CULTURAL LEAF: A LOD PORTAL FOR EXPLORING THE CULTURAL HERITAGE

Ilie Cristian DOROBĂȚ[1], Vlad POSEA[2], Radu BONCEA[3]

The paper introduces the Cultural Leaf Semantic Portal whose aim is to provide a consolidated user-friendly solution for exploring the 36,132 Romanian artifacts and the 1,086 cultural entities that house them. Using the eCHO framework, the data collected from official sources are subjected to a transformation process through which they are sanitized, temporal expressions are normalized, thus reducing the number of forms in which the centuries and millennia are expressed from 1,733 to only 67, and then they are transformed into semantic structures. Finally, the resulting data is accessible via both a SPARQL endpoint and a semantic portal.

Keywords: Cultural Heritage, Linked Open Data, Semantic Web.

1. Introduction

Nowadays, as the need for readily available information continues to rise, the prominence of digitalization is steadily increasing across different sectors. Although digital libraries cannot entirely substitute traditional heritage collections, they serve as a bridge between cultural institutions and information consumers. The significance of digital libraries is apparent both for everyday individuals who are searching for particular Cultural Heritage Objects (CHOs) that intrigue them, and for the professionals in the field who need extensive and easily attainable sources of information.

Over time, various platforms and tools have been developed to facilitate data migration into LOD. BookSampo [1] is one of the longest-standing Digital Humanities projects in the Sampo portals series[4]. Launched as early as 2011, BookSampo was developed around a proprietary knowledge graph that enables the representation of features describing Finnish fiction literature. In the year 2022, this portal adopted a new look, migrating from the UI developed in Drupal, which implemented a simple text-search engine, to a new semantic UI with

[1] Ph.D. candidate, Dept. of Computer Science, National University of Science and Technology POLITEHNICA Bucharest, Romania, e-mail: ilie.dorobat@stud.acs.upb.ro
[2] Conf., Dept. of Computer Science, National University of Science and Technology POLITEHNICA Bucharest, Romania, e-mail: vlad.posea@cs.pub.ro
[3] Head of Research Department, National Institute for Research and Development in Informatics - ICI, Bucharest, Romania, e-mail: radu.boncea@ici.ro
[4] https://seco.cs.aalto.fi/applications/sampo/, accessed: September 2023

integrated data analytics tools and faceted search [2], developed through the integration of the Sampo-UI Framework[5] [3]. WarVictimSampo 1914–1922 [4] is another relatively new project within the same series, which, through the Sampo-UI Framework, implements an UI similar to the one implemented by BookSampo. The only difference lies in the knowledge graph used and the available facets.

The present work examines the study undertaken to consolidate all the data sources presented above into a single web portal designed to facilitate users' access to cultural data. To achieve this, using the eCHO framework[6] developed as part of a previously undertaken study [5], the data sources describing CHOs and cultural institutions are consolidated and transformed into Linked Open Data (LOD), allowing us to enrich them with terms from the DBpedia knowledge graph and standardize temporal expressions [6][7]. The data thus produces is stored in a triplestore that provides users with access for querying through a SPARQL endpoint. Additionally, to meet the needs of data consumers for swiftly navigating without requiring a strong technical background, the *Cultural Leaf Semantic Portal* was developed. It offers a user-friendly environment that allows both general users and those in the field to discover details related to the Romanian cultural heritage.

2. Data Considerations

The National Institute of Heritage (INP) is the public institution in Romania entrusted with the task of establishing and overseeing databases as well as digital assets pertaining to tangible, intangible, and digital cultural heritage [8]. Despite INP's recent and significant efforts to revitalize online portals that host cultural information, navigating through these portals still remains a demanding task due to the diverse nature of the published data which is spread across numerous portals, encompassing at least 5 resources categorized into the following 3 distinct groups based on their particular characteristics:

i. CHOs: the portal clasate.cimec.ro [7] and a collection of 10 datasets [8] structured according to LIDO XML, a widely accepted harvesting schema [9] that provides a more comprehensive view of CHOs.
ii. Cultural Institutions: the portal ghidulmuzeelor.cimec.ro [9] and a tabular dataset [10] which offers additional information concerning the cultural institutions in Romania.

