Knowledge Organization

Knowl. Organ. **2025**; 52(3): 39046
https://doi.org/10.31083/KO39046

Article

From Editorial Records to Structured Provenance Information: Documenting Warrant in Knowledge Organization Systems Using Large Language Models

Yi-Yun Cheng[1,*,†], Inkyung Choi[2,†]

[1]School of Communication and Information, Rutgers University, New Brunswick, NJ 08901, USA
[2]OCLC Research, Dublin, OH 43017, USA
*Correspondence: yiyun.cheng@rutgers.edu (Yi-Yun Cheng)
[†]These authors contributed equally.

Academic Editor: Natália Tognoli
Submitted: 15 March 2025 Revised: 19 April 2025 Accepted: 28 April 2025 Published: 25 June 2025

Yi-Yun Cheng is an assistant professor at the School of Communication and Information, Rutgers, the State University of New Jersey. Her research and teaching focuses on reconciling interoperability problems in taxonomies and other knowledge organization systems (KOS) in biodiversity and geographic contexts. Her work has been published in *JASIST, Information Processing & Management Journal, Journal of Documentation, Knowledge Organization Journal*, and she has co-authored a book about provenance metadata. Cheng is a leading PI of a IMLS grant project about provenance; as well as a co-PI on a U.S.DOT grant project about building a user-centered bridge technology taxonomy.

Inkyung Choi is an associate research scientist at OCLC Research, specializing in metadata research. Choi earned her PhD in Information Science from the University of Wisconsin-Milwaukee, and she currently serves on the DCMI education committee. Her work has been published in *JASIST, Journal of Documentation, Cataloging & Classification Quarterly,* and *Knowledge Organization Journal*. At OCLC Research, Choi applies a range of data analytics techniques to enhance mappings between vocabularies grounded in digital content descriptions, aiming to address domain-specific challenges. Her work also centers on embedding contextual references within classification systems to support more nuanced knowledge organization practices.

Abstract

Editorial records of a knowledge organization system (KOS) are useful for tracking the provenance of how and why a concept evolves over time. In this study, we explore the use of Large Language Models (LLMs) in obtaining structured provenance information from editorial records of KOSs. Specifically, this study focuses on one type of provenance, namely, "Warrant", which refers to external sources for decision-making on KOS changes. This study presents examples based on four Dewey Decimal Classification (DDC) editorial exhibits, each exhibit containing a discussion and proposed actions for change on a DDC topic. To explore whether LLMs can be used to extract Warrant from these DDC exhibits, we design experiments to test the models' consistency and factual accuracy. We use GPT-4o-mini with the Retrieval Augmented Generation (RAG) approach. For the system and user instructions, chain-of-thought and few-shot prompting strategies are used. For consistency, we test the impact of repeated prompting on ChatGPT's performance; for factual accuracy, we assess whether varying temperature settings yield divergent outputs. Our findings show that consistency and factual accuracy are maintained on most categories of warrant information (Document, Literature, and Concept Scheme), with an average F1-score greater than 70%. For extracting "Concept", the performance is low, with an average F1-score ranging from 30–40%. This demonstrates that ChatGPT is promising for extracting most warrant information from editorial records but still requires a human-in-the-loop verification step to fact-check the concepts extracted. Finally, a recommended process for KOS provenance documentation using LLMs is provided.

Keywords: knowledge organization systems; provenance; large language models; GPT

1. Introduction

Changes to Knowledge Organization Systems (KOSs) often require extensive discussions among editorial teams, working groups, or individuals curating the KOSs (LoC, n.d.; OCLC, n.d.). These efforts have usually been documented in various types of editorial notes, records, or meeting minutes. Some KOS editors have directly documented vocabulary changes in structured, machine-readable formats (Adolpho et al, 2023; Homosaurus, n.d.); however, most have likely archived these editorial notes informally following editorial meetings.

KOS editorial records are valuable resources for researchers and practitioners alike in understanding the provenance of a specific KOS and making more informed decisions about KOSs in the future. By "provenance", we mean the whole connected web of vocabulary changes, reasons for changes, supporting evidence, and the agents involved. Having such provenance information about a KOS can help us learn why a concept was historically accepted or resisted and how it has evolved over time (e.g., due to paradigm shifts (Tennis, 2012) in different political climates, etc.). For instance, the "Gulf of Mexico" has been the recognized vocabulary in geographic KOSs historically until recently. Being able to trace such changes through editorial documents can inform users of KOSs if and when a change occurs to "Gulf of America" due to pressure from a new United States administration (Campbell, 2025). Subsequent provenance queries that a future editor or researcher might make include: "What warranted the change of this term?" and "What was the evidence for changing the term back or changing it to another form?"

