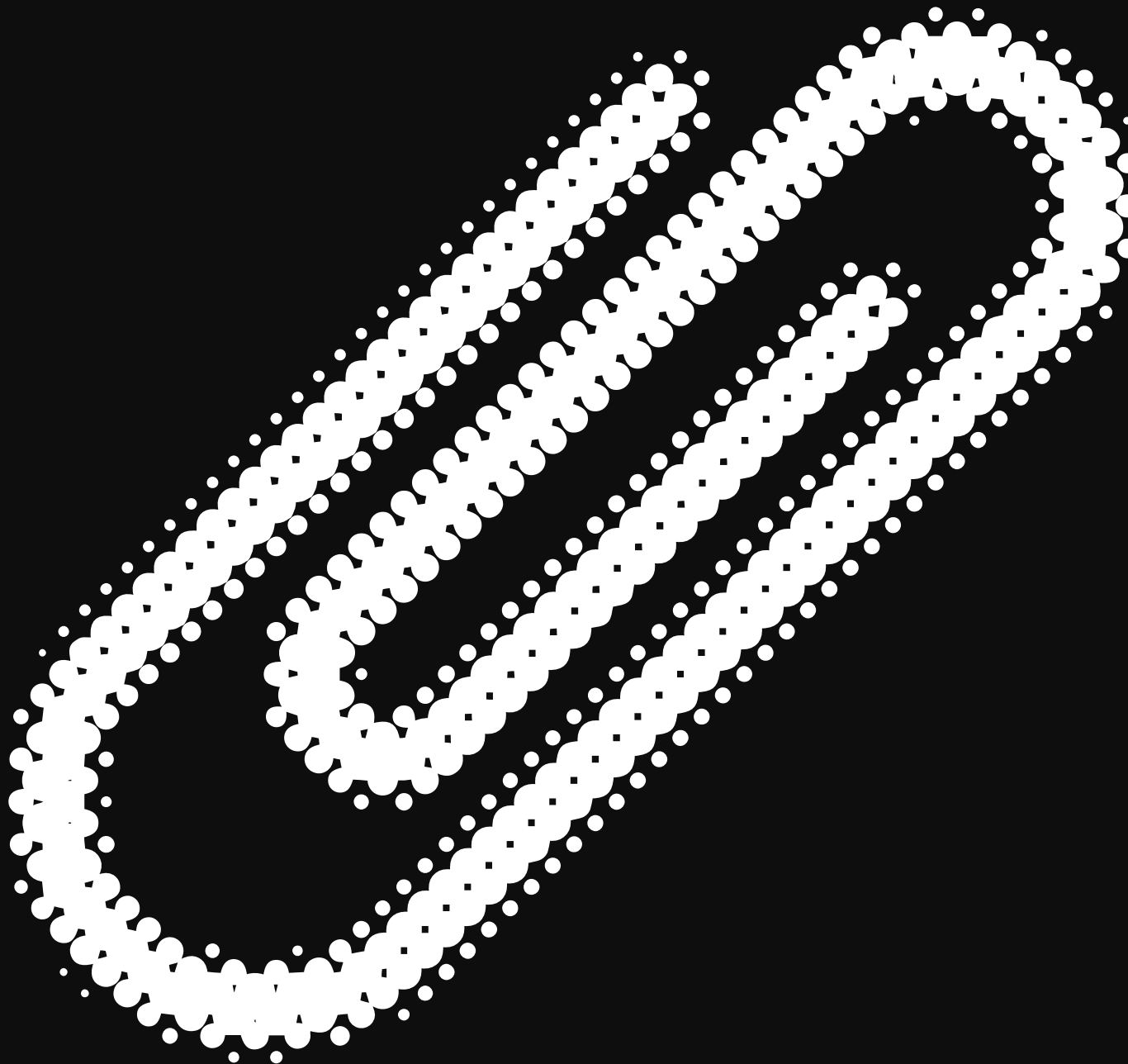


Metadata
Enrichment of
Digitally Converted
Slovak Ethnographic
Research Collection
Documents Using
Generative Artificial
Intelligence

Andrej
Gogora

Institute
of Ethnology
and Social
Anthropology
Slovak Academy
of Sciences,
Klemensova 19,
813 64,
Bratislava,
Slovakia



Abstract



Keywords

metadata enrichment,
generative artificial
intelligence,
ChatGPT,
ethnographic research
collection,
descriptive metadata,
digitally converted collection

This chapter focuses on the use of generative artificial intelligence (GenAI) in the documentary process of metadata restoration. The primary aim is to test the semi-automated methods of metadata extraction and metadata enrichment applied to a machine-readable text sample of digitally converted documents from Slovak ethnographic research collections using a turnkey GenAI model (ChatGPT). It also compares the outputs with traditional methods of creating metadata entries, evaluating the efficacy and accuracy of using GenAI for ethnographic research collection documentary procedures. The chapter explores the benefits and challenges of employing GenAI in this context, offering introductory insights into its potential for improved discoverability, better resource management, and enhanced user experience with research collections.

“Continuous text is not suitable for automated processing.”
Emília Hrubcová, 1973

“Continuous text can be more challenging for computational automated processing, but it is not unsuitable, especially with the use of advanced techniques and models.”
ChatGPT-4o, 2024

Compared to other texts in this collected volume, this chapter is characterized by its thematic specificity, as it does not deal directly with ethnographic or anthropological topics. Nonetheless, it is functionally linked to these fields. Disciplinarily, it occupies a space at the intersection of digital curation, digital archiving, and documentation practices in ethnography. Despite its formalized and technological approach, the focus on ethnographic research collection and its further research use plays a key role here. This approach legitimizes the place of this chapter in the collected volume, as well as its importance for ethnography itself.

