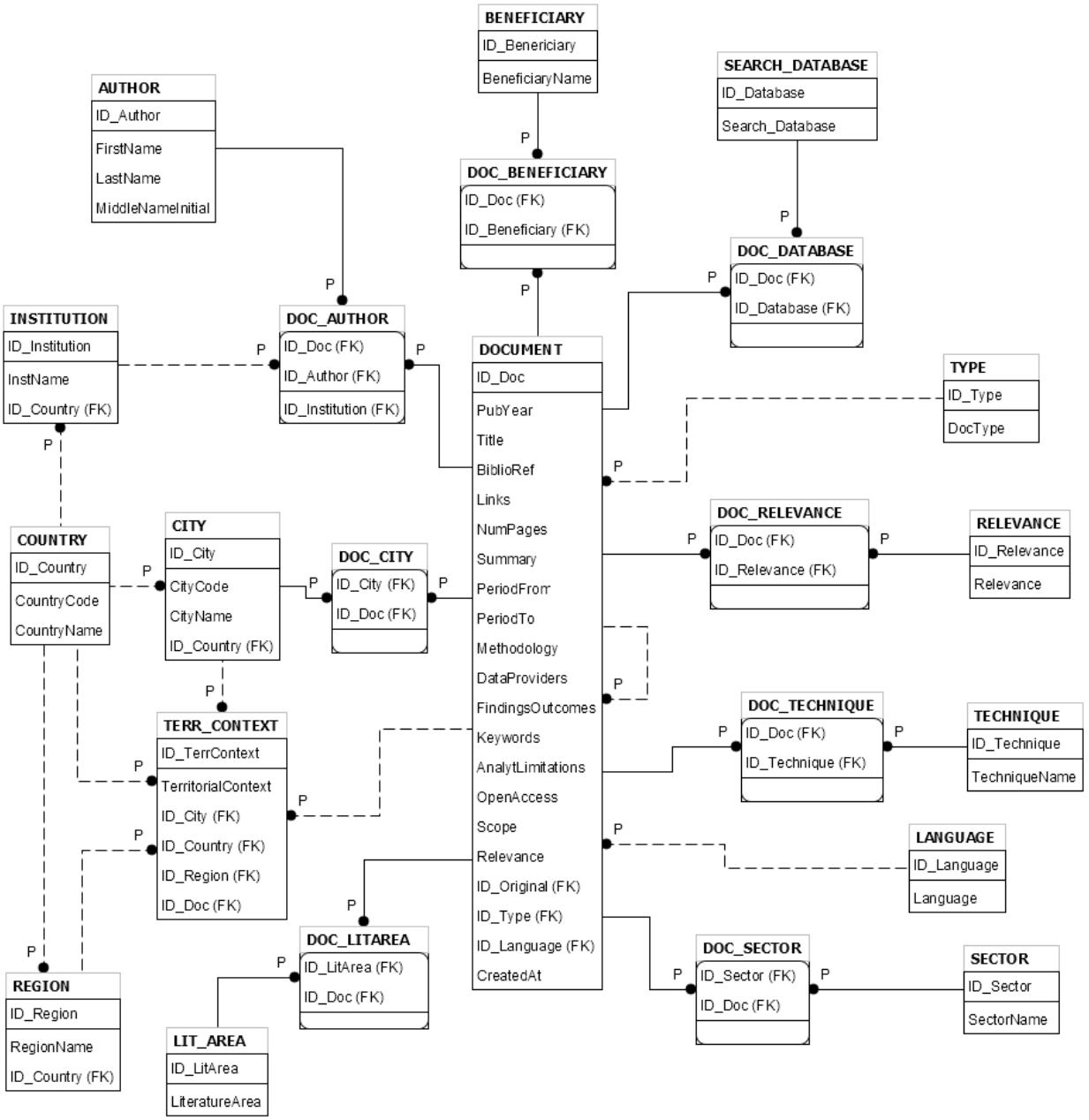

Fig. 2. ER model based on user requirements.

Entity-Relationship (ER) model was drawn in MySQL Workbench, using Integration DEFinition for Information Modeling (IDEF1X) notation. IDEF1X is a method for designing relational databases and a data modelling language for the development of graphical models which represent the structure and semantics of information within an environment or system [17]. IDEF1X model is very similar to the implementation of tables in the database - just by looking at the model, anticipated contents of the relational table in the database can be visualized. This is a big advantage in the context of database modelling and distinguishing physical and conceptual/logical representation. This

is also the reason for choosing IDEF1X to be used for our ER model. ER model was developed based on user requirement analysis and is shown in Fig. 2. The model contains all the relevant types of entities for a future document repository.

3.3. Relational model for document storage

Relational model was created using MIRIS methodology [18] and is shown in Table 2. It shows all the relations for a future database, along with their final set of attributes. Primary keys are underlined in a solid line and foreign keys in a dashed line.

