Towards Providing Analytical Services in the Web-Based Platform for Intelligent Cultural Content Management CultIS

Maxim Goynov [1], Detelin Luchev [1], Desislava Paneva-Marinova [1], Radoslav Pavlov [1], Konstantin Rangochev [1]

[1] *Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, bl. 8, Acad. Georgi Bonchev, Str., 1113 Sofia, Bulgaria*

Abstract **– This paper aims to present the efforts to design and develop the analytical services of the *CultIS* - a web-based platform for intelligent cultural content management. The *CultIS* Analytical Module's purpose is to provide advanced statistical data and analysis for the content usage and user interests in digital libraries powered by *CultIS*. Based on some of the top rated open-source solutions, the module provides both flexibility and high performance in data loading, report management and visualization. The paper discusses the motivation for the development of the *CultIS* Analytical Module, the selection of relevant technologies for the development, some specifics of the design and the implementation and some legal issues of data analytics. The proposed solution aims to face the problems and fill the gaps which exist in the current area of analytical solutions and platforms.**

Keywords **– CultIS, digital humanities, digital content management systems, analytical services, CLaDA-BG, research infrastructure.**

DOI: 10.18421/TEM144-05
https://doi.org/10.18421/TEM144-05

Corresponding author: Desislava Paneva-Marinova, *Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, bl. 8, Acad. Georgi Bonchev, Str., 1113 Sofia, Bulgaria*
Email: dessi@cc.bas.bg

Received: 30 January 2025.
Revised: 23 June 2025.
Accepted: 27 June 2025.
Published: 27 November 2025.

1. Introduction

The intelligent cultural content management platform *CultIS* [1] is a complex web-based environment for storing, managing and retrieving data from the social sciences and humanities. This web-based platform contains a rich set of technologies for the purpose of maintaining and managing a wide variety of digital cultural entities. Descriptive cataloging, flexible storage, standard and complex search and grouping in different intersections, virtual interactive presentation of objects and collections, are part of *CultIS* features and components. They have been developed in response to the growing expectations and functional needs of users of museums, libraries and other cultural institutions providing content from the historical and cultural heritage. The functions and components of *CultIS* also allow for flexible structuring and management of metadata, data, and objects, as well as subject indexing, intelligent data retrieval, collections curation and development, *etc*. Digitized copies of cultural and historical heritage objects of various types (such as text, graphics, audio, video, 3D formats or other media objects), and their corresponding metadata, can be stored and managed by the basic *CultIS* prototype.

As an infrastructure component of the Bulgarian National Interdisciplinary Research e-Infrastructure CLaDA-BG [2], part of the EU infrastructures CLARIN and DARIAH, the *CultIS* platform supports its mission to develop a technological national infrastructure for resources and technologies in the field of linguistic and cultural heritage.

Current implementations of *CultIS* are created in accordance with the needs and requirements of the National Library "Ivan Vazov" in Plovdiv, the "Peyo Yavorov" Regional Library - Burgas, the Central Library of the Bulgarian Academy of Sciences, the Encyclopaedia Slavica Sanctorum project, the Digital Library "Virtual Encyclopedia of Bulgarian Iconography", etc. [3], [4], [5].

This paper aims to present the efforts to design and develop the analytical services (viz. Analytical Module) of *CultIS* platform. The "analytics" is used to describe "statistical and mathematical data analysis that clusters, segments, scores, and predicts what scenarios are most likely to happen" [6]. Data analytics and business analytics are processes where, through the use of quantitative methods, the extraction of meaning from data is achieved to make informed decisions in various fields [7], [8], [9], [10], [11], [12]. Data science focuses on deriving meaning from systematized data using algorithms, statistical models, and computer programming, while business analytics is more concerned with extracting meaningful insights and visualizing data to facilitate the decision-making process, but both, in their effectiveness, seek to increase the quality and efficiency of decisions in a variety of systems. Four methods of business analytics can be identified as the main ones: Interpreting data over time to identify trends and patterns (descriptive), interpreting data over time to determine why something did or did not happen (diagnostic), using statistically analyzed data to predict future outcomes (predictive), generating test scenarios and applying other techniques to determine which actions will have the best outcome in a given scenario (prescriptive). Business analytics also includes data management, data visualization, business intelligence, etc. [13], [14], [15].