In recent years, the public has increasingly regarded GenAI as an ambitious and productive technology in both everyday applications and research contexts (Leontidis & Royal Society Invited Scientific Contributors, 2024). This also applies to the practice of GLAM (galleries, libraries, archives, museums) institutions dealing with cultural heritage (Thiel & Bernhardt, 2023). Therefore, it is necessary to test and evaluate the use of GenAI as a methodological tool in processes of archiving and curating digital research collections. In this case, it will be specifically applied to the digitized Slovak textual ethnographic research collection, where the original metadata records are often incomplete or unavailable. Completing the metadata manually is a time-consuming and demanding task. Semi-automated metadata enrichment with GenAI has significant potential to simplify this process.

From this context, the following research questions arise:

- How can GenAI contribute to the digital conversion of textual documents in ethnographic research collections, especially in the process of metadata enrichment?
- To what extent is the use of GenAI effective compared to the traditional documentation and metadata processing of ethnographic research collections?
- Are the outputs of GenAI metadata enrichment in ethnographic research collections trustworthy and comprehensive?
- What are the main advantages and disadvantages of employing GenAI for metadata processing?

The primary aim is to test semi-automated methods for metadata extraction and enrichment applied to a machine-readable text sample of digitally converted Slovak ethnographic research collection documents using a turnkey GenAI model (ChatGPT) and to evaluate the efficacy and accuracy of these methods in the context of ethnographic research collection documentation procedures. It is important to point out that the function of GenAI is not to replace the traditional process of assigning metadata in close collaboration with the researcher submitting the resources. Instead, it aims to streamline various manual and time-consuming tasks. This endeavor represents an introductory evaluation of GenAI's application in metadata enrichment for ethnographic research collections, utilizing a commonly available, non-customizable GenAI model on a limited sample of text documents. The output will be a procedural and methodical evaluation of the quality of semi-automated metadata generation within the established metadata scheme and a comparison with original metadata records. The purpose of this chapter is to explore the benefits and challenges of using GenAI in this context, offering curatorial insights into its potential for improved discoverability, better resource management, and enhanced user experience with research collections. To clarify, the ambition of this chapter is not to directly contribute to the production of ethnological or anthropological knowledge. Instead, within the framework of digital curation as an auxiliary discipline, it seeks to analyze the preparation of a more precise and granular metadata that can support archival or other types of research within these scientific fields (for the reciprocal relationship between research collection and ethnographic research, see: Gogora, 2023). In this case, GenAI is applied to generate answers to curatorial and documentation issues, which are methodically linked to the process of generating answers to ethnographic or anthropological questions.

This part is dedicated to the current state of research on the given topic and provides an overview of AI and machine learning technology, followed by a brief summary of their use in digital archiving and curation. GenAI, as a subset of artificial intelligence (AI) systems, refers to the capability to generate new content or solutions based on the data it has been trained on. Unlike traditional AI, which operates based on predefined rules or simply processes data, GenAI is designed to create novel outputs, such as text, images, or music, that were not explicitly programmed into it. This is achieved through advanced machine learning techniques, particularly deep learning, which allow models to learn complex patterns, structures, and relationships from extensive datasets. These capabilities are widely used in applications, such as natural language processing (NLP), automated content creation, creative arts, and design generation.

GenAI has begun to play a significant role in the digital curation of libraries, archives, and collections, including research ones, by automating various labor-intensive tasks. Initially, GenAI was practically implemented in the processes of digital preservation and conversion of archives, especially cultural heritage artefacts, for some time (Ferro et al., 2023; Das et al., 2022; Teel, 2024). The most common processes in which these technologies are applied include optical character recognition (OCR) and handwritten character recognition (HCR), image enhancement techniques to improve the quality of scanned images, segmentation of digitized documents into individual sections or components using automatic computer vision methods, and automated quality control checks during the digitization process. In general, by leveraging GenAI in mass digital conversion, institutions accelerate the digitization process, improve the quality of postprocessing, and streamline workflow management.

In addition to being commonly used in digital preservation, GenAI has recently been implemented in the broader digital curation workflow (Neudecker, 2023). There are theoretical reflections on changes in archival and digital curation as such (Cushing & Osti, 2023) and on the future challenges of these disciplines (Colavizza et al., 2021). In the curatorial process itself, GenAI is primarily used for information management and cataloging (e.g., automated data classification, tagging, sorting, and clustering based on similarities; automatic markup;

metadata enhancement; content summarization; sentiment analysis; duplicity or anomaly detection and removal; automated statistical data extraction). For the purpose of collection users, GenAI has begun to be used to offer interactive and personalized environments to improve services and enhance visitors' experiences. Moreover, recommendation systems are applied to suggest relevant items to further engage the research community and public. GenAI is able to automatically collect statistical data on visitors, analyze user behavior, and evaluate the collection's impact (Caramiaux, 2023).

In digital curation, metadata enrichment refers to the process of enhancing incomplete metadata to improve the discoverability, usability, and overall value of digital assets. This process involves adding more detailed, accurate, and contextually relevant information to the metadata associated with digital objects. The use of GenAI can significantly enhance the metadata of digitally converted research collections, mainly through metadata extraction (e.g., the automatic generation of key metadata elements by content analysis of the full text, and the automatic identification and categorization of named entities such as people, places, and institutions within the text). Another form of using GenAI is semantic enrichment, for example, in the form of ontology linking (mapping metadata to existing ontologies or linking to standardized vocabularies that improve semantic relationships and interoperability) (Randby & Marciano, 2020) and concept tagging (GenAI, through deep learning mechanisms, tags documents with relevant concepts, even if they are not explicitly mentioned). In addition to these essential applications, GenAI can automatically provide metadata quality assurance, error correction, and anomaly detection for subsequent human review.

This section focuses on the introduction of the *Collection of Research Reports* (CRR), previous historical attempts to computationally process its items, the current state of its digital conversion, the respective metadata schema, as well as the identification of metadata issues along with potential solutions using GenAI.