Table 2. Relational model.

RELATION (Primary Key, Attribute, Attribute, …., Foreign Key)
DOCUMENT(ID_Doc, PubYear, Title, BiblioRef, Links, NumPages, Summary, PeriodFrom, PeriodTo, Methodology, DataProviders, FindingsOutcomes, Keywords, AnalytLimitations, OpenAccess, Scope, Relevance, ID_Original, ID_Type, ID_Language, ID_Sector, ID_Area, CreatedAt)
BENEFICIARY(ID_Beneficiary, BeneficiaryName)
DOC_BENEFICIARY(ID_Doc, ID_Beneficiary)
SEARCH_DATABASE(ID_Database, Search_Database)
DOC_DATABASE(ID_Doc, ID_Database)
TYPE(ID_Type, DocType)
RELEVANCE(ID_Relevance, Relevance)
DOC_RELEVANCE(ID_Doc, ID_Relevance)
TECHNIQUE(ID_Technique, TechniqueName)
DOC_TECHNIQUE(ID_Doc, ID_Technique)
LANGUAGE(ID_Language, Language)
SECTOR(ID_Sector, SectorName)
DOC_SECTOR(ID_Doc, ID_Sector)
LIT_AREA(ID_LitArea, LiteratureArea)
DOC_LITAREA(ID_Doc, ID_LitArea)
REGION(ID_Region, RegionName)
CITY(ID_City, CityCode, CityName, ID_Country)
COUNTRY(ID_Country, CountryCode, CountryName)
DOC_CITY(ID_Doc, ID_City)
TERR_CONTEXT(ID_TerrContext, TerritorialContext, ID_City, ID_Country, ID_Region, ID_Doc)
INSTITUTION(ID_Institution, InstName, ID_Country)
AUTHOR(ID_Author, FirstName, LastName, MiddleNameInitial)
DOC_AUTHOR(ID_Doc, ID_Author, ID_Institution)

Model is then expanded to include user access rights and metadata about users who create documents or users who access documents to perform keyword search. Also, metadata to facilitate keyword search is added into the model. Expanded ER model is shown in Fig. 3. In Table 3 relations added into a relational model are shown.