To retrieve any meaningful information from KOS editorial records, the records themselves must be in a sufficiently structured format. Extant efforts have focused on extending the use of provenance models such as PROV (https://www.w3.org/TR/prov-overview/) in metadata descriptions (Li and Sugimoto, 2018) or applying PROV to classifications in Linked Open Data formats (Lodi et al, 2014), so that descriptions can be more effectively structured. Other group efforts, like the Simple Knowledge Organization Systems (SKOS), support structured properties such as changeNote or editorialNotes (Isaac and Summers, 2009). Our prior work focuses on obtaining provenance information in a structured, ontology-compliant format that supports efficient provenance retrieval (Choi and Cheng, 2025). We model the change activities a KOS takes and suggest elements and attributes to document changes in a structured format. However, existing efforts and our own prior work mostly rely on KOS editorial teams' awareness of the model and their willingness to adopt the structured formats in their future updates to the KOSs. Editorial notes from the past that must be retroactively documented often remain buried in digital repositories.

In this study, we explore the use of Large Language Models (LLMs) to obtain structured information from KOS editorial records. LLMs, in conjunction with chatbots, make working with large texts more intuitive for users with varying levels of computational training (Aljanabi et al, 2023; Bail, 2024). This creates opportunities to examine how LLMs can support KO-related tasks while lowering technical barriers. The overarching goal of this line of research is to explore whether new, i.e., emerging, technologies such as Generative Artificial Intelligence (AI) can be systematically and automatically applied in KO practices.

Specifically, we focus first on one type of provenance information, "warrant", using sample editorial exhibits drawn from the Dewey Decimal Classification. In this study, "warrant" is defined as the "external sources used to support decision-making on KOS changes", which comprise four categories: Document, Literature, Concept, and Concept Scheme (Choi and Cheng, 2025). Each of these categories is defined in detail in the Method section. Our research questions are:

RQ1: In what ways can LLMs be leveraged to obtain warrant information from KOS editorial records? This includes two sub-questions:

RQ1.1: How can consistency be maintained when obtaining warrant information using LLMs?

RQ1.2: How does the temperature setting of an LLM affect the factual accuracy of extracted warrant information?

RQ2: What recommendations can be made for using LLMs in documenting warrant information in KOSs?

In this work, consistency in LLM evaluation is defined as providing responses that convey the same meaning in response to similar prompts, regardless of syntactical or formatting differences. If the responses diverge in meaning, they are considered inconsistent (Patwardhan et al, 2024). Further, factual accuracy in information has been a critical concern, particularly involving the veracity of data-to-text generation and online content (Goodrich et al, 2019; Lucassen and Schraagen, 2011; Thomson et al, 2023). Factual accuracy is defined as the ability of text generated by LLMs to stay true to the source facts provided.

2. Literature Review

2.1 Knowledge Organization (KO) in the Age of AI

Automating metadata and knowledge organization tasks has long been a goal in information science, as these processes are often repetitive and labor-intensive. As early as the mid-20th century, researchers discussed the potential for automatic indexing and classification to assist in organizing vast amounts of information (Greenberg et al, 2021). By the early 21st century, the desire for automation had only increased, with the advent of the Web, the Semantic Web, and Linked Data (Berners-Lee and Hendler, 2001; Lassila et al, 2001). By then, KO researchers and information scientists had begun to struggle with the organization of born-digital content and the exponential growth of web-driven information.

In the KO community, there were discussions about how to navigate between Knowledge Representation (KR) and KO and how KR's focus on ontology reasoning might enable a more automated and dynamic approach to organizing vocabularies (Giunchiglia et al, 2014; Qin, 2020). Other KO researchers leveraged probabilistic computational approaches, such as machine learning (ML), natural language processing (NLP), and clustering, rather than relying on deterministic reasoning with ontologies. These approaches were used to automate vocabulary generation, selection, domain analysis, document classification, and indexing (Greenberg et al, 2021; Roitblat et al, 2010; Smiraglia and Cai, 2017; Westin, 2024). Under both approaches, the objectives were mostly to provide structured, machine-interpretable content, concepts, or vocabularies to enhance the organization and retrieval of documents, collections, or items.

Artificial Intelligence (AI) has long existed as a research field focused on enabling machines to perform tasks that require human-like intelligence (McCarthy et al, 2006). Broadly speaking, approaches and technologies in the data sciences (e.g., ML, NLP, deep learning, etc.) all constitute AI technologies. With the recent launch of ChatGPT by OpenAI, the notion of AI was popularized among the general public and has since largely been equated with generative models (GenAI), GPT, and LLMs. Though GenAI and LLMs remain in their nascent phase, researchers are increasingly experimenting with LLMs, seeking to automate the creation of KOSs such as taxonomies or ontologies (Sun et al, 2024). Many studies outside of KO—for example, of the use of LLMs to review research articles or serve as "AI" reviewers—suggest that LLMs can generate generic feedback and evaluations, though human input is still needed (Liang et al, 2024; Zhou et al, 2024). Whether LLMs can fully replace human effort remains an open and debated question for KO scholars to continue exploring.