The benefits of using and providing analytical services in the web-based platform for intelligent cultural content management may include better user insights, user personalization, improved operational efficiency, improved performance, more effective presentation of information, new opportunities for dissemination of information, risk management, *etc*., as it is in other business areas [16], [17].

One of the modern computing methods that allows users to easily and selectively extract and query data to analyze it from different perspectives is OLAP (Online Analytical Processing) [18], [19]. Following the development of the standard database concept of OLTP (Online Transaction Processing) as a term, OLAP denotes a database technology that has been optimized for querying and reporting, as opposed to the processing transactions focus of OLTP. Used exclusively for data extraction from analytical repositories to support the decision-making process, OLAP tools enable users to quickly perceive and interpret the data and, as a consequence, to make fact-based decisions [20]. Since its appearance in the 1990s [21], the development of OLAP technological lines of research and application have gone in two directions - adapting OLAP to certain data formats and hybridizing it with other techniques for specific purposes [22].

CultIS Analytical Module is an extension for the *CultIS* platform, which aim is to provide advanced statistical data and analysis for the content usage and user interests in digital libraries powered by *CultIS*. Based on some of the top rated open-source solutions, the module provides both flexibility and high performance in data loading, report management and visualization.

The motivation of the development of the *CultIS* Analytical Module is presented in section 2 of this paper. The selection of relevant technologies for the development of the *CultIS* Analytical Module is discussed in section 3. Some specifics of the design and the implementation are demonstrated in section 4. Section 5 discusses legal issues of data analytics. Some future directions in the *CultIS* Analytical Module development are given in conclusion.

2. Motivation

Nowadays most of the analytical services available on the web are very powerful, easy to set up, easy to use, highly customizable and are perfect solution both for specific and general-purpose analysis. However, for some specific domains or report requirements, such solutions can be inconvenient for implementation, use and maintenance. Also, the complexity of such solutions can impact the performance of the web platform being tracked.

Another issue may be the fact, that all data is stored on third party servers, which usually is related to privacy issues and concerns.

And finally, the most popular analytical solutions on the web are just being blocked by default from most of the privacy-oriented web browsers and web extensions.

The *CultIS* analytical module tries to face all the above-mentioned issues, trying to provide highly customizable reporting, storing data on the same DL server infrastructure making it as secure as the main DL is. The module architecture guarantees zero additional network loads and performance issues for the end users, because the whole data collection process is executed in a separate thread at the *CultIS* back-end.

CultIS addresses data analysis challenges: Data accessibility (adequate storage and maintenance of data for use by less experienced data professionals and analysts in humanitarian institutions); maintenance of data quality (optimization of time, effort and resources for proper maintenance and use of data); data security; choosing the right tools for data analysis (meeting the needs and infrastructure of the users), and in this way *CultIS* provides the users of the system with the most adequate solution for their needs.

3. Selection of Relevant Technologies for Analytical Services Implementation

As a part of the *CultIS* strategy to stick to open-source technologies, the *CultIS* analytical module was built using two open-source technologies – an online analytical processing (OLAP) database and a data visualization tool.

In order to achieve optimal performance, research from a short-list of the most popular and high rated technologies was executed and the following tools were chosen:

- ClickHouse [23] is an open-source column-oriented OLAP database that enables its users to generate powerful analytics, using SQL queries, in real-time. As of 2024 ClickHouse is rated as one of the fastest and most powerful OLAP (online analytical processing) databases [24].
- Superset [36] is a modern data exploration and data visualization platform. Superset can replace or augment proprietary business intelligence tools for numerous teams. Superset integrates well with a variety of data sources.

4. Design and Implementation

The *CultIS* core platform uses MongoDB as a main database engine. Data in MongoDB is structured, but its structure is usually very complicated, which makes it hard to achieve good analytical performance. In order to make data more analysis-friendly, non-relational data and structures had to be "flattened" using relational models.

