Information Research - Vol. 31 No. iConf (2026)

Multi-viewpoint knowledge graphs for minority cultural heritage: the case of online Jewish museum collections

Sara Minster, Maayan Zhitomirsky-Geffet, and Inna Kizhner

DOI: https://doi.org/10.47989/ir31iConf64200

Abstract

Introduction. This study addresses the representation of ethnic minority cultures in online museum collections, which often reflect diverse viewpoints. We propose a data-driven methodology to construct a large-scale multi-viewpoint knowledge graph, using Jewish cultural heritage as a case study.

Method. We developed an LLM-based pipeline that combines object typing, named entity recognition, relation extraction, enrichment, and clustering.

Results. An analysis of 647,951 records and 178,444 extracted subjects from the collections of Jewish museums across the globe revealed diverse thematic emphases: Israel and the Netherlands prioritised religious themes, while others highlighted everyday life. Surprisingly, only Australia emphasised the Holocaust.

Conclusions. The central contribution of this study is the development of a knowledge organisation system capable of tracing major trends and identifying patterns in the polyvocality of perspectives. The methodology provides quantifiable, scalable analysis of multi-viewpoint cultural heritage, extendable to other minorities.

Introduction

As a result of the large-scale digitisation of cultural heritage data by GLAM institutions (Galleries, Libraries, Archives, and Museums) and commercial entities, numerous isolated cultural heritage databases have been created. Knowledge about the cultures of ethnic minorities is inherently diverse and multi-viewpoint; however, while some perspectives dominate, others remain underrepresented or concealed. Several recent studies have examined the representation of ethnic minorities in online museum collections (Kizhner et al., 2022; Zhitomirsky-Geffet et al., 2023). Nevertheless, these analyses were limited to only two or three major collections. Moreover, cultural data from the many GLAM institutions worldwide has never been represented through a multi-viewpoint ontology and knowledge graph.

Ontologies and knowledge graphs are increasingly employed to represent cultural heritage information, standardising and linking data available in heterogeneous formats, languages, and structures to enable cross-database integration and analysis. They support heterogeneous information representation and interoperability (Dijkshoorn et al., 2018; Shoilee et al., 2023). Two foundational ontologies were developed by large international research groups: CIDOC-CRM, the outcome of a decade-long project (Doerr, 2003), and the Europeana data model (Doerr, 2011). To incorporate the idea of perspectivism into knowledge organisation systems (KOS), multi-perspective and multi-viewpoint KOS and ontology models have recently been introduced in the literature (Baclawski et al., 2018; Gnoli, 2011; Zhitomirsky-Geffet, 2019). Ideally, an inclusive KOS should represent both past and present discourses as well as the range of possible contemporary contexts and worldviews (Zhitomirsky-Geffet, 2019). Examples of ontologies designed to capture polyvocality and multiple perspectives in relation to cultural heritage include the '*upper-level cultural ontology*' (Vlachidis et al., 2018) and an ontology supporting multi-perspective provenance research (Shoilee et al., 2023).

In this study, we propose a new paradigm for investigating the cultural heritage of ethnic minorities by means of a large-scale multi-viewpoint knowledge graph that both organises and integrates data while reflecting the multiplicity of perspectives on minority cultures. To this end, we developed a generic methodology for the automated data-driven construction of a multi-viewpoint ontology that enables the characterisation, comparison, contrast, and synthesis of different perspectives on the representation of an ethnic minority in online museum collections worldwide. The constructed ontology serves as the basis for a systematic comparative analysis of changes in representation across geographies, time periods, and subjects in different countries. Each museum collection is evaluated in terms of its representational balance and diversity relative to other collections and to the ontology as a whole, using distributional analysis of the ontological data.