[5] https://github.com/SemanticComputing/sampo-ui, accessed: September 2023
[6] https://github.com/iliedorobat/enriching-cultural-heritage-metadata, accessed: September 2023
[7] http://clasate.cimec.ro/, accessed: September 2023
[8] https://data.gov.ro/dataset?organization=institutul-national-al-patrimoniului, accessed: September 2023
[9] http://ghidulmuzeelor.cimec.ro/, accessed: September 2023
[10] https://data.gov.ro/dataset/ghidul-muzeelor-din-romania, accessed: September 2023

iii. Map: the portal map.cimec.ro[11] which offers a cartographic representation of 5 databases, encompassing the database for Romanian cultural institutions among them.

The cultural data consists of 36,132 descriptions of CHOs, identifying the following 3 types of events in which they were involved: i) collecting; ii) finding; iii) production. Fig. 1 showcases the statistical data related to these events, categorized based on the domain of the CHOs and the event type while the Table 1 encompasses the specific values used to create the chart. It is noteworthy that events classified as *collecting* are exclusively associated with CHOs belonging to the *Natural Sciences* domain. This pertains to the phase wherein biological material such as living organisms (reptiles, insects, crustaceans, dicotyledons, monocotyledons, etc.) and physical objects (native elements, silicates, sulfides, sulfates, phosphates, carbonates, tectosilicates, etc.) are *collected* for purposes of preservation, research, and eventual presentation in an organized exhibition.

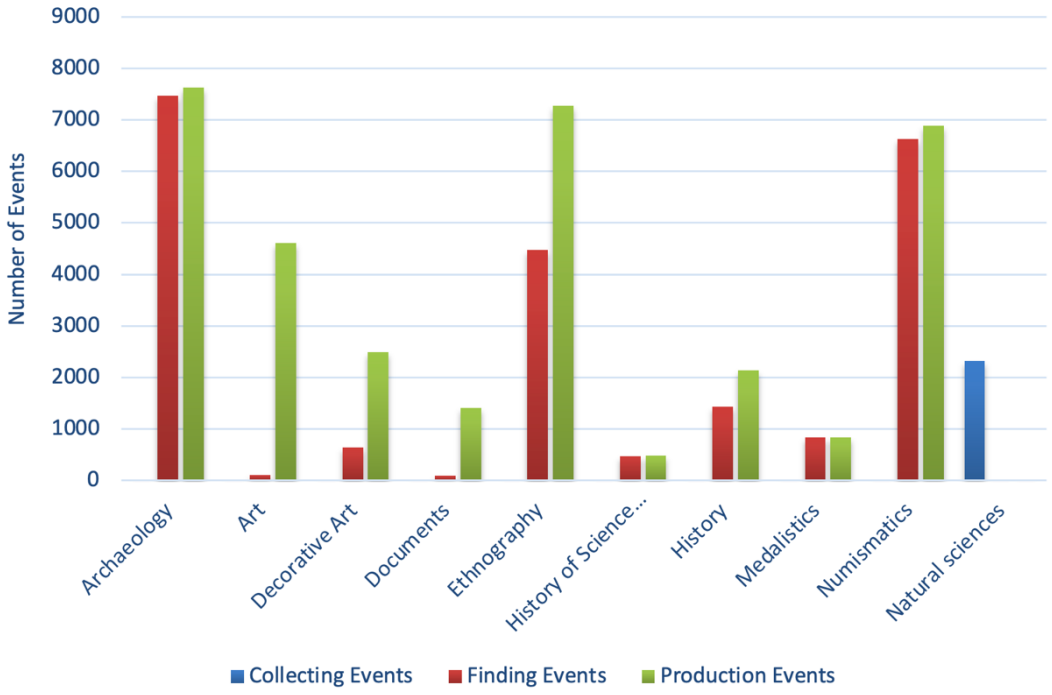

Fig. 1. Types of events associated with each field of study.

For instance, in the case of insect studies, the entomologist embarks on this process by first identifying the specific group of insects under investigation and conducting preliminary research (referred to as the insect detection stage). Subsequently, the entomologist proceeds to observe and capture the insects

[11] https://map.cimec.ro/Mapserver/, accessed: September 2023

without causing them physical harm (referred to as the insect observation and capture stage), and only then selects adult specimens for in-depth examination (referred to as the insect collection stage). Following collection, the insects undergo specialized treatment (referred to as the preservation and preparation stage), and only then can the specialist process the acquired data [10]. In practice, each stage of the study of biological materials constitutes an individual event that can be documented by researchers, but regrettably, as indicated in Table 1, the analyzed datasets exclusively encompass one advanced stage of this procedure, namely the collection stage.