The CRR has been systematically built by experts in ethnography at the *Institute of Ethnology and Social Anthropology Slovak Academy of Sciences* (IESA SAS) since 1946. It contains over 1,520 documents recorded in field and archival research by the staff of the IESA SAS. The CRR comprehensively covers cultural heritage themes throughout Slovakia, such as employment, production, food, clothing, types of housing, social practices, customs, and folk arts (Prandová, 1982). The collection includes passport letters, archival cards, questionnaires, research reports, transcripts, media excerpts, photo documentation, and sheet music. The diversity of document types is considerable; they are handwritten, typewritten, and produced by various types of printing methods such as needlepoint, laser, inkjet printing, and newspaper clippings. There are also technical drawings and sketches. The CRR is an important collection from the perspective of documenting traditional folk culture and the history of ethnography in Slovakia.

The CRR is currently in an advanced state of analog-to-digital conversion (Kubisa & Gogora, 2018), but the first reflections on the computational processing of this and other collections of IESA SAS date back to the 1970s. Initially, it involved familiarizing with the issue and the first theoretical attempts to apply statistical processes, already employed in the exact sciences using mainframe computers, to ethnography. The main aim was to simplify and streamline manual documentation work while integrating a massive amount of data into a universal documentation system (Hrubcová, 1973; Prandová, 1978). Productive use of computer technology began at IESA SAS in the latter half of the 1980s, particularly in the recording and processing of documentary material on traditional culture (Zajonc, 2006). The characteristic feature of these efforts was the experimental creation of the first domestic ethnographic metadata forms and the search for shared terminological standards.

In the early 1990s, specific emphasis was placed on the relational database approach and a strong push toward creating unified database systems for different

types of collection resources (Zajonc & Uhlíková, 2003), alongside the first complex analysis of the connection between computing technology and the research process in ethnology in Slovakia (Zajonc, 1996). However, most of these efforts remained theoretical or incompletely implemented due to insufficient technological infrastructure and a lack of knowledge about advanced computational processes, and dependence on other technological workplaces and experts.

The first personal computers appeared in the workplace at the end of the 1990s, and at the beginning of the new millennium, funds were raised for the first scanning devices. This led to tangible outputs in the form of large-scale digitization and post-processing, primarily of visual resources, as well as to the systematic conceptualization of technological procedures (Ratica, 2003), and the development of more advanced electronic catalogs (Popelková & Zajonc, 2008). It can be said that the use of technology in documentation practice at the IESA SAS has been historically approached in the right way and with a strong focus on ethnography itself. However, a more formalized technological aspect, as well as knowledge of international standards and the field of digital curation, was lacking.

The complex digitization and computational metadata processing of the IESA SAS textual research reports only began in 2017. So far, the entire collection has been chemically cleaned and archivally repackaged. A development policy document and an internationally standardized metadata schema have been prepared. The manual transcription of the incremental book (39,350 entries) and the electronic subject catalog (40,023 entries) have been completed. All 1,520 documents have been digitized, containing approximately 250,000 unit pages, and 710 documents have been post-processed, though without OCR. Continuous post-processing, metadata completion, and conceptual preparation for the creation of a digital repository system are currently underway. This forms the basis for testing GenAI through the sample of textual field research reports.

In this section, a uniform, standardized metadata schema intended for describing the CRR textual documents will be presented. The schema was pre-established in the *Development Policy* document for CRR before the launch of digitization and metadata processing. It is based on the international standard DublinCore Metadata Terms. The primary reason listed here is that its elements will later be used in testing GenAI for CRR metadata enrichment. The metadata schema is composed of descriptive, administrative, and technical parts:

Descriptive metadata

Category	DublinCore Term
Identifier	http://purl.org/dc/terms/identifier
Title	http://purl.org/dc/terms/title
Subtitle	http://purl.org/dc/terms/alternative
Author	http://purl.org/dc/terms/creator
Researcher	http://purl.org/dc/terms/creator (qualified DC element)
Originator	http://purl.org/dc/terms/creator (qualified DC element)
Keywords	http://purl.org/dc/terms/subject
Abstract	http://purl.org/dc/terms/abstract
Municipality	http://purl.org/dc/terms/Location (qualified DC element)
Part of the municipality	http://purl.org/dc/terms/Location (qualified DC element)
District	http://purl.org/dc/terms/Location (qualified DC element)
Region	http://purl.org/dc/terms/Location (qualified DC element)
Country	http://purl.org/dc/terms/Location (qualified DC element)
Locality	http://purl.org/dc/terms/Location (qualified DC element)
Date of realization	http://purl.org/dc/elements/1.1/date
Project	http://purl.org/dc/terms/isPartOf

Administrative and technical metadata

Category	DublinCore Term
Type	http://purl.org/dc/elements/1.1/type
Size	http://purl.org/dc/terms/SizeOrDuration
Language	http://purl.org/dc/terms/language
Method of accrual	http://purl.org/dc/terms/accrualMethod
Date of submission	http://purl.org/dc/terms/dateSubmitted
Format	http://purl.org/dc/elements/1.1/format
Institution	http://purl.org/dc/elements/1.1/publisher
Access rights	http://purl.org/dc/terms/accessRights
License	http://purl.org/dc/terms/license
Bibliographic citation	http://purl.org/dc/terms/bibliographicCitation
Note	http://purl.org/dc/terms/description

Table 1 CRR standardized metadata schema

In the current phase of digital conversion, after the completion of digitization and transcription of original metadata, it is appropriate to apply the process of metadata enrichment so that CRR documents are fully prepared for access and retrieval. From the schema, a group of descriptive metadata has been identified as problematic and potentially in need of additional semi-automated completion. Based on the archival analysis of CRR documents, the following deficiencies have been identified:

- Metadata are incomplete and unspecified (mainly ‘Title’, ‘Keywords’, ‘Location’).
- Metadata are not available at all in the original catalogs (‘Subtitle’, ‘Abstract’).
- Metadata are arbitrarily assigned by the documentarian, not respecting the source document (‘Keywords’, ‘Title’, ‘Note’).
- Metadata do not correctly refer to the context of the document (e.g., position of a keyword or locality in the document).
- Metadata transcription requires a final check.