As a case study, this paper examines the representation of Jewish culture in dedicated Jewish museums with publicly accessible online collections worldwide. Jewish communities have been established across numerous countries since the Babylonian exile in the sixth century BCE (https://www.nli.org.il/en/discover/judaism/jewish-communities), creating a long and complex history of cultural dispersion. Consequently, the question of what constitutes the culture of an ethnic minority such as Judaism is inseparable from the processes of collecting, digitising, tagging, linking, and publishing artworks online—processes that frequently take place in diverse, and at times even hostile, cultural environments (Kizhner et al., 2022). In our study, we conceptualise these environments as distinct viewpoints determined by the countries of origin of the museum collections. Building on this conceptualisation, we address the following interrelated research questions: (1) How can a multi-viewpoint knowledge graph be systematically constructed from online museum collections? (2) Which subjects and locations are most and least frequently represented across the multi-viewpoint knowledge graph as a whole and within each museum and country? and (3) Which source countries exhibit similar representational patterns?

Institutional background and research data

Jewish museums

The first Jewish museums were established at the end of the nineteenth century in Central Europe. The Jewish Museum in Vienna was founded in 1895, followed by Budapest in 1896, Frankfurt and Hamburg in 1897, and Prague in 1906. Even within the relatively homogeneous agenda of Jewish museums, there were notable differences in emphasis across countries: the Jewish museum in Prague was created in response to anti-Semitic events, whereas the Vienna museum sought to minimise references to antisemitism in its displays.

Throughout the twentieth century, Jewish museums presented a diverse array of themes reflecting both the priorities of Jewish communities and broader developments in museum practices. Some collections focus on ritual objects as aesthetically significant items or as evidence of disappearing communities. Others showcase contemporary Jewish artists to engage secular audiences. Certain museums document Jewish communities with attention to varied cultural and geographical backgrounds, while others preserve artifacts reflecting local Jewish experiences, such as the Jewish-American experience. These collections often include historical documents and family archives (Kochavi, 2022).

Research data

The analysis draws on object metadata published on the websites of fourteen Jewish museums, encompassing over half a million records (Table 1). While Jewish museums exist globally, data collection was limited in cases where museums lacked online collections, denied permission to scrape data, or prohibited data extraction under their terms of use. This study focuses exclusively on publicly accessible digital collections, as the research question concerns the multi-perspectival and polyvocal knowledge available on museum websites. Consequently, internal museum databases were not used, even when such data were potentially available. Metadata for the retrieved objects were systematically scraped and exported as a CSV file for analysis. The museum data comprise multiple metadata fields associated with each individual record, capturing heterogeneous information about various descriptive entities, including locations, time periods, and subjects. Importantly, the data enable the representation of these entities and their interrelationships as knowledge graphs, both at the level of individual museums and across the entire set of museums.

Museum country of origin	Jewish museum name	Number of metadata records for the museum
Denmark	Danish Jewish Museum	3,513
Ukraine	Chernivtsi Museum of the History and Culture of Bukovinian Jews	413
Brazil	Museu Judaico de São Paulo	9,757
Australia	Sydney Jewish Museum	114,329
Israel	Museum of Jewish People in Tel Aviv	11,416
Canada	Jewish Museum and Archives of British Columbia	44,177
Estonia	Estonian Jewish Museum	4,354
Netherlands	Jewish Museum Amsterdam	71,086
Belgium	Jewish Museum of Belgium	73,083
Germany	Jewish Museum Berlin	135,783
Greece	Jewish Museum of Greece	104,776
USA	Jewish Museum New York	20,001
Switzerland	Jewish Museum of Switzerland	5,484
South Africa	South African Jewish Museum	49,779
Total number of records for all Jewish Museums		647,951

Table 1. A list of the Jewish museums under study and the number of records for each of them.

Method

Data acquisition and pipeline design

The initial step in our workflow involved crawling and scraping fourteen museum websites to extract raw data from their publicly accessible online collections. This process yielded heterogeneous metadata that required extensive transformation before it could support advanced querying and analysis. To address this, we developed a multi-stage pipeline that integrates large language model (LLM) services, Python-based data processing libraries, and enrichment techniques. The central objective was to convert raw museum metadata into a coherent and interconnected knowledge graph, which would then serve as the basis for systematic comparative analysis.