While collecting events are exclusive to CHOs involving the preservation of biological materials and physical items, finding and production events are distinctive to CHOs belonging to the other 9 study areas. From a chronological perspective, production events mark the initial stages in which these CHOs are engaged. As the name suggests, these types of events offer insights into the creation process of CHOs, encompassing details like the location and date of manufacture, materials utilized, and information about the creator of the piece, among others. Conversely, finding events are outcomes of actions that occur after the production stage, supplying information about the process of uncovering and highlighting CHOs such as the event's location and date, information about the individuals involved, and more.

Table 1

Types of Events Associated with Each Field of Study

Category	Total Number of CHOs	Event Type		
		Collecting	Finding	Production
Archaeology	7,644	0	7,473	7,632
Art	4,613	0	11	4,610
Decorative art	2,502	0	647	2,497
Documents	1,407	0	10	1,407
Ethnography	7,269	0	4,482	7,269
History of science and technology	486	0	469	484
History	2,141	0	1,434	2,140
Medalistics	844	0	843	844
Numismatics	6,886	0	6,633	6,886
Natural sciences	2,340	2,327	0	0
TOTAL	**36,132**	**2,327**	**22,002**	**33,769**

Of course, as outlined in Ordinance no. 43 of January 30, 2000 on the protection of archaeological heritage and the declaration of archaeological sites as areas of national interest [11], activities that occur after the production stage encompass a wide range of activities. These range from surveying, identification, archaeological excavations, inquiry, harvesting, to recording and scientific exploitation. It's also worth noting that, for newly discovered objects to become

part of the national heritage, Law 182 dated October 25, 2000 on the protection of the mobile national cultural heritage [12], mandates their classification and registration within either the Inventory of the National Cultural Heritage Fund or the Inventory of the National Cultural Heritage Treasury. Subsequent to classification, further steps include preservation, safekeeping, security of mobile cultural artifacts, conservation, restoration, and circulation (*"lending"* cultural artifacts for exhibitions or cultural initiatives).

Regarding the data comprising cultural institutions, a total of 1,087 museums and public collections were found with descriptions ranging from basic information such as name and address to more intricate details like the history of the institutions and the buildings that house them, etc.

3. Normalization of Temporal Expressions

By leveraging the third-party TENF library[12] developed as part of eCHO framework, the standardization of temporal expressions became a very easy task. To accomplish this task, the TENF library implements an approach of recognizing temporal expressions based on regular expressions. The standardization process begins, as expected, with the extraction of temporal expressions from datasets, then continuing with classification of temporal expressions into the following categories according to their degree of similarity: i) statements whereby the time periods referred to cannot be accurately distinguished (e.g. "dinastia xxv", "nesemnat", etc.); ii) periods of time that are subject of interpretation (e.g. "pleistocen', "epoca de bronz", etc.); iii) calendar dates; iv) calendar years; v) shapes of representation of centuries and millennia.

After classification, temporal expressions have undergone a sanitization process consisting in the elimination of accents specific to the Romanian language, the elimination of inaccurate time periods (e.g.: "anul 15=1802/1803"), the standardization of some less frequently used forms of temporal expressions ("s:", "sc", "se." and "sex" have been replaced by "sec.") and the notations expressing anno domini (e.g.: "d. hr", "p. hr", etc.) and before Christ (e.g.: "î. hr", "î. chr", "a. chr", etc.) have been replaced by the following unique notations: "__AD__" and "__BC__".

During the next stage, regular expressions are manually build, and then applied to sanitized values to standardize them using a single pattern, the DBpedia resources (e.g.: dbr:19th_century, dbr:20th_century, etc.). To gain a clearer insight into the consequences of this conversion, Table 2 consolidates the data on the frequency of the five temporal expression categories (epoch, unknown, date, year, timespans – centuries and millenniums) during both stages, before and after standardization.