Therefore, descriptive items, including ‘Title’, ‘Subtitle’, ‘Keywords’, ‘Abstract’, and all location metadata, will be individually tested using GenAI. Processing these deficiencies manually is time-consuming and resource-intensive, especially given the large volume of documents. Moreover, the researchers who authored the CRR research reports are mostly deceased or unable to edit the metadata records. For this reason, it is assumed that the use of GenAI in CRR metadata clarification, completion, and correction will simplify and make the resolution of this issue more efficient, at least partially.

CRR metadata enrichment

For testing metadata enrichment using GenAI, a sample of six CRR documents was selected totaling 271 unit pages and 514,686 characters (excluding spaces). This sample size is sufficient, as the aim is not to assess GenAI's performance in automated processing of massive amounts of data, but rather to conduct a micro-analysis of selected metadata items. The selection of the number of units is also based on the nature of the ethnographic research collection, which requires more detailed curation and greater attention to the uniqueness of the selected entries, as most documents are neither formally nor content-homogeneous.

The sample documents were not selected randomly; they were chosen based on the following criteria, with justification provided for each:

- A4 paper format (other formats, such as archive cards or notebooks with a specific page layout, pose challenges for document layout analysis and subsequent OCR conversion for now),
- typewritten or printed text (in the current phase of CRR processing, OCR conversion has not yet been applied; this and the previous criteria were selected to reduce time and capacity requirements for implementing this process),
- Slovak language (foreign language and dialect entries were filtered out to simplify input analysis),
- continuous text ranging 10 to 50 pages (to ensure a representative amount of text for semi-automatic NLP processing),
- problematic original metadata entries (to allow for a comparison between the deficient original metadata and the metadata generated by GenAI).

None of these criteria are defined by a specific ethnographic subject, theme, or other descriptive metadata unit, as this is a formal curatorial evaluation of the potential application of GenAI for metadata enrichment, considering the current stage of CRR processing. Based on these criteria, the following CRR documents were selected:

Identifier	Title	Author	Summary
AT1408	Urban winegrowers in the 1 st half of the 20 th century from the archive	Popelková, Katarína	Total words: 29,892 Vocabulary Density: 0.313 Readability Index: 11.206 Average Words Per Sentence: 36.7
AT1414	Transformations of the local community in the 20 th century	Beňušková, Zuzana	Total words: 10,031 Vocabulary Density: 0.335 Readability Index: 6.849 Average Words Per Sentence: 7.3
AT1422	Socio-economic-cultural changes in the Slovak countryside	Benža, Mojmir	Total words: 12,326 Vocabulary Density: 0.330 Readability Index: 10.780 Average Words Per Sentence: 12.1
AT1426	Non-agricultural employment in the villages of the Nitra district in the first half of the 20 th century	Faltánová, Ľubica	Total words: 28,835 Vocabulary Density: 0.238 Readability Index: 8.210 Average Words Per Sentence: 11.3
AT1430	Transformation of the countryside	Danglová, Oľga	Total words: 3,840 Vocabulary Density: 0.486 Readability Index: 12.664 Average Words Per Sentence: 14.5
AT1462	Personalities of Slovak ethnology: Interview with Š. Mruškovic	Slavkovský, Peter	Total words: 8,452 Vocabulary Density: 0.400 Readability Index: 14.867 Average Words Per Sentence: 15.7

Table 2 Sample of CRR documents

Scanned and post-processed document images were converted to plain text format using the Tesseract Open Source OCR Engine software, version 5.4.1., with AI integration via an LSTM neural network and support for Slovak. Subsequently, language and formal errors were corrected using ChatGPT and finally reviewed manually. Even at this stage, the use of GenAI in the form of text normalization (and later in language translation from Slovak to English during real-time querying) proved effective, leading to more accurate outputs.

In this main part, the application of GenAI to selected items of the CRR metadata scheme is presented. Among the available options (ChatGPT, Claude, Gemini, Copilot, Perplexity, etc.), the chatbot ChatGPT, developed by OpenAI, was chosen in its GPT-4o model (using a Plus account). The reasons for selecting ChatGPT are as follows:

- GPT-4o has an architecture that is among the most advanced in terms of Natural Language Processing (NLP), enabling it to process complex texts and generate highly accurate data, including nuanced topics, themes, and keywords.
- It is trained on a diverse and extensive dataset from various fields and contexts, making it capable of handling ethnographic field research documents and generating relevant metadata across different domains.
- It supports multiple languages, including Slovak.
- It can be flexibly customized with specific prompts to align with particular requirements (this means adapting to various metadata schemes and enrichment strategies).
- It is widely used and integrated into various platforms, making it more accessible to integrate into existing workflows for metadata enrichment.

In the future, it would be useful to compare the results of several chatbots. For now, GPT-4o's advanced capabilities, user-friendliness, efficiency, and low entry costs make it a strong choice for initial metadata enrichment in research collections (Dwivedi et al., 2023). The testing procedure involved uploading the plain text of the document to the ChatGPT platform and querying the deficient metadata items in real time. To maintain consistency, the wording of the queries was strictly based on the phrasing used in the official CRR metadata filling manual. The outputs of the metadata enrichment for selected descriptive items are as follows (the interpretations and implications of the results are provided in the *Discussion* section):

Title and subtitle

Query: “Based on the attached text, create a title and subtitle for the document, according to the subject, keywords, object, captured period or the location of the research, so that it describes the document in terms of content and theme (avoid using overly long or vague formulations).”