The process of transferring data from the online transaction processing (OLTP) database (MongoDB) to the OLAP database is called an ETL (Extract, Transform, Load) process.

A relational schema which includes the necessary data from the OLTP DB is defined. A very common high-performance solution using normalized data is to use the so-called snowflake schema (Figure 1).

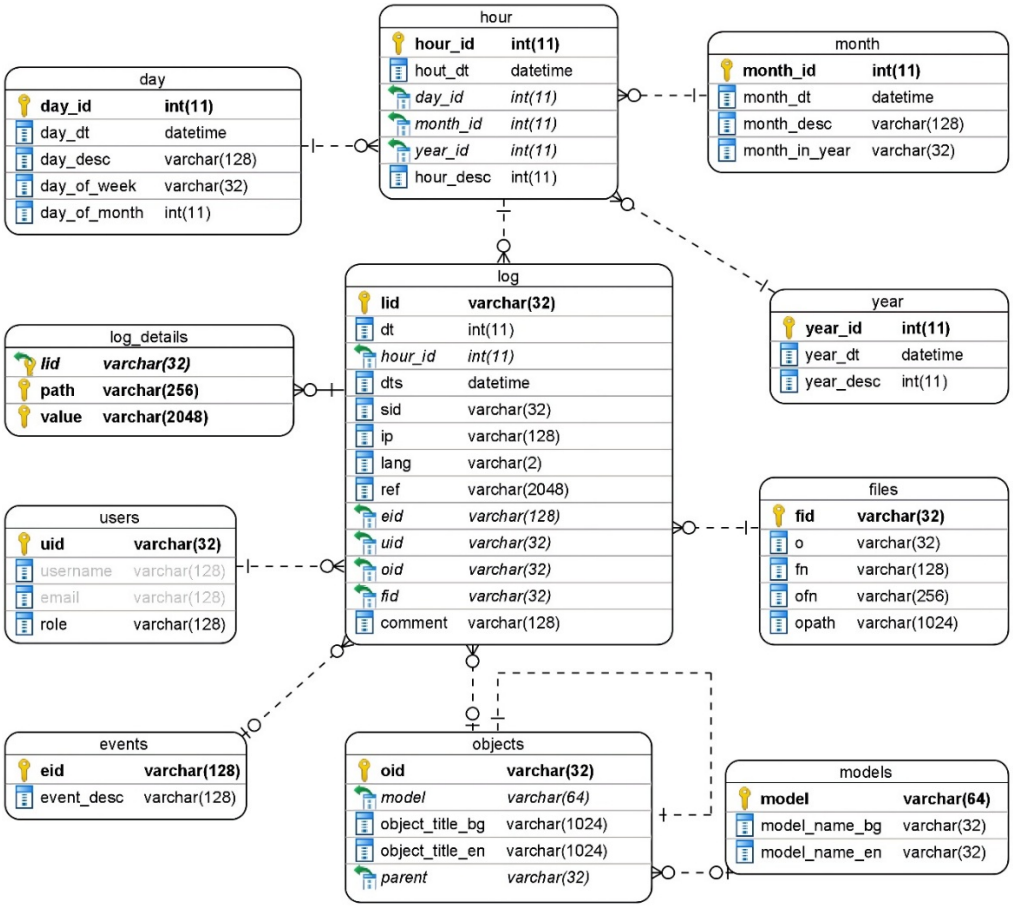

Figure 1. Snowflake diagram of the proposed relational mode

The snowflake model is an extension of the star schema model. Like its name suggests, its structure is based on one main table (fact table), related to many secondary tables (dimensions).

In a snowflake model there is always more than one level of nesting, unlike the star schema, where the nesting level is always single.