[12] https://github.com/iliedorobat/timespan-normalization, accessed: September 2023

Table 2

Incidence of Time Period Types

Type of Time Period	Unique Occurrences of Unformatted Temporal Expressions	Unique Occurrences of Formatted Temporal Expressions
epoch & unknown	240	240
date	2,702	*903
year	2,015	
centuries	1,619	61
millenniums	114	6
TOTAL	****6,690**	**1,210**

*Calendar dates have been transformed into annual calendar units
**The overall count of unique occurrences is 6,630. However, owing to certain temporal phrases encompassing both timespans (centuries and millenniums) and dates/years (e.g. *"s: xvii; 1685, decembrie 23"*), they have been segregated into two categories.

It is evident that the standardization process has notably reduced the frequency of notations denoting centuries and millennia. Initially, with 1,619 distinct occurrences for expressions indicating centuries and 114 for millennia, standardization has ultimately condensed these figures to a mere 61 notations for centuries and 6 for millennia. Essentially, this process has substantially reduced the count of distinct notations representing the constructions used to define centuries and millennia by over 96%, a discrepancy originating from the manner in which these notations were recorded by the human operators responsible for collecting and digitizing data pertaining to CHOs in the Romanian national heritage.

For instance, for time intervals like *"4/4. sec. xix.- 1/4. sec. xx"*, *"4/4 sec.xix. sfârșitul sec.al xix-lea și începutul sec.al xx-lea."* the DBpedia concepts dbr:19th_century and dbr:20th_century will be employed. Similarly, for the interval *"sec. xvii-xix"* all relevant DBpedia concepts denoting the centuries encompassed in this range will be utilized, namely dbr:17th_century, dbr:18th_century, and dbr:19th_century. Ultimately, users interested in querying the system to discover CHOs associated with the 19th century no longer need to construct queries targeting the specifics of every notation containing the 19th century. Instead, they only have to refine their search results by a single resource, which in this instance is dbr:19th_century.

The entire set of 6,630 temporal expressions extracted from the LIDO-compliant datasets were manually validated, the result being shown in Fig. 2. As it can be seen, only 4.22% of the temporal expressions were erroneous, partially wrong, or partially valid, of which:

i. 3.62% are either invalid (unknown) or subject to specialized investigation (epoch), being effectively excluded from this process.
ii. 0.35% are partially normalized or partially wrong.
iii. 0.25% are totally wrong.

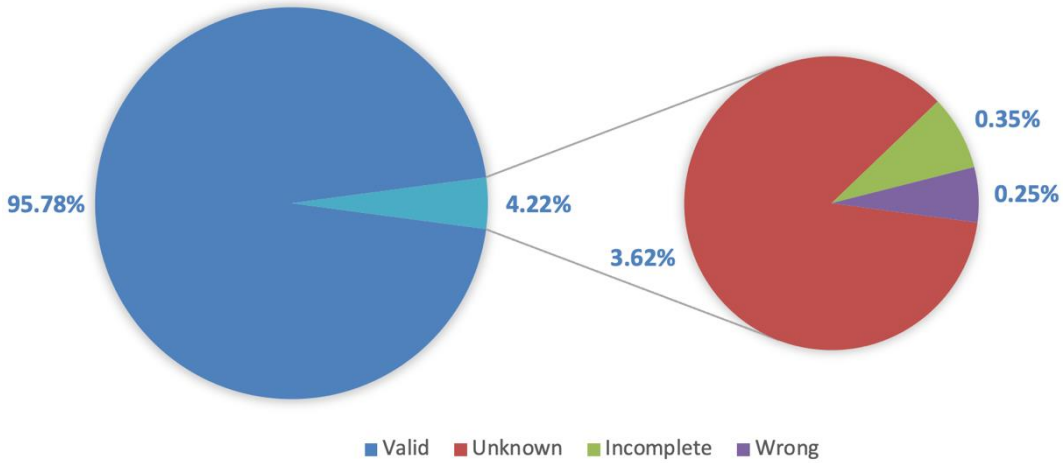

Fig. 2. Statistics of errors occurred in the normalization of unique temporal expressions.