Identifier	Title / subtitle (documentarist)	Title / subtitle (author)	Title / subtitle (ChatGPT)
AT1408	Urban winegrowers in the 1 st half of the 20 th century from the archive / –	Urban winegrowers as a group within the organism of an interwar city / Modra, 1 st half of the 20 th century – material from archival research	Winegrowers of the city of Modra in the first half of the 20 th century / Social and economic aspects of viticulture in Modra based on archival sources
AT1414	Transformations of the local community in the 20 th century / –	Transformations of the local community in the 20 th century, BANKA – BÁNK, Hungary – Novohrad / Field research material	Transformations of the local community in Banka during the 20 th century / Socio-cultural changes and influences on the Slovak minority in Hungary
AT1422	Socio-economic-cultural changes in the Slovak countryside / –	Socio-economic-cultural changes in the Slovak countryside – Velušovce / –	Socio-economic and cultural development of the village Velušovce in the 20 th century / Transformation of the countryside and collectivization in Central Ponitrie
AT1426	Non-agricultural employment in the villages of the Nitra district in the first half of the 20 th century / –	<i>Id. / –</i>	Non-agricultural employment in the villages of the Nitra district / Socio-economic aspects and living conditions of the inhabitants in the first half of the 20 th century

AT1430	Transformation of the countryside / –	<i>Id. / –</i>	Socio-economic changes and local self-government in the municipalities of the Pezinok district after 1989 / Case studies from Dolany and Štefanová
AT1462	Personalities of Slovak ethnology: Interview with doc. PhDr. Š. Mruškovic / –	<i>Id. / Interview with Doc. PhDr. Štefan Mruškovič, CSc. About ethnology and museology in Slovakia in the second half of the 20th century</i>	Conversations on ethnology and museology: Štefan Mruškovič's reflections on the development of the disciplines in the second half of the 20 th century / Personal testimonies on the formation of ethnology and museology in Slovakia

Table 9 Metadata enrichment of 'Title' and 'Subtitle'

Due to discrepancies in the records, the table shows metadata recorded by the documentarian in the physical register as well as the records in the original source material made by the authors. These two records are illogically not always identical (*Id.*), and the subtitles are completely absent from the physical registers. As a result, they are all compared with ChatGPT outputs. It is evident that in at least three cases (AT1414, AT1422, AT1430), the titles created by ChatGPT are more precise and comprehensive in key data such as subject, locality, and time period. Given the absence of subtitles, their generation by GenAI represents a significant contribution in these cases. Furthermore, there is an obvious formal and content continuity between the title and the subtitle.

Keywords

Query: "From the attached text, generate a set of thematic keywords by subject indexing that semantically characterize the overall content of the document (4–8 keywords)."

Identifier	Keywords (documentarist)	Keywords (author)	Keywords (ChatGPT)
AT1408	viticulture, urban winegrowers (population employment, production, transport and trade)	-	viticulture, winemaking in Modra, interwar period, self-government and economic issues, social and cultural customs, political and religious organizations, social and economic changes, chronicles and archival records, Jewish community, anti-Semitism
AT1414	-	mine, farmhand, wealth, Budapest, gypsies, church, children, house, cooperative, folklore festival, Count, farming, hygiene, cottagers, lake, priest, horses, pub, christening, kulaks, culture, bathing, religion, unemployed, harvests, trade, money, entrepreneurship, work, vacationers, repatriation, fish, Slovaks, Slovak, associations, wedding, holidays, school, dance, market, vineyards, death, mutual aid, education, land, mixed marriages, railway worker, Jews	ethnography of Slovak communities, historical and social changes, cultural heritage of Slovaks in Hungary, repatriation and migration, religious and social customs, life in the countryside, mixed marriages, recreational areas and social ties

AT1422	social relations, socio-economic and cultural changes (social systems, groups, social and family relations and customs)	-	socio-economic changes in the countryside, demographic and migration trends, cultural and social traditions, environment and infrastructure, education and employment, agriculture and local economy
AT1426	employment of the population, non-agricultural jobs (population employment, production, transport and trade)	non-agricultural jobs, agricultural workers, forestry workers, maids, industrial workers, large estate	non-agricultural jobs, rural villages, agricultural labor, socio-economic conditions, history of the Nitra district, living conditions in the 20 th century, migration and employment, cultural traditions and customs
AT1430	social relations, transformation of the countryside (social systems, groups, social and family relations and customs)	-	transformation of the countryside, demographic changes, agriculture and viticulture, regional development and cooperation, social policy, cultural and sports activities, business in the countryside, municipal government and infrastructure
AT1462	auxiliary scientific disciplines, personalities of Slovak ethnology, doc. PhDr. Štefan Mruškovič, CSc. (history of ethnography and folk culture, magazine covers, press production)	history of science, ethnology, museology	ethnology in Slovakia, museology and museum studies, personalities of Slovak science, history of folk culture research, development of Slovak museology, Štefan Mruškovič, cartography of folk culture, institutional development of SAS

Table 4 Metadata enrichment of 'Keywords'

In the case of keywords, there can also be several inconsistencies and omissions in the original records. It is evident that ChatGPT demonstrated the capacity to expand and specify the limited list of keywords (AT1408, AT1430, AT1462) and meaningfully reduce the redundant number of keywords while emphasizing essential topics (AT1414). In addition, in AT1408, the chatbot was able to capture a theme—Jewish community, anti-Semitism—that the documentarian did not notice but which is emphasized several times in the document in the context of the interwar period. It follows that the ChatGPT application is suitable here for second-level specification, expansion and addition of keywords, as well as for the statistical capture of apparently minor topics.

Abstract

Query: “From the attached text, generate a brief and descriptive summary (abstract) of the text source (3–6 sentences).”

Identifier	Abstract (ChatGPT)
AT1408	The document focuses on economic and social factors affecting viticulture and winemakers in the town of Modra during the first half of the 20 th century. The research includes an analysis of archival documents that reveal the impact of these factors on the self-government, culture, and social structure of the community. Special attention is given to the differentiating factors within the urban community and their impact on the life of winegrowers as a social group. The document also captures the wider socio-political and economic impacts on the local community in Modra.
AT1414	The document focuses on the transformation and social changes in the village of Banka in Hungary during the 20 th century, with an emphasis on the Slovak community. It describes demographic trends, migration, mixed marriages, and changes in ethnic identity. Additionally, it addresses economic and cultural transformations, including the impact of recreational activities and changes in the village’s infrastructure. The document also analyzes the role of local governance and relations with vacationers, which contributed to overall social ties and regional development.
AT1422	The document analyzes the socio-economic and cultural changes that have influenced the Slovak countryside, with a focus on the village of Velušovce. It describes the demographic development, employment of the population, cultural traditions and activities, as well as the development of the village’s infrastructure. The document focuses on transformations in agriculture, entrepreneurship, and social life, while also examining the impact of modernization on the living conditions of local residents. It also deals with

the cooperation of local institutions and the population in the development of the municipality and its integration into the wider regional context.