Analytical data logged by the *CultIS* platform is stored in only one MongoDB collection. However, it is not possible to map it to a single relational table, because for some data events, like user searches, one element of this collection may contain arrays of data, related to the specific search criteria. So, the minimalistic working solution was to split the log collection into two relational tables, the main one (fact table) – log and a dimension table – log_details, containing all additional data in a key-value pair manner. The typical structure of a log data row always includes information about the date and time the event was logged, user session id, type of event, objects related to the event (if any), files related to the event, user id related to the event (user personal data is anonymized), *etc*. So, it is intuitive to build the other part of the snowflake schema (the other dimensions), in accordance with the above-mentioned relations.

The first try for executing the ETL process was by using MongoDB as the main OLAP database. Data loading of more than 4M rows of data took about ten minutes, which is relatively good result for such volume of data. However, building the main dataset by joining all of the tables presented on the diagram, showed extremely bad results and it was necessary to switch to the ClickHouse analytical database. The ETL process using ClickHouse was executed in less than a minute. Joining all tables and visualizing data now takes less than one second and this directly contributes to the good performance of the data visualization tool – Apache Superset [36].

4.1. Data Visualization – Diagrams and Reports for Specific Digital Library

The following diagrams are small part from all reports implemented for the digital library of Public Library “Ivan Vazov” – Plovdiv. Reports are based on data collected between 10.2022 (the official start of the platform) and 11.2024.

Visualization tools support decision-making processes and improve information services in libraries, are similar to other institutions [25].

All reports can be filtered by date, user roles, languages, events, etc. Apache Superset allows one filter to be applied at the same time for all charts having the same dataset.

Therefore, the relational model and dataset is structured in a way that all of the charts created for a specific DL use the same dataset. Thus, the users who view the reports will have a convenient and consistent data filtering features.

Using data analysis, Public Library "Ivan Vazov" – Plovdiv will be able to obtain additional valuable information about users’ behavior, usage patterns and collection trends, leading to more efficient and personalized services. This would help uncover patterns and trends to inform decision-making processes, identify areas for improvement and reallocate resources accordingly. On this basis, models could be developed to forecast demand, recommend resources, and increase user satisfaction and library functionality [26], [27]. With the increasing integration of digital technologies and the effective use of measurement approaches, especially data analytics, the value and impact of libraries in the digital age is increased and the library itself is more successfully included in the concept of "Library 4.0" [28], [29].

Figure 2 shows the main user activities such as searching, viewing, creating and updating object during the last year. There were two peaks in 01.2024, which were related to Google Search indexing (Googlebot). It turned out that the current implementation is not able to tell whether a request is from a regular user or from a search engine crawler. So, there will be a future enhancement in order to discriminate bots from other users.

Figure 2. User activities over time

Figure 3 is similar to Figure 2, but it shows in a clearer way the user requests distribution among days in a given period.

It is convenient to see the direct relations and outcomes between user activities in the DL and social network campaigns, physical or virtual events, feasts, etc.

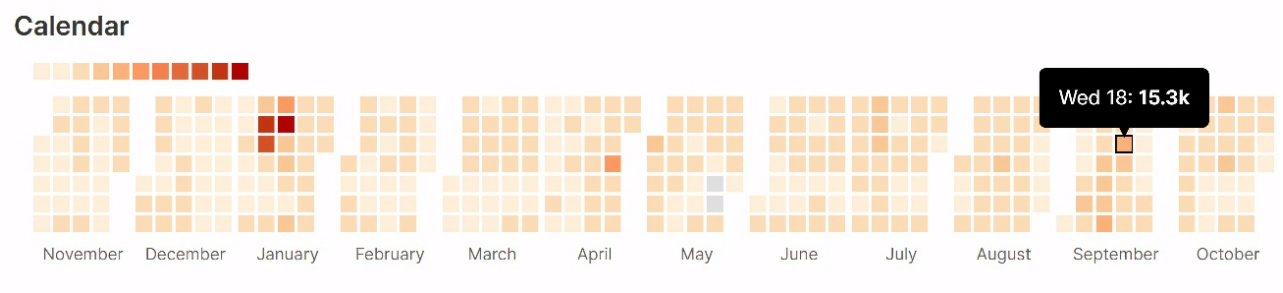

Figure 3. Library usage by days (01.11.2023-31.10.2024)

One of the most important functional features in the digital library of Public Library "Ivan Vazov" is the ability to search in the text documents, books, periodicals, archives, etc.