The pipeline is implemented entirely in Python and relies on several key libraries. *Instructor*, a Python library designed for extracting structured data from LLM outputs, was employed in combination with *Pydantic* to guarantee type safety and schema validation. Pandas provided functionality for initial data ingestion and preprocessing of CSV exports from the crawlers. The *networkx* library was central to the project, serving as the primary data structure for constructing and manipulating the knowledge graph; it enabled graph operations such as node relabelling, merging, and traversal. To streamline usability, we adopted *Typer*, which facilitated the creation of a user-friendly command-line interface for running different pipeline stages. The workflow also leveraged the *Gemini 2.0 Flash LLM*, which was responsible for classification, entity recognition, relation extraction, and temporal interpretation tasks.

Knowledge graph extraction

The extraction process involved four key stages: object typing, named entity recognition (NER), relation extraction (RelEx), and graph assembly.

Object typing: Each record's metadata was passed to Gemini 2.0 Flash to classify its fundamental type (e.g., '*postcard*,' '*photograph*,' '*newspaper*'). This classification added an essential semantic layer to the dataset.

Named entity recognition (NER): Record text was processed with the LLM to identify and categorise key entities. The system extracted persons, locations, organisations, periods, materials, and subjects. For materials specifically, a secondary LLM validation step ensured that identified items corresponded to physical substances such as *paper*, *bronze*, or *ink*.

Relation extraction (RelEx): Entities and the original text were then processed through an LLM prompt to establish directed relationships. Relations included '*made of*' (record → material), '*depicts*' (record → subject), and '*created in*' (record → period).

Graph assembly: Each record was transformed into a small graph of nodes and edges. These were incrementally merged into a single knowledge graph, with duplicate nodes consolidated during a final merge step to ensure consistency.

Knowledge graph enrichment

After extraction, the graph underwent a sequence of enrichment operations to standardise and refine the data.

Data cleaning: Placeholder entities with values such as *unknown*, *n/a*, or *anonymous* were identified and removed.

Title generation: For records lacking descriptive titles, Gemini 2.0 Flash generated new ones based on available metadata. These were clearly marked with a [GENAI] prefix.

Period standardisation: Ambiguous or fuzzy temporal expressions (e.g., '*Late 19th Century*,' '*circa* 1920') were interpreted by the LLM and converted into the standardised <start_year>-<end_year> format.

Period merging: Records assigned to overlapping time spans were merged by computing the intersection of ranges, producing consolidated period nodes.

The enrichment stage critically depended on carefully engineered prompts, structured into *System Prompts* (high-level task rules) and *User Prompts* (record-specific data). For example:

NER prompt:

> System: Extract entities from the document. Rules: materials must be physical (paper, bronze, ink); subjects must be concise; object type must match common museum descriptors.
>
> User: {record_metadata}

RelEx prompt:

> System: Extract record → entity relations. Apply *'made of'* exclusively to record–material pairs.
>
> User: {record_metadata} with {list_of_entities}

Preliminary results of this pipeline, including frequency distributions of entities and relationships, are visualised at https://jewish-bubbles.jhn.ngo/.

Subject distribution analysis

A central question of our study concerns how Jewish culture is thematically represented across museums. Each institution provides subject metadata for its items, but this information is recorded as uncontrolled free-text terms, which produces substantial variation and redundancy. For example, similar items might be tagged *'synagogue,' 'Jewish temple,'* or *'house of prayer.'* Such inconsistency makes traditional summarisation approaches impractical.

Generative topic models such as latent Dirichlet allocation (LDA) (Blei et al., 2003) are widely used for discovering themes, but they depend on word co-occurrence patterns within longer documents. Since our corpus consists primarily of short terms (1–3 words), co-occurrence structure is sparse, and LDA performs poorly (Fan et al., 2023). Following recent best practices for short-text analysis (Likhitha et al., 2019), we instead adopted a three-stage workflow:

Vector representation: Subject terms were embedded using *SentenceTransformers* (model *all-MiniLM-L6-v2*), which is optimised for short, domain-specific phrases. Embeddings were normalised to improve comparability.