AT1426	The document deals with non-agricultural employment in the villages of the Nitra district in the first half of the 20 th century. The research focuses on the socio-economic conditions of rural inhabitants who engaged in various forms of agricultural work, industrial labor, and other non-agricultural activities. It analyzes the impact of these jobs on living conditions, culture, and everyday life in specific municipalities. The document also maps changes in work habits and living conditions during the transition from traditional means of livelihoods to new economic models.
AT1430	The document deals with the transformation of the countryside in the municipalities of Dolný and Štefanová, analyzing the demographic, socio-economic and cultural changes that affected these municipalities during the 20 th century. It examines the development of infrastructure, local entrepreneurship, agriculture and cultural activities, emphasizing the importance of regional cooperation and the impact of wider social processes on local communities. The document also assesses local governments’ efforts to improve living conditions and preserve cultural heritage, as well as the challenges they face in maintaining traditional crafts and agriculture.
AT1462	The document contains an interview with Štefan Mruškovič, a prominent Slovak ethnologist and museologist, who reflects on the development of ethnology and museology in Slovakia in the second half of the 20 th century. The interview covers his personal and professional experiences, including his work as director of the Slovak National Museum and the Slovak National Gallery. Mruškovič talks about his work on the institutional development of museums, the cartography of folk culture and the creation of museum exhibitions, emphasizing the importance of methodological and political influences on the development of these disciplines in Slovakia.

Table 6 Metadata enrichment of ‘Abstract’

Abstracts are not present in the original CRR documentation, so these metadata were newly generated by ChatGPT. Compared to the content of the documents and the processing of previous metadata, their logical and semantic quality is relatively high. Creating abstracts by someone other than the author is professionally demanding and time-consuming, so this represents a fundamental contribution of GenAI to metadata enrichment.

Locality

Query: “From the attached file, generate a list of all recognized localities (from official municipalities to parts of municipalities), along with their respective District, as well as Region, and Country, as well as any other localities mentioned in the text.”

Identifier	Locality (documentarist)	Locality (author)	Locality (ChatGPT)
AT1408	Municipality: Modra District: Pezinok Region: Bratislava Country: Slovakia	–	Municipalities: Modra, Pezinok, Trnava, Hrubý Šúr, Šenkvice, Bratislava (all with their respective district, region, country) Localities within Modra: Harmónia, Mittelberg, Trlinka
AT1414	Municipality: Bánk Country: Hungary	District: Rétság	Municipalities: Bánk, Rétság, Vacov, Romháň (district and region not specified) Localities within Bánk: Tó utca (street name)
AT1422	Municipality: Velušovce District: Topoľčany Region: Nitra Country: Slovakia	–	Municipalities: Velušovce, Topoľčany, Dubnica nad Váhom, Partizánske, Prievidza, Nováky, Prašice, Závada, Podhradie, Tesáre, Nemečky, Tvrdomestice, Jacovce, Kuzmice, Nemčice (all with their respective district, region, country) Locality: Podhradská dolina, Remote location near Velušovce: Lazarčík

AT1426	District: Nitra Region: Nitra Country: Slovakia Locality: municipalities of the Nitra district	Municipalities: Cabaj-Čápor, Mojmírovce, Veľká Dolina, Dražovce, Veľký Cetín	Municipalities: Cabaj-Čápor, Mojmírovce, Veľká Dolina, Dražovce, Jarok (all with their respective district, region, country) Locality within Cabaj-Čápor: Hrušov, Locality connected to Hrušov: Párovské Háje, Part of Nitra: Čermáň
AT1430	Municipality: Doľany / Štefanová District: Pezinok Region: Bratislava Country: Slovakia	–	Municipalities: Doľany, Limbach, Čierna Voda, Zeleneč, Štefanová, Budmerice, Horné Orešany, Dolné Orešany, Červený Kameň (all with their respective district, region, country)
AT1462	Municipality: Bratislava / Martin District: Bratislava / Martin Region: Bratislava / Žilina Country: Slovakia	Bratislava, Martin – ethnological institutions	Municipalities: Bratislava, Martin, Smolenice, Šaštín, Trnava, Partizánske, Banská Štiavnica, Sedliacka Dubová (all with their respective district, region, country) Locality: Záhorie (a region, not a specific locality), Locality: Molpir (associated with Smolenice, archaeological site)

Table 6 Metadata enrichment of ‘Locality’

ChatGPT offers expanded lists of locations that were not registered by the original documentarian's procedures. Furthermore, it presents these locations in a consistent format—municipality, district, region, country. Using Named Entity Recognition (NER) technology, it also identifies unofficial locations and provides explanations, which is helpful for users researching local sites.

In addition to testing the selected metadata items, supplementary analyses were also conducted, such as the generation of structured document contents, named entity recognition, generation of bibliographic records of documents, normalization of metadata entries according to standardized formats, and a final check (without detailing specific results here). From both a methodological and practical perspective, the automated recognition of persons appears to be the most interesting for further archival and curatorial processing. ChatGPT effectively identified occurrences of names and surnames in the sample, along with brief descriptions. This is useful not only from a semantic and descriptive point of view but also for the required document anonymization. For online publication of documents that do not contain informed consent, the implementation of an anonymization process is crucial. Instead of the labor-intensive manual search for mentioned persons, the use of GenAI with advanced NER technology is very promising for the future.

Discussion

Based on the obtained output data, the interpretation of the results is presented in the following section, organized by individual metadata items.