The chart on Figure 4 shows the most popular phrases used in the full-text search feature.

Figure 4. Most popular search phrases

Data on Figure 5 shows the most used metadata attributes for content searches. For example, the first row, containing "$s.issue.year.≥" means that 26300 searches were made with the query "show me all periodical issues issued after certain year".

The selected row "$s.identification.author" shows the number of searches performed over the author attribute.

Search attributes

Key	Request count
$s.issue.year.≥	26.3k
$s.issue.year.≤	19.8k
$s.issue.no	7.06k
$s.identification.publishing.year.≥	4.21k
$s.issue.month.=	3.65k
$s.identification.name	3.26k
$s.identification.publishing.year.≤	2.76k
$s.identification.kind	1.99k
$s.identification.author	1.7k
$s.identification.publishing.place	1.45k
$s.identification.subcollection	801

Figure 5. Most used metadata search criteria

5. Legal Issues

The issue of data security is important in the development and implementation of online analytical processing (OLAP) systems and data warehouses [30], and *CultIS* strives to comply with all requirements and trends in this area [31], [32], [33], [34], [35].

Data for the analysis is collected by *CultIS* during the consumption of APIs by the UI. The whole data collection process is aligned with the privacy and data regulations available in the EU and worldwide. Detailed privacy policy documents and technical explanations are available for the users of the specific digital libraries. User consent is taken into account when applying data collection rules. All statistics provided by the module are anonymous in terms of user accounts and user data.

6. Conclusion

Current research in cultural heritage content management systems is primarily focused on the creation of large collections of multimedia resources and regular tools for their indexing and retrieval. However, these systems need to provide more than advanced content maintenance and retrieval services. They should support users in monitoring content, acquiring knowledge, and better satisfying their needs, interests and desires. This paper presents an extension of the current functionality with content analytic services. The main goal is to access implicit and hidden data, content, rules and facts, dependencies and trends valid for the content in the repository, to synthesize and summarize the collected data to be used in various research and learning.

The presented study contribution is related to the creation of a robust architecture of a highly-customizable analytical module based on validated techniques and models by defining and implementing efficient OLAP and ETL processes and the integrations between the target components. The proposed solution aims to face the problems and fill the gaps which exist in the current area of analytical solutions and platforms. As a result, an easy and intuitive system is implemented which allows researchers not to worry about the data retrieval and data credibility, but to focus on the meaning of the data and to examine it through different perspectives and points of view. The paper presents the *CultIS* Analytical Module, which is in a stage of development and implementation in some of the current release of CLaDA-BG digital libraries. So far, there are some development challenges, for example, those related to distinguishing a human from a bot, which are under investigation.

It was also found useful to provide analytics according to geo-location, which is under design. Another important part of the goal in the future is to provide not only user-oriented analysis, but also to implement object-centric analysis, closely related to digital library objects, their metadata, indexed content, dependencies, etc.

Acknowledgements

The research, presented in this article, was carried out within ClaDA-BG, the Bulgarian national research infrastructure for resources and technologies for language, cultural and historical heritage, integrated within CLARIN EU and DARIAH EU, funded by the Ministry of Education and Science of the Republic of Bulgaria (support for the Bulgarian National Roadmap for Research Infrastructure), Grant number: D01-324/01.12.2023, https://clada-bg.eu/bg/.