For instance, when normalizing the expression *"2 a.chr - 14 p.chr"* the result leads to the DBpedia resource dbr:14 which provides information about significant events and figures from the year 14 AD. This outcome is partially correct because in this case the year 14 AD is within the processed interval of 2 BC to 14 AD. However, in the case of the expression *"sec. xx; [22]33"* the result is partially incorrect because even if the resource dbr:20th_century, providing details about the 20th century, is correctly extracted, the resource dbr:19 is also extracted, when in fact, the resource dbr:1933 should have been extracted instead.

While in the case of the two classifications presented above, the distinction lies in whether an incorrect resource is identified or not, the third classification pertains to cases where no valid resource can be pinpointed. A clear instance of this can be observed when normalizing the temporal expression *"s: 19; 4/4"*, which leads to the extraction of the resource dbr:19 which is entirely incorrect as this resource pertains to the year 19 AD rather than the 19th century.

Despite these instances, the outcomes confirm the proficiency of the TENF library not only in processing straightforward temporal expressions but also in handling intricate time intervals which may encompass abbreviations, ordinal number suffixes, as well as both Roman and Arabic numerals.

4. The Cultural Leaf Semantic Portal

Cultural Leaf Semantic Portal [13] is an environment designed to simplify the access to data describing CHOs and cultural institutions for various audiences including researchers, students, and the general public. While the portal doesn't replace traditional research methods [14], users benefit from an intuitive and exclusive environment that allows them to focus their efforts on more significant activities, rather than juggling between different portals and datasets.

Fig. 3 illustrates the underlying workflow of the *Cultural Leaf Semantic Portal*. The eCHO framework is employed for the parsing, standardization, and transformation of the resources into LOD, subsequently loaded into a triplestore accessible through a SPARQL endpoint. While users navigate within the *Cultural Leaf Semantic Portal*, their interactions initiate diverse processes. These user requests are converted into SPARQL queries which are used to fetch data from the triplestore, and the outcomes are processed before being sent to the users. All these requests are performed by the LOD service[13] that serves the portal.

The *Cultural Leaf Semantic Portal* is crafted as a single-page application, developed utilizing Angular 15 [14]. It incorporates the Leaflet library [15] for designing the Romanian map and Bootstrap 5 [16] for seamless integration of a modern design. In terms of functionality, the portal provides three distinct perspectives for exploring the underlying knowledge graph, namely: i) exploring cultural institutions; ii) exploring CHOs; iii) statistics on the time periods during which CHOs have been involved.

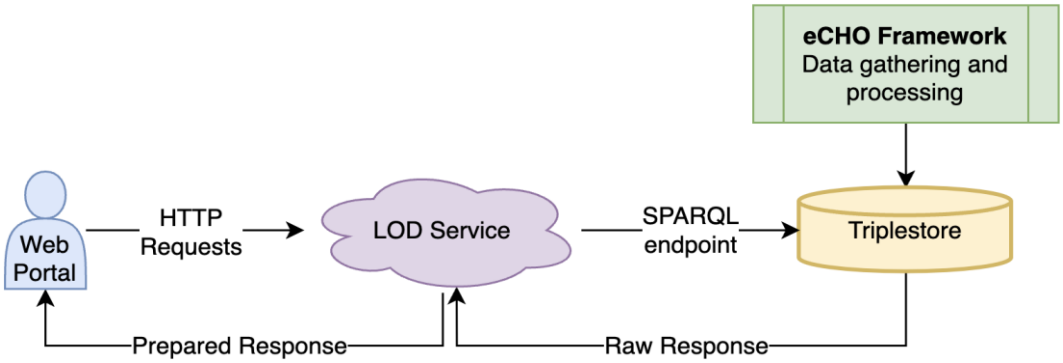

Fig. 3. The workflow of the Cultural Leaf Semantic Portal.

The first perspective shown in Fig. 4 is readily accessible from the portal's main entry-point. It showcases a map of Romania, outlining its counties and employing a range of blue shades to indicate the total number of collections and museums from each county's territorial jurisdiction. For instance, Sălaj County is depicted with a pale shade of blue, almost resembling white, signifying its relatively lower count of only 4 collections and museums. Conversely, counties like Alba, Suceava, Neamț, Harghita, Prahova, and Dâmbovița, which accommodate a significant number of such institutions (ranging from 42 to 61 within each county), are visualized in a notably deeper hue of blue.