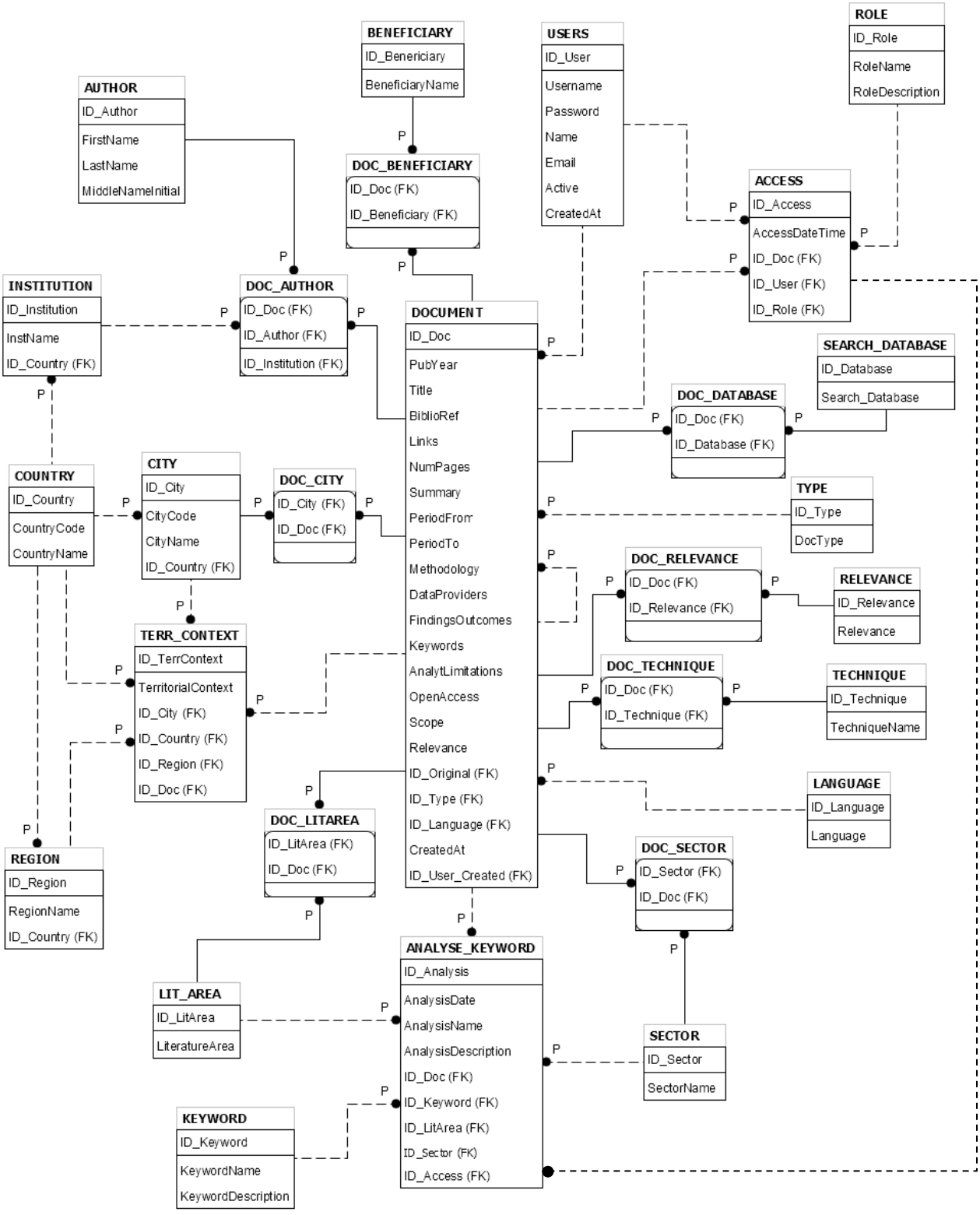

Fig. 3. Expanded ER model.

Table 3. Expanded relational model.

RELATION (Primary Key, Attribute, Attribute, …., Foreign Key)
USERS(ID_User, Username, Password, Name, Email, Active, CreatedAt)
ROLE(ID_Role, RoleName, RoleDescription)
ACCESS(ID_Access, AccessDateTime, ID_Doc, ID_User, ID_Role)
KEYWORD(ID_Keyword, KeywordName, KeywordDescription)
ANALYSE_KEYWORD(ID_Analysis, AnalysisDate, AnalysisName, AnalysisDescription, ID_Doc, ID_Keyword, ID_LitArea, ID_Sector, ID_Access)
DOCUMENT(ID_Doc, PubYear, Title, BiblioRef, Links, NumPages, Summary, PeriodFrom, PeriodTo, Methodology, DataProviders, FindingsOutcomes, Keywords, AnalytLimitations, OpenAccess, Scope, Relevance, ID_Original, ID_Type, ID_Language, CreatedAt, ID_User_Created)

3.4. Database development and beta test

MESOC document repository is based on the relational database MySQL, PHP programming language and Apache server, while Phalcon framework was used to create the web application. As the development of a web application is not the topic of this paper, we will describe only the technologies used to create a repository. PHP and MySQL are often used together to create dinamic web aplications because MySQL is easily accessed from PHP. There is a standard interface for calling MySQL procedures from PHP, so you don't need to know details of how the PHP language interfaces with the MySQL database. Also, they all (MySQL, PHP and Apache server) run on various computer types and operating systems, all are open source projects and have community support. In short, they work well together. [19][20]. After development of the first prototype, it was used to test and refine user requirements.

The prototype repository [21] currently contains 800 documents, classified into 11 predefined cultural sectors. The documents are also classified into 3 literature areas covering three crossover themes of the new European Agenda for Culture: 1) health and well-being, 2) urban and territorial renovation and 3) people's engagement and participation. 250 documents belong to two or more cultural sectors, and 50 belong to more crossover themes.

4. Conclusions and future work

This paper describes a research and innovation project MESOC that aims to propose, test and validate an innovative and original approach to measure the societal value and impact of culture, cultural policies and cultural practices. For this purpose, a document repository and a web application are developed. The main contributions of this paper are: a) a set of user requirements in the cultural domain collected from international users, b) a data model and database for the document repository, and c) a metadata model used to satisfy the user requirements.

Complex systems cannot be created without gathering and understanding user requirements. The users of this system are international, they do not work in the same organization, but it has been possible to elevate their tasks and requirements beyond the national level and meet the needs of all. Given the specificity of the system, user requirements were collected through user queries. The user queries were written in the form of specifying the data that the users needed from the system.

Based on the queries, the structure of the repository (conceptual model) was defined. After the database was built, the first prototype was created using selected tools. Based on the prototype, users were able to revise and add their requirements, obtaining a system that met all their needs. The tests of the prototype were performed with a set of 800 documents initially stored in the repository.

The limitation of this version of the system is its current audience. The system is intended for a wider audience, i.e. the public, but the tests of the prototype have been carried out so far only with a small group of users. On the other hand, the main benefit of the system is that it offers international users the opportunity to work together on a unique system within the framework of the European Agenda for Culture. This benefit has been achieved precisely through the use of metadata to provide additional descriptions of the documents in the repository, thus providing the basis for their retrieval against defined criteria and interpretation in the context of their use.