References:

[1]. CultIS. (n.d.). *CultIS*. Institute of Mathematics and Informatics, Bulgarian Academy of Sciences. Retrieved from: https://CultIS.math.bas.bg/en [accessed: 22 December 2024]

[2]. CLaDA-BG. (n.d.). *E-infrastructure for Bulgarian language and cultural heritage.* Bulgarian Academy of Sciences & CLARIN-DARIAH Bulgaria. Retrieved from: https://clada-bg.eu/bg/ [accessed: 23 December 2024]

[3]. Paneva-Marinova, D., et al. (2024). Full-Fledged Access and Usability of Content in a Digital Cultural Heritage Library: Approaches, Paradigms, and Implementation. *ACM Journal on Computing and Cultural Heritage*, *17*(1), 1-12. Doi: 10.1145/3631135

[4]. Dimitrova, M., et al. (2023). "Encyclopaedia Slavica Sanctorum" Ten Years Later: Main Trends and Questions. *Digital Presentation and Preservation of Cultural and Scientific Heritage*, *13*, 87-96. Doi: 10.55630/dipp.2023.13.8

[5]. Goynov, M., et al. (2024). CultIS: Web-based Platform for Intelligent Cultural Content Management. *Digital Presentation and Preservation of Cultural and Scientific Heritage*, *14*, 19-36. Doi: 10.55630/dipp.2024.14.1

[6]. Gartner. (n.d.). *Analytics*. Information Technology Glossary. Retrieved from: http://www.gartner.com/it-glossary/analytics [accessed: 24 December 2024]

[7]. The Investopedia Team. (2025). *Data Analytics: What It Is, How It's Used, and 4 Basic Techniques*. Investopedia. Retrieved from: https://www.investopedia.com/terms/d/data-analytics.asp [accessed: 25 December 2024]

[8]. Madugula, S., et al. (2023). Big data for the comprehensive data analysis of IT organizations. *The Journal of High Technology Management Research*, *34*(2), 100465. Doi: 10.1016/j.hitech.2023.100465.

[9]. Hussain, A. A., Al-Turjman, F., & Sah, M. (2021). Semantic web and business intelligence in big-data and cloud computing era. *Innovations in Smart Cities Applications Volume 4: The Proceedings of the 5th International Conference on Smart City Applications*, 1418-1432. Doi: 10.1007/978-3-030-66840-2_107

[10]. Acito, F., & Khatri, V. (2014). Business analytics: Why now and what next?. *Business Horizons*, *57*(5), 565-570. https://doi.org/10.1016/j.bushor.2014.06.001

[11]. Nam, D., Lee, J., & Lee, H. (2019). Business analytics adoption process: An innovation diffusion perspective. *International Journal of Information Management*, *49*, 411-423. https://doi.org/10.1016/j.ijinfomgt.2019.07.017

[12]. Schmitt, M. (2023). Deep learning in business analytics: A clash of expectations and reality. *International Journal of Information Management Data Insights*, *3*(1), 100146.

[13]. Yalcin, A. S., Kilic, H. S., & Delen, D. (2022). The use of multi-criteria decision-making methods in business analytics: A comprehensive literature review. *Technological forecasting and social change*, *174*, 121193.

[14]. Holsapple, C., Lee-Post, A., & Pakath, R. (2014). A unified foundation for business analytics. *Decision support systems*, *64*, 130-141. https://doi.org/10.1016/j.dss.2014.05.013

[15]. Pinder, J. P. (2022). *Introduction to business analytics using simulation*. Academic Press.

[16]. Ghasemaghaei, M., Hassanein, K., & Turel, O. (2017). Increasing firm agility through the use of data analytics: The role of fit. *Decision Support Systems*, *101*, 95-105. https://doi.org/10.1016/j.dss.2017.06.004

[17]. Ashrafi, A., et al. (2019). The role of business analytics capabilities in bolstering firms' agility and performance. *International journal of information management*, *47*, 1-15. https://doi.org/10.1016/j.ijinfomgt.2018.12.005

[18]. Biscobing, J. (2023). *OLAP (online analytical processing)*. TechTarget. Retrieved from: https://www.techtarget.com/searchdatamanagement/definition/OLAP [accessed: 27 December 2024]

[19]. Sharma, P., & Girdhar, P. (2014). Online Analytical Processing (OLAP). *Journal of Advance Research in Computer Science & Engineering*, *1*(3), 1–4. https://doi.org/10.53555/nncse.v1i3.520