[13] https://github.com/iliedorobat/cultural-leaf-service, accessed: September 2023
[14] https://angular.io, accessed: September 2023
[15] https://leafletjs.com, accessed: September 2023
[16] https://getbootstrap.com/docs/5.2/getting-started/introduction, accessed: July 2023

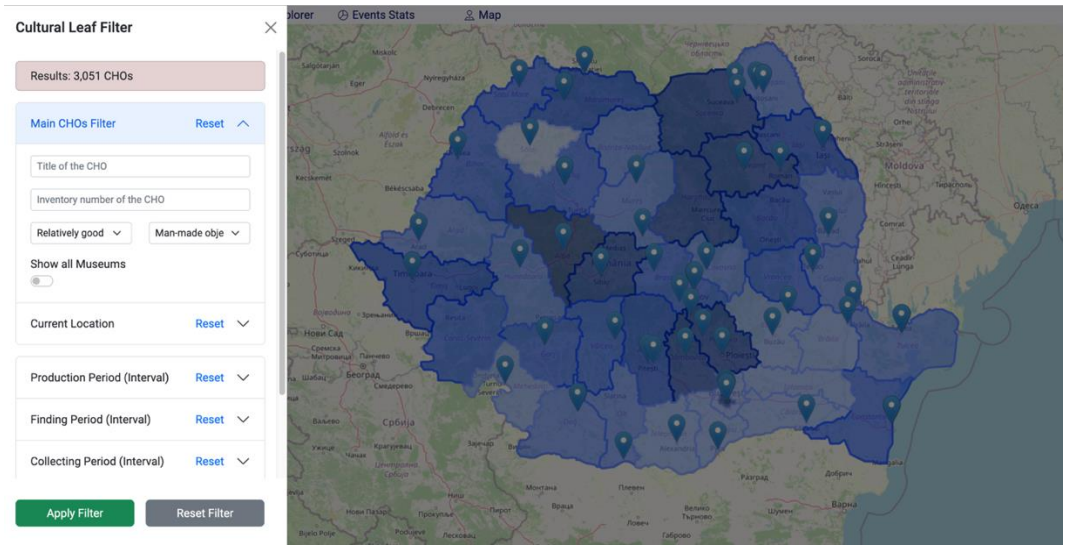

Fig. 4. The map displaying the markers of cultural institutions hosting the filtered CHOs [13].

The markers shown on the map correspond to recognized collections and museums, and their placement is determined by their geographical coordinates, but if the source data lacks this information, the markers are positioned based on the geographical coordinates of the localities where the cultural institutions are located. Clicking on a marker triggers a popup that displays the name of the cultural institution it represents and a button enabling users to view comprehensive details about that institution.

This perspective also contains the global search system which is operationalized via a sidebar that encompasses diverse facets [2], enabling users to refine their searches according to general features like title, inventory number, status and current location of CHOs and some specific to certain categories such as: i) medal shape applied only to CHOs in the medal category; ii) age, era and gender of biological objects. Upon applying the facets, the total count of identified CHOs is showcased at the upper section of the sidebar, and solely the distinct markers of these CHOs persist on the map. However, if the *"Show all Museums"* option is selected will cause all markers describing cultural institutions to be exhibited on the map.

The second perspective shown in Fig. 5 allows the exploration of CHOs and can be accessed either from the sidebar by clicking on the filtered result or from the main menu by choosing the *"CHOs Explorer"* option. In either scenario, the facets are retrieved from the global search system and employed within the exploration interface, which comprises a table showcasing the names, categories, and inventory numbers of the CHOs, in addition to the global search system. Finally, *Cultural Leaf Semantic Portal* allows users to explore statistics about the top 10-time intervals in which CHOs were involved.