Future work includes further development of the system and its testing with a wider audience, implementation of additional features to facilitate user interaction with the system, easier keyword searching and document analysis, and knowledge extraction in general.

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 under project "uniri-drustv-18-182".

References

[1] MESOC project page. Last accessed 31.01.2022: https://www.mesoc-project.eu/

[2] Zhao, Man, Han-Teng Liao, and Si-Pan Sun. (2020) "An education literature review on digitization, digitalization, datafication, and digital transformation." *6th International Conference on Humanities and Social Science Research (ICHSSR 2020):* 302-306.

[3] Leonardi, Paul M., and Jeffery W. Treem (2020). "Behavioral visibility: A new paradigm for organization studies in the age of digitization, digitalization, and datafication." *Organization Studies* **41(12):** 1601-1625.

[4] Adane, Alehegn, Assefa Chekole, and Getachew Gedamu. (2019) "Cultural Heritage Digitization: Challenges and Opportunities." *International Journal of Computer Applications* **178 (33)**.

[5] Manikowska, Ewa. (2019) "Digitization: Towards a European Cultural Heritage." *Cultural Heritage in the European Union*: 417–444.

[6] Abbasova, Vasila Soltanaga. (2020) "Main Concepts of the Document Management System Required for Its Implementation in Enterprises." *ScienceRise* **(1):** 32-37.

[7] Manoj, G., Deep, I., Sahana, K. C., and Madhavi, R. P. (2015) "Online Document Repository System." *International Journal of Advance Research in Computer Science and Management Studies* **3(3):** 74-80.

[8] Parkoła, Tomasz, and Błażej Betański. (2020) "Towards a new generation of Digital Repository of Scientific Institutes." *Qualitative & Quantitative Methods in Libraries* **9(3)**.

[9] Panoutsopoulos, Hercules, Christopher Brewster, and Spyros Fountas. (2021) "A Semantic Data Model for a FAIR Digital Repository of Heterogeneous Agricultural Digital Objects". *2nd Integrated Food Ontology Workshop, Bolzano, Italy*.

[10] Milanovic, Nikola, Ralf-Detlef Kutsche, Timo Baum, Mario Cartsburg, Hatice Elmasgünes, Marko Pohl, and Jürgen Widiker. (2008) "Model&metamodel, metadata and document repository for software and data integration." *International Conference on Model Driven Engineering Languages and Systems*: 416-430. Springer, Berlin, Heidelberg.

[11] Verdugo, Priscila, Catalina Astudillo-Rodriguez, Jackelín Verdugo, Juan-Fernando Lima, and Santiago Cedillo. (2020) "Documentation and Scientific Archiving: Digital Repository." *International Conference on Applied Human Factors and Ergonomics*: 296-302. Springer, Cham.

[12] Courage, Catherine, and Kathy Baxter. (2005) "A practical guide to user requirements." Morgan Kaufmann.

[13] Cherotich Ronoh, Lilian, Geoffrey Muketha Muchiri, and Franklin Wabwoba. (2015) "Factors Affecting Requirements Elicitation for Heterogeneous Users of Infomation Systems." *International Journal of Computer Science Engineering and Technology* **5(3):** 35-39.

[14] Van Vliet, Hans. (2008) "Software Engineering - Principles and Practice, 3rd Edition." John Wiley&Sons, Chicester, UK.

[15] Wu, Mingfang, Fotis Psomopoulos, Siri Jodha Khalsa, and Anita de Waard. (2019) "Data Discovery Paradigms: User Requirements and Recommendations for Data Repositories." *Data Science Journal* **18(1)**: 3.

[16] Boukhari, Ilyes, Stéphane Jean, and Idir Ait-Sadoune, and Ladjel Bellatreche. (2018) "The role of user requirements in data repository design." *Int J Softw Tools Technol Transfer* **20**: 19–34.

[17] Computer Systems Laboratory of the National Institute of Standards and Technology. (1993) "Integration Definition for Information Modeling (IDEF1X) Standard" *FIPS Publication 184*.

[18] Pavlić, Mile. (2009) "Informacijski sustavi (Information systems)." *Department of Informatics, University of Rijeka*. Rijeka, Croatia.

[19] Kromann, Frank M. (2018) "Beginning PHP and MySQL: from novice to professional." Apress.

[20] Davis, Michele E., and Jon A. Phillips. (2007) "Learning PHP & MySQL: Step-by-Step Guide to Creating Database-Driven Web Sites." O'Reilly Media, Inc.

[21] Kovacic, Bozidar, Vanja Slavuj, and Martina Asenbrener Katic. (2022) "Analyzing the benefits of using a document repository to aid decision-making in the field of culture". *iSCSi – International Conference on Industry Science and Computer Sciences Innovation, Porto, Portugal*.