Regarding the generation of **titles** and **subtitles** using ChatGPT, it was confirmed that these outputs are actually usable and semantically correct. In the outputs, the most important descriptive data are very clearly and concisely incorporated. However, in the context of digitally converted research collections, it is necessary to critically consider their application, particularly in relation to the archival principle of integrity. According to this principle, the institution must preserve collection documents in their original state, without changes, in order to protect their original value and information. Editing the title, in this context, constitutes a fundamental challenge to integrity as well as ethical standards in digital curation. This raises questions about the usability of GenAI in the creation of titles for already registered documents. However, it is clearly useful when titling new documents or found documents where no title exists. In cases of unclear or contradictory document titles, it is acceptable to clarify or shorten the title using GenAI, provided that the correctness of the new title is verified and does not compromise the document's integrity (ideally with consultation and consent from the author). For subtitles, which are mostly absent in the case of CRR, the application of GenAI is justified. By generating semantically correct units, it can enhance the research usability of the document. Finally, the testing revealed that the practical reason for the imprecise wording of the original titles—as well as other metadata—is mainly due to insufficient guidance for the author of the research report on how to create metadata units.

As previously mentioned, GenAI's ability to add, specify, and expand **keywords** was reliably verified. However, for entirely new documents, it is currently more appropriate to rely on the author's creation of keywords, along with an instructional manual for filling in metadata, and possibly supplemented by the additional use of GenAI. Editing keywords does not pose a problem from the perspective of the principle of integrity, as their creation and processing are expected to involve the active intervention of documentarian. In addition, the existing semantic system of keywords in the given collection (in the case of CRR, the so-called *Classifier*) is part of this process, and it represents the main obstacle to the

arbitrary use of GenAI. To maintain the coherence of this system (whether in a hierarchical, linear, or hybrid form), a correlation between assigned entries and the established system of keywords is assumed. From a practical standpoint, the generation of keywords using GenAI contradicts the need to maintain a controlled dictionary of subject keywords. This could lead to undesirable situations where multiple synonymous or similar terms are used to denote a single phenomenon, or where ambiguous and imprecise terms are generated freely. This issue can be technically solved by linking a controlled wordlist in GenAI as a reference semantic list for generating keywords. However, this solution currently requires a higher level of technical, professional, and conceptual preparation. Therefore, the challenge remains, from a practical standpoint, how to combine the existing keyword system with keyword generation by ChatGPT.

Given its demanding manual preparation, **abstract** generation using ChatGPT has been proven to be expertly consistent and efficient. This is especially true for original converted research collections, where the authors usually did not create document summaries. It streamlines the workflow for the documentarian, and when the documents are made accessible, the generated abstracts serve as a suitable aid for better orientation for the user. The generation of abstracts is clearly the most effective use of GenAI for metadata enrichment in this context. A potential future challenge would be to develop a more specific query that would instruct GenAI to generate abstracts in a structured format, containing the goals, methods, purpose, brief content of the document and findings. However, in the case of generated abstracts from the CRR sample, the issue of GenAI's creativity emerges. Upon examination, it is evident that the generated texts display signs of genericity and lack of specificity, while they often repeat formulations and stylistic structures. Even though it can be argued that they are concise, they evidently lose their uniqueness as well as aspects of the human interpreter. It can be assumed that, with the continued improvement of GenAI models and the amount of training data used, the issue will be mitigated. In this context, however, a critical issue regarding the sustainability of the outputs arises. Due to the rapid development of ChatGPT (as well as the models from other companies), it is necessary to consider regularly updating the results obtained during the widespread deployment of this technology. For more sophisticated tasks, newer-generation models are expected to provide better results compared to those of previous generations. This phenomenon significantly complicates the large-scale and fully automated integration of GenAI into the documentary process of handling research collections.

As previously mentioned, the main advantage of ChatGPT for the enhancement of **location** metadata is its ability to detect and formally classify location records presented in the sample. However, upon rechecking the outputs, it turned out that ChatGPT identified additional locations that had been overlooked. In addition, these extensive lists have no internal ordering. It is evident that some localities are

central to the content while others are marginally mentioned. Therefore, further consideration is needed to prioritize items according to their significance to the document. The query should specify that the found locations be arranged by semantic relevance, so less important location records are automatically filtered out.

Finally, it is important to note that the results can vary even when the same instruction is entered, which means that working with the ChatGPT model requires an increased level of attention and an advanced ability to formulate appropriate requirements in natural language. It is necessary to acquire advanced skills in querying, to systematically test a larger number of instructions, and examine the outputs to ensure that the requirements are correctly set. Additionally, during work with ChatGPT, there were no significant issues with generating nonsensical or inaccurate outputs, commonly referred to as AI hallucinations.

Throughout the processing of the results, the personal-professional aspect emerged multiple times. This is especially true for digitally converted collections, where the current documentarian is entirely dependent on the original records made in the past. The examples clearly highlight several inconsistencies caused by documentarian errors rather than system failures. In addition, it is evident that, in the case of research collections, apart from documentarian control and clear instructions, much depended on the ability and willingness of the author-researcher to correctly fill in the important metadata items. A practical takeaway from this testing is emphasis of the necessity for repeated documentarian oversight of both document content and metadata records.

In line with the defined aim, it can be concluded that GenAI, and specifically the ChatGPT model (GPT-4o), is currently efficient and accurate for metadata enrichment of research collections and primarily for certain types of metadata items and under the condition of a thorough documentarist inspection and revision. This is particularly true for documentary preparation of metadata and for incomplete or problematically filled metadata where manual processing would be inefficient. Practically, it is a highly accessible and cost-effective solution that is simple and intuitive, even for ordinary users, reducing the time and cost associated with manual creation and maintenance. However, to achieve adequate results, it requires systematic documentary preparation, conceptualization of procedures, basic knowledge of GenAI backend processes, experience in formulating precisely targeted requirements, and testing multiple outputs. Based on the testing of a sample of Slovak ethnographic research reports, it is necessary to point out that using GenAI as a turn-key chatbot may not be suitable for fully automated and mass use in enriching metadata across extensive collections. The very nature of research collection documents requires individual processing of source material and metadata units (rather than high-volume ones), as well as detailed micro-analysis, consistent control, and careful curatorial and ethical consideration of each record in consultation with the authors or experts from the given field. For a curation process organized in this way, GenAI has a valid role and will undoubtedly contribute to the restoration of original resources. The next challenge in the near future is to test the usability of a customizable GenAI model and custom training data as well as external data integration, especially in the case of keywords and locations metadata. This approach would ensure more accurate metadata enrichment, greater control over computational processes, as well as additional contextual relationships and information.