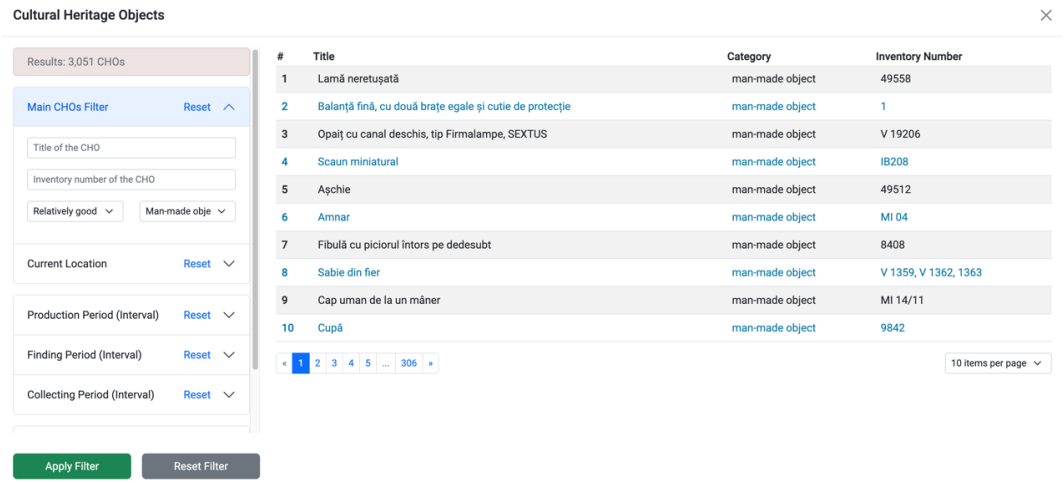

Fig. 5. CHOs Explorer after applying the filter.

5. Conclusions

One prominent limitation of both the cultural data stored in DSpace archives and LIDO-compliant datasets is the absence of connections between the represented resources, which is natural considering that in both cases the data is represented through XML data structures that do not offer such capabilities. While these methods have become the industry standard for representing cultural data, with LIDO offering the possibility of rich descriptions of cultural assets across various domains like art, architecture, cultural history, history of technology, and natural history, they primarily serve as a common ground for metadata sharing. Migrating them to EDM and ultimately publishing them as LOD brings about a plethora of advantages such as:

- flexibility – enables real-time updates to the schema of a triplestore without the need for any downtime or redesign and removes the necessity of creating additional entities like tables in SQL to establish many-to-many relationships.
- efficiency – SPARQL offers a direct approach to managing complex queries, in contrast to SQL queries, which tend to become convoluted and inefficient when the database lacks appropriately designed columns for joins and lacks proper indexing [15].
- integration – through the utilization of a shared framework (RDF and SPARQL) and standardized formats like N-Triples or N-Quads, the process of importing/exporting data and migrating from one storage system to another is greatly simplified.

- discoverability – triplestores facilitate inferencing, enabling the derivation of additional knowledge from the existing dataset based on a predefined set of inference rules [16].

By using the eCHO framework, the data sourced from INP sources was consolidated and transformed into LOD. This procedure led to an enhancement of the vocabulary by incorporating terminology from the DBpedia knowledge graph, alongside the standardization of temporal expressions using a unique format, dramatically decreasing the number of expressions detailing centuries and millennia from 1,733 to a mere 67. Ultimately, the outcome of this process is stored within a triplestore, and users can access them via a SPARQL endpoint. Additionally, by utilizing a triplestore such as GraphDB, users have access to an automatic inference mechanism, allowing them to deduce additional knowledge.

The *Cultural Leaf Semantic Portal* allows users to explore data concerning CHOs and cultural institutions without having to navigate through various individual portals and datasets provided by INP. To facilitate navigation, the portal provides a map, a tabular view containing comprehensive information about CHOs, and illustrative charts that aid in statistical analysis. Furthermore, the portal has integrated a faceted search system that empowers users to refine their searches, including the ability to apply filters based on specific time periods associated with CHOs. This feature holds significant importance in the analysis of cultural data, being absent in the portals provided by INP due to the heterogeneous nature of temporal expressions.

Acknowledgements

The results presented in this article have been funded by the Ministry of Investments and European Projects through the Human Capital Sectoral Operational Program 2014-2020, Contract no. 62461/03.06.2022, SMIS code 153735.

R E F E R E N C E S

[1] *A. Ahola, E. Hyvönen, H. Rantala,* A user interface model for digital humanities research: Case BookSampo – Finnish Fiction Literature on the Semantic Web, in: Proceedings of ESWC 2023, poster and demo papers, Springer, 2023.

[2] *D. Tunkelang,* Faceted search, in: Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool, Palo Alto, California, 2009.