[20]. Burstein, F., et al. (2008). Online analytical processing (OLAP) for decision support. *Handbook on Decision Support Systems 1: Basic Themes*, 259-276. https://doi.org/10.1007/978-3-540-48713-5_13

[21]. Codd, E. F., Codd, S. B., & Salley, C. T. (1993). *Providing OLAP to User-Analysts: an IT mandate*. E.F. Codd Associates. Retrieved from: http://www.estgv.ipv.pt/paginaspessoais/jloureiro/esi_aid2007_2008/fichas/codd.pdf [accessed: 28 December 2024]

[22]. Darmont, J., Novikov, B., Wrembel, R., & Bellatreche, L. (2022). Advances on data management and information systems. *Information Systems Frontiers*, *24*(1), 1-10.

[23]. ClickHouse (2024). Retrieved from https://clickhouse.com/ [accessed: 12 December 2024]

[24]. RisingWave. (2023). *A Comprehensive Overview of Top 8 Real-Time OLAP Databases*. [24]. RisingWave. Retrieved from: https://risingwave.com/blog/a-comprehensive-overview-of-top-8-real-time-olap-databases/ [accessed: 29 December 2024]

[25]. Zakaria, M. S. (2021). Data visualization as a research support service in academic libraries: An investigation of world-class universities. *The Journal of Academic Librarianship*, *47*(5), 102397. https://doi.org/10.1016/j.acalib.2021.102397

[26]. Oakleaf, M., et al. (2017). Academic libraries & institutional learning analytics: One path to integration. *The Journal of Academic Librarianship*, *43*(5), 454-461. https://doi.org/10.1016/j.acalib.2017.08.008

[27]. Litsey, R., & Mauldin, W. (2018). Knowing what the patron wants: Using predictive analytics to transform library decision making. The Journal of Academic Librarianship, 44(1), 140-144. https://doi.org/10.1016/j.acalib.2017.09.004

[28]. Labangon, D., & Manabat, A**.** (2019). Establishing connections, bridging the gap: Library 4.0 and its role in digital humanities. *Proceedings of the DLSU Research Congress 2019*, 1–9. Retrieved from: https://www.dlsu.edu.ph/wp-content/uploads/pdf/conferences/research-congress-proceedings/2019/lli-II-015.pdf [accessed: 10 January 2025]

[29]. Noh, Y. (2023). A study on the discussion on Library 5.0 and the generation of Library 1.0 to Library 5.0. *Journal of Librarianship and Information Science*, *55*(4), 889-905. https://doi.org/10.1177/09610006221106183

[30]. Wang, L., & Jajodia, S. (2008). Security in Data Warehouses and OLAP systems. *Handbook of Database Security: Applications and Trends*, 191-212. https://doi.org/10.1007/978-0-387-48533-1_8

[31]. Hussain, F., et al. (2020). Enterprise API security and GDPR compliance: Design and implementation perspective. *IT professional*, *22*(5), 81-89. https://doi.org/10.1109/MITP.2020.2973852

[32]. Doty, P. (2020). Library analytics as moral dilemmas for academic librarians. *The Journal of Academic Librarianship*, *46*(4), 102141. https://doi.org/10.1016/j.acalib.2020.102141

[33]. Abdmeziem, F., Boukhedouma, S., & Oussalah, M. C. (2021). On the data security of information systems: Comparison of approaches and challenges. *Computational Science and Its Applications – ICCSA 2021*, 240-255. Springer. https://doi.org/10.1007/978-3-030-86970-0_18

[34]. Prashar, N., Hooda, S., & Kumar, R. (2022). Current status of challenges in data security: A review. *International conference on cybersecurity in emerging digital era*, 3-13. Springer. https://doi.org/10.1007/978-981-99-5080-5_1

[35]. OWASP. (2023). *OWASP Top 10 API Security Risks – 2023*. Open Web Application Security Project. Retrieved from: https://owasp.org/API-Security/editions/2023/en/0x11-t10/ [accessed: 14 January 2025]

[36]. Superset. (2024). *Home page.* Superset Apache. Retrieved from https://superset.apache.org/ [accessed: 18 January 2025]