This paper was produced with the support of the scientific research project *Revitalization of cultural heritage in public space as a reflection of global influences on rural and urban communities*, VEGA 2/0058/22 (2022–2025).

- Caramiaux, B. (2023). AI with Museums and Cultural Heritage. In S. Thiel & J. C. Bernhardt (Eds.), *AI in museums* (pp. 117–130). transcript Verlag. <https://doi.org/10.1515/9783839467107-011>
- Colavizza, G., Blanke, T., Jeurgens, C., & Noordegraaf, J. (2021). Archives and AI: An Overview of Current Debates and Future Perspectives. *Journal on Computing and Cultural Heritage*, 15(1). <https://doi.org/10.1145/3479010>
- Cushing, A., & Osti, G. (2023). “So how do we balance all of these needs?”: How the concept of AI technology impacts digital archival expertise. *Journal of Documentation*, 79(2), 12–29.
- Das, D., B., Maringanti, H., & Dash, N. (2022). Role of Artificial Intelligence in Preservation of Culture and Heritage. In *Digitalization of Culture Through Technology* (pp. 92–97). Routledge. <https://doi.org/10.4324/9781003332183-16>
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., ... Wright, R. (2023). Opinion paper: “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642. <https://doi.org/10.1016/j.ijinfomgt.2023.102642>
- Ferro, S., Pelillo, M., & Traviglia, A. (2023). AI-assisted digitalisation of historical documents. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-M-2–2023, 557–562. <https://doi.org/10.5194/isprs-archives-XLVIII-M-2-2023-557-2023>
- Gogora, A. (2023). Digital Research Collection – Principles and the Case of the Ethnological Collection of Research Reports. *Slovenský národopis / Slovak Ethnology*, 71(2), 107–125. <https://doi.org/10.31577/SN.2023.2.13>
- Hrubcová, E. (1973). Dokumentácia národopisných zbierok a literatúry pomocou samočinného počítača. *Informácie Slovenskej národopisnej spoločnosti*, 1(1–2), 52–59.
- Kubisa, T., & Gogora, A. (2018). Stav textového archívu Ústavu etnológie SAV a možnosti jeho digitálneho spracovania = The situation of the text archive of the Institute of Ethnology SAS and the possibilities of its digital processing. *Slovenský národopis / Slovak Ethnology*, 66(1), 97–106.
- Leontidis, G., & Royal Society Invited Scientific Contributors. (2024). Science in the age of AI: How artificial intelligence is changing the nature and method of scientific research. The Royal Society. <https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/>
- Neudecker, C. (2023). Digital Curation and AI: Opportunities and Risks for Cultural Heritage Institutions. In S. Thiel & J. C. Bernhardt (Eds.), *AI in Museums* (pp. 149–162). transcript Verlag. <https://doi.org/10.1515/9783839467107-013>
- Popelková, K., & Zajonc, J. (2008). Po prvej tisícke... Skúsenosti z tvorby elektronického katalógu obrazových dokumentov Ústavu etnológie SAV. *Slovenský národopis / Slovak Ethnology*, 56(4), 445–466.
- Prandová, E. (1978). Súčasná a perspektívne úlohy dokumentácie v Národopisnom ústave SAV. *Slovenský národopis / Slovak Ethnology*, 26(3), 509–512.
- Prandová, E. (1982). *Dokumentačné fondy Národopisného ústavu SAV 1946–1977*. Krajské osvetové stredisko v Bratislave.
- Randby, T., & Marciano, R. (2020). Digital Curation and Machine Learning Experimentation in Archives. In *2020 IEEE International Conference on Big Data (Big Data)* (pp. 1904–1913). <https://doi.org/10.1109/BigData50022.2020.9377788>
- Ratica, D. (2003). Možnosti digitalizácie archívnych dokumentov v Ústave etnológie SAV. In *Neničte Archívy! Současné metody dokumentace lidové kultury a digitalizace získaného materiálu či údajů o něm. Sborník příspěvků 19. Strážnického symposia konaného ve dnech 17. – 18. září 2003* (pp. 46–49).
- Teel, Z.A. (2024). Artificial Intelligence’s Role in Digitally Preserving Historic Archives. *Preservation, Digital Technology & Culture*, 53(1), 29–33. <https://doi.org/10.1515/pdtc-2023-0050>
- Thiel, S., & Bernhardt, J. C. (2023). *AI in Museums: Reflections, Perspectives and Applications*. transcript Verlag. <https://doi.org/10.1515/9783839467107>
- Zajonc, J., & Uhlíková, L. (2003). Univerzálna karta – Dve podoby jednej databáze. In *Neničte archívy! Současné metody dokumentace lidové kultury a digitalizace získaného materiálu či údajů o něm. Sborník příspěvků 19. strážnického symposia konaného ve dnech 17. – 18. září 2003* (pp. 39–50). <http://www.nusl.cz/ntk/nusl-22101>
- Zajonc, J. (1996). *Niektoré možnosti aplikácie výpočtovej techniky vo vedeckovýskumnom procese etnológie na Slovensku (s príkladmi z textilnej kultúry)*. Ústav etnológie SAV.
- Zajonc, J. (2006). Budovanie elektronických dokumentačných a bibliografických databáz v Ústave etnológie SAV. *Etnologické rozpravy*, 13(1), 30–47.