Clustering: We experimented with three clustering algorithms—*k-means*, *HDBSCAN*, and *Hierarchical Agglomerative Clustering* (HAC)—on sample data from three museums. HAC consistently produced the most coherent and stable groups, so it was selected for the full-scale analysis. For comparability, we fixed 20 clusters per museum and assigned each subject a cluster_id in {0,…,19}.

Cluster labeling: Cluster descriptors were generated with ChatGPT V4, producing concise, human-readable names.

To compare across institutions, we applied the same workflow at an aggregate level: cluster labels from individual museums were pooled, re-embedded, and re-clustered into 20 cross-museum clusters. These were harmonised through LLM-generated names and subsequently spot-checked. Validation followed a two-part regime modelled on the correctness protocol (Ullmann et al., 2022). First, we conducted a targeted review of terms flagged as potential misfits by the LLM, with manual reassignment where appropriate. Second, we performed random audits within each museum cluster to identify systematic errors.

Results

The results demonstrate both the diversity of entities and their inter-relationships and the heterogeneity of the minority culture representations across museums. Figure 1 illustrates the number of extracted locations, time periods, and persons, as connected to other entities. While subjects and locations are comparatively well represented, persons and organisations appear less frequently. Nevertheless, periods and persons can still be meaningfully analyzed in aggregated form and across museums (Figure 1).

To enable systematic comparison, we summarise each museum's distribution of subjects across clusters using a 100% stacked bar chart, which facilitates direct cross-institutional analysis of thematic emphases (Figure 2). In total, more than 178,000 subjects were extracted from the museum datasets. Notably, only six of the fourteen museums included a dedicated metadata field for subjects, underscoring the variability in cataloguing practices.

As shown in Figure 2, the distribution of subjects varies substantially across museums, reflecting the plurality of viewpoints on Jewish culture represented in each institution. Museums in the Netherlands and Israel display a strong emphasis on religious and identity-related themes. In contrast, the Danish museum presents the lowest representation of religious and traditional aspects. Furthermore, the Danish, Estonian, and Australian museums exhibit a heterogeneous distribution of subjects, covering a wide range of clusters (10–12 out of 20). These subjects predominantly highlight everyday aspects of community life, including medicine, economic documentation, food, arts and media, family, youth and education, clothing, migration, and even sports.

A particularly unexpected finding concerns the representation of the Holocaust. Contrary to expectations, only the Australian museum places substantial emphasis on this subject. Other museums, including those located in Europe and Israel, devote comparatively less attention to Holocaust-related materials. This discrepancy highlights not only institutional differences in curatorial priorities, but also broader cultural and geographical variations in how Jewish heritage is framed and presented to the public.