[3] *E. Ikkala, E. Hyvönen, H. Rantala, M. Koho,* Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces, in: Semantic Web - Interoperability, Usability, Applicability 13(1), 2022, pp. 69–84. doi:10.3233/SW210428.

[4] *H. Rantala, I. Jokipii, M. Koho, E. Ikkala, J. Tuominen, E. Hyvönen,* Building a Linked Open Data Portal of War Victims in Finland 1914-1922, in: DHN 2020 Digital Humanities in the Nordic Countries. Proceedings of the Digital Humanities in the Nordic Countries 5th Conference, **vol.** 2612, 2020, pp. 310-317.

[5] *I. C. Dorobăț, V. Posea,* Enriching the Cultural Heritage Metadata Using Historical Events: A Graph-Based Representation, in: Doucet, A., Isaac, A., Golub, K., Aalberg, T., Jatowt, A. (eds) Digital Libraries for Open Knowledge, TPDL 2019, **vol.** 11799 of Lecture Notes in Computer Science, Springer, Cham, 2019, pp. 344-347. doi:10.1007/978-3-030-30760-8_30.

[6] *I. C. Dorobăț, V. Posea,* Raising the Interoperability of Cultural Datasets: The Romanian Cultural Heritage Case Study, in: Themistocleous, M., Papadaki, M., Kamal, M.M. (eds) Information Systems, EMCIS 2020, **vol.** 402 of Lecture Notes in Business Information Processing, Springer, Cham, 2020, pp. 35-48. doi:10.1007/978-3-030-63396-7_3.

[7] *I. C. Dorobăț, V. Posea,* The Power of Regular Expressions in Recognizing Dates and Epochs, in: The 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), Pitesti, Romania, 2021, pp. 1-3. doi:10.1109/ECAI52376.2021.9515139.

[8] Government Decision no. 593/2011, 2011. URL: https://legislatie.just.ro/Public/DetaliiDocument/129426.

[9] *R. Stein, O. Balandi,* Using LIDO for Evolving Object Documentation into CIDOC CRM, in: **vol.** 2, no. 1 of Heritage, 2019, pp. 1023-1031. doi:10.3390/heritage2010066.

[10] *L. Pârvulescu,* Studiul Entomologic. Tehnici de Colectare și Prelucrare a Materialului, Entomologie - Lucrări Practice (Material Collection and Processing Techniques, Entomology - Practical Works), West University of Timișoara. [Online] Available from https://biologie.uvt.ro/laboratoare/Ento_1.pdf [accessed: March 2023].

[11] The Romanian Government, Ordinance no. 43 of January 30, 2000 on the Protection of Archaeological Heritage and the Declaration of Archaeological Sites as Areas of National Interest. [Online] Available from https://www.cdep.ro/pls/legis/legis_pck.htp_act_text?idt=22093 [accessed: April 2023].

[12] The Romanian Parliament, Law 182 of October 25, 2000 on the Protection of the Mobile National Cultural Heritage. [Online] Available from https://www.cdep.ro/pls/legis/legis_pck. htp_act_text?idt=24709 [accessed: April 2023].

[13] *I. C. Dorobăț, V. Posea, R. Boncea,* Enhancing the Semantic Access and Visualization for Cultural Heritage: The Romanian Case Study, in: The 19th International Scientific Conference eLearning and Software for Education (eLSE), Bucharest, Romania, 2023.

[14] *H. Rantala, E. Ikkala, I. Jokipii, M. Koho, J. Tuominen, E. Hyvönen,* WarVictimSampo 1914–1922: A Semantic Portal and Linked Data Service for Digital Humanities Research on War History, in: Harth, A., et al. The Semantic Web: ESWC 2020 Satellite Events, ESWC 2020, **vol.** 12124 of Lecture Notes in Computer Science, Springer, Cham, 2020, pp. 191-196. doi:10.1007/978-3-030-62327-2_33.

[15] A mapping of SPARQL onto conventional SQL. (n.d.). [Online] Available from https://www.w3.org/2008/07/MappingRules/StemMapping [accessed: September 2023].

[16] What is Inference? (2022, May 30). Ontotext. [Online] Available from https://www.ontotext.com/knowledgehub/fundamentals/what-is-inference/ [accessed: September 2023].