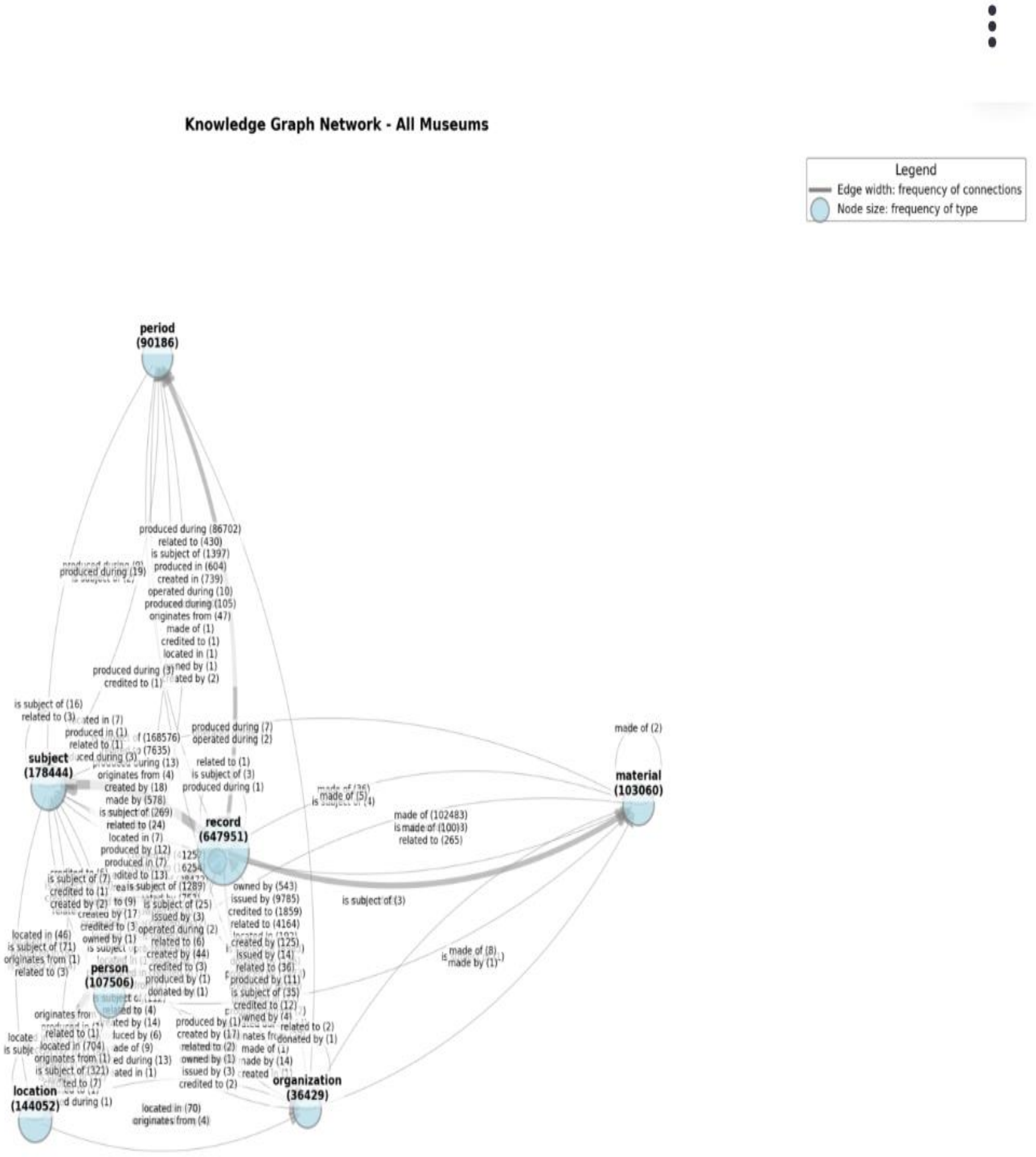

Figure 1. Jewish museum cultural heritage knowledge graph. The graph presents connected data showing relations between records and entities for over half a million records of objects disseminated through the websites of Jewish museums.

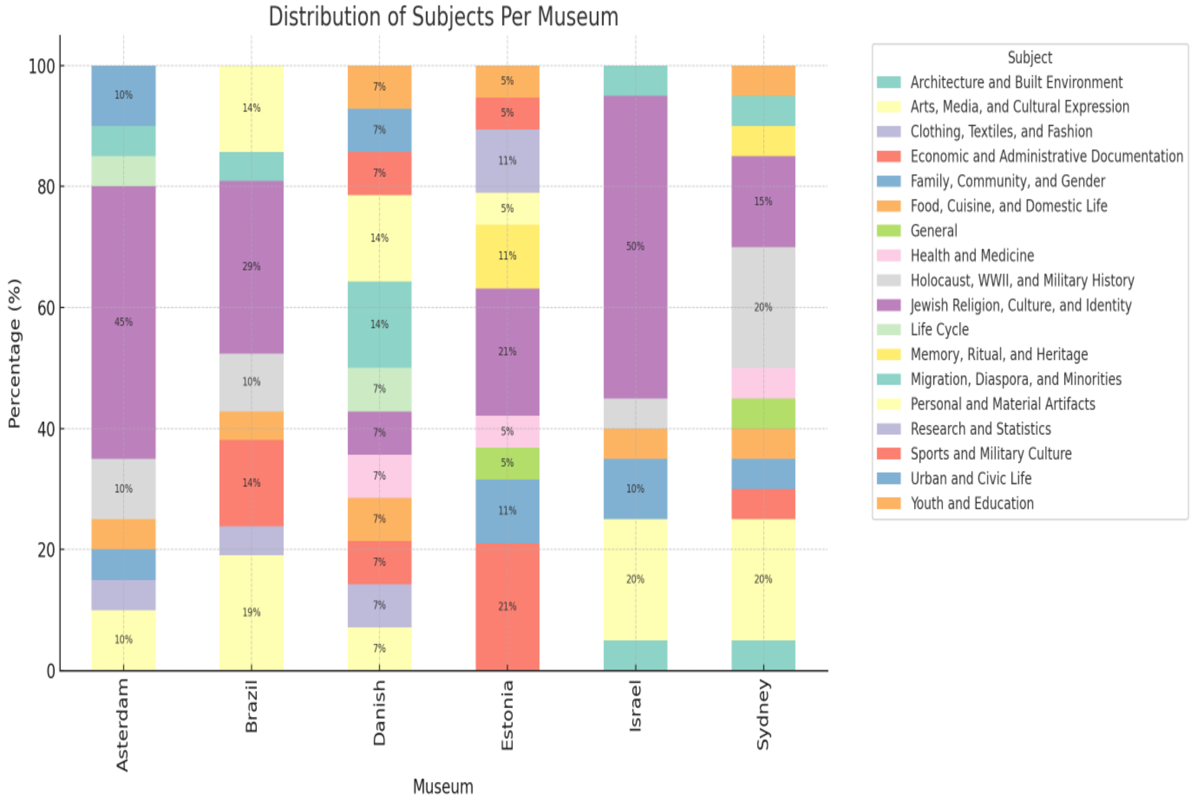

Figure 2. Subject distribution in each museum.

Conclusions

This paper presents a methodology and preliminary findings for constructing a multi-viewpoint ontology to represent a minority culture as reflected in dedicated online museum collections. By integrating advanced large language model (LLM) capabilities with Python-based graph libraries and clustering algorithms, the methodology transforms heterogeneous museum metadata into a structured, multi-viewpoint knowledge graph. This framework enables systematic cross-institutional comparisons of geographic and thematic representations while also providing a scalable foundation for analyzing how ethnic minority cultures are depicted across global museum collections.

Our results on Jewish cultural heritage reveal that the perspectives represented are predominantly those of the Global North—with the exceptions of Brazil and South Africa—since online collections are currently available for only fourteen countries out of nearly 200 recognised by the United Nations. Despite this limitation, the methodology demonstrates the ability to capture multiple perspectives, including those highlighting Jewish culture as expressed by contemporary artists, migration histories, and Holocaust remembrances. These findings align with earlier qualitative evaluations of Jewish museums (Kochavi, 2022), while extending them by offering quantifiable and measurable assessments of diverse viewpoints across more than half a million records.

The central contribution of this study is the development of a knowledge organisation system capable of tracing major trends and identifying patterns in the polyvocality of perspectives that cannot be adequately captured through close reading alone. We plan to extend the analysis to additional entity types, including periods, persons, and organisations. Incorporating these dimensions will allow for a more nuanced account of representational multiplicity, which may subsequently be contextualised through the study of acquisition histories, collection histories, and documentation practices (Turner, 2020; Stevenson, 2022). Looking forward, the methodology presented here can be adapted for the multi-viewpoint knowledge representation of other ethnic minorities, offering a generalisable framework for cultural heritage research.

Acknowledgements

This research was supported by the Israel Science Foundation grant no. 249-2023-1202040.

About the authors

Sara Minster is a PhD student in the Department of Information Science and Applied Artificial Intelligence, Bar-Ilan University, Israel. Her research interests include knowledge graphs, applied data science and AI for cultural heritage. She can be contacted at sara.minster@biu.ac.il

Maayan Zhitomirsky-Geffet is a Professor in the Department of Information Science and Applied Artificial Intelligence, Bar-Ilan University, Israel. She received her PhD in computer science from the Hebrew University of Jerusalem, Israel. Her research interests include automatic text analysis, knowledge graphs, computational social sciences and humanities. She can be contacted at maayan.zhitomirsky-geffet@biu.ac.il

Inna Kizhner is a postdoctoral researcher in the Department of Information Science and Applied Artificial Intelligence, Bar-Ilan University, Israel. She received her PhD from Siberian Federal University. Her research interests include digital humanities, cultural analytics and museum digitisation practices. She can be contacted at inna.kizhner@biu.ac.il

References

Baclawski, K., Bennett, M., Berg-Cross, G., Casanave, C., Fritzsche, D., Luciano, J., Schneider, T., Sharma, R., Singer, J., Sowa, J., Sriram, R.D., Westererinen, A., and Whitten, D. (2018). 'Ontology summit 2018 communique – contexts in context', Applied Ontology 13(3), pp. 181-200.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning research, 3(Jan), 993-1022.

Dijkshoorn, C., Aroyo, L., Van Ossenbruggen, J., & Schreiber, G. (2018). Modeling cultural heritage data for online publication. Applied Ontology, 13(4), 255-271.

Doerr, M. (2003). 'The CIDOC Conceptual Reference Module - an ontological approach to semantic interoperability of metadata', AI Magazine 24(3), pp. 75-92.

Doerr, M., Kritsotaki, A., and Boutsika, A. (2011). 'Factual argumentation - a core model for assertions making', Journal on Computing and Cultural Heritage (JOCCH) 3(3), p. 34.

Fan, Y., Shi, L., & Yuan, L. (2023). Topic modeling methods for short texts: A survey. Journal of Intelligent & Fuzzy Systems, 45(2), 1971-1990.

Kizhner, I., Terras, M., Afanaseva, J., Pusenkova, D., Sherer, M., and Skorinkin, D. (2022). 'The culture of very rich and very poor: Do digital museum collections tell us anything about Jewish culture', in: Zaagsma et al. (eds), Jewish Studies in the Digital Age, Berlin: de Gruyter.

Kochavi, S. (2022). Jewish museums in the United States: Foundations and changes in the twentieth century. Ars Judaica: The Bar Ilan Journal of Jewish Art, 18(1), 145-156.

Likhitha, S., Harish, B. S., & Kumar, H. K. (2019). A detailed survey on topic modeling for document and short text data. International Journal of Computer Applications, 178(39), 1-9.

Shoilee, S. B. A., de Boer, V., & van Ossenbruggen, J. (2023). Polyvocal knowledge modelling for ethnographic heritage object provenance. In Knowledge graphs: Semantics, machine learning, and languages (pp. 127-143). IOS Press.

Stevenson, A. (2022). Egyptian archaeology and the twenty-first century museum. Cambridge University Press.

Turner, H. (2020). Cataloguing culture: Legacies of colonialism in museum documentation. UBC Press.

Ullmann, T., Hennig, C., & Boulesteix, A. L. (2022). Validation of cluster analysis results on validation data: A systematic framework. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12(3), e1444.

Vlachidis, Andreas, et al. 'CrossCult D2. 5 Upper-level Cultural Heritage Ontology.' (2018).

Zhitomirsky-Geffet, M. (2019). 'Towards a diversified knowledge organisation system – An open network of inter-linked subsystems with multiple validity scopes', Journal of Documentation 75(5), pp. 1124-1138.

Zhitomirsky-Geffet, M., Kizhner, I., and Minster, S. (2023). 'What do they make us see: a comparative study of cultural bias in online databases of two large museums', Journal of Documentation (preprint). https://doi.org/10.1108/JD-02-2022-0047.