In adopting GenAI and LLMs, KO scholars must consider various ethical concerns that have emerged. One long-standing research problem in KO, surrounding the misrepresentation, underrepresentation, or absence of minorities in vocabularies (Higgins, 2016; Littletree and Metoyer, 2015), is echoed and perpetuated in ML, NLP, and LLM technologies. For instance, Rosa et al. (2024)'s work examining GenAI's capacity for knowledge representation of images showed that AI-generated images often replicate and reinforce existing societal biases. Moreover, growing reliance on AI-generated suggestions has raised concerns about its role in decision-making processes (Buçinca et al, 2021). How can we trust GenAI when the LLMs behind it are susceptible to "hallucination", biases, and incorrectness? Or, how can governance be established to guide the use of AI within a healthy KO ecosystem, ensuring that emergent technologies are applied in sustainable and responsible KO practices (Bagchi, 2021)?

2.2 Provenance and Warrant

Extant studies emphasized the broader utility of provenance in KOS. For instance, Turner, Bruegeman, and Moriarty (2024) highlighted how provenance captures the cultural and epistemological factors shaping knowledge systems when reconstructing historical warrants and contextualizing cataloging decisions. Similarly, Zhang et al. (2020) formalized provenance graphs in NLP to trace the origins and evolution of claims and found that provenance graphs are reliable to support the verification process in digital systems. Together, these approaches illustrate how provenance not only documents change but also facilitates trust in knowledge systems.

In the field of KO, Warrant has historically been a critical evaluative mechanism in the development and maintenance of a KOS. For example, literary warrant justified the inclusion of concepts based on published materials, while user warrant focused on concepts according to user needs (Beghtol, 1986). In recent years, KO researchers have been studying how Warrant guides classification systems and vocabularies, particularly in justifying changes prompted by cultural, technological, or contextual shifts (Dobreski, 2020; Martínez-Ávila and Budd, 2017). This dual role of Warrant, as a decision-making and record-keeping mechanism, naturally overlaps with provenance in documenting the rationale for changes.

Building on this foundation, our prior work incorporated the principles of Warrant in KO with provenance tracking to allow for more transparent documentation of changes in KOSs. By incorporating provenance information on change activities and supporting evidence, our model not only captures the rationale behind changes but also clarifies how a KOS evolves over time (Cheng et al, 2024; Choi and Cheng, 2025). Standards like World Wide Web Consortium (W3C) PROV and Resource Description Framework (RDF) enable automated systems to trace modification histories, while warrant information offers a contextual understanding of those changes (Hartig and Zhao, 2010; Markovic et al, 2014). For instance, the SC-PROV vocabulary proposed by Markovic et al. (2014) captured provenance in social computations and demonstrated its ability to assess digital interactions. Zhang et al. (2020) highlighted the potential of natural language processing to extract structured provenance data from unstructured sources.

2.3 The Use of LLMs

In this work, we leverage LLMs while also experimenting on their consistency and factual accuracy in extracting information. Here, we provide a brief background about LLM fine-tuning: Research indicates that LLM performance can be improved through fine-tuning of the model's parameters; for tasks requiring the extraction of factual information, adjusting LLMs to a moderate or lower "temperature" setting can help minimize creative outputs

(Peeperkorn et al, 2024; Renze, 2024; Zhao et al, 2025). However, while LLMs show potential for automating KO practices, their reliability in extracting warrant information requires further investigation.

In this study, we explore the utility of LLMs in extracting structured warrant information from diverse sources and investigate their consistency and factual accuracy. By leveraging LLMs, we investigate the potential of using LLMs to automate the extraction of warrant information from KO editorial records.

3. Method

3.1 Data

Four Editorial Policy Committee (EPC) documents were used in this study. These documents were collected from the Dewey Editorial Policy Committee webpage, with each containing information about changes proposed to a Dewey Decimal Classification (DDC) topic. Each of the four documents includes mentions of the categories Document, Literature, Concept, and Concept Scheme—the essential warrant entities defined in our prior model (Choi and Cheng, 2025). We refer to these documents as "Exhibits" in the following sections. To construct ground truth data for what constitutes Document, Literature, Concept, and Concept Scheme, the authors reviewed the four exhibits and annotated each instance of warrant information accordingly.

3.2 Experimental Design

Our experiments involve extracting relevant concepts from the Exhibits according to the classes in our provenance-based KO model. The main goals of this experiment were to: (1) document the warrant for changes to KOS concepts, as indicated in the Exhibits; and (2) explore the feasibility of extracting unstructured free-text warrant information and converting it into structured provenance information.

We used GPT-4o mini in our experiments, with Retrieval-augmented generation (RAG), which retrieves and generates answers based only on the four Exhibit files provided. LLMs demonstrate strong factual recall in controlled environments but struggle with nuanced fact-checking tasks, leading to concerns about hallucinations and over-reliance by users (Laban et al, 2023; Jiang et al, 2024). The RAG approach, which retrieves and answers only from the sources or training data, can help improve factual accuracy compared to standalone LLMs (Wang et al, 2023).

We conducted two experiments:

Experiment 1 tested the consistency of LLM prompt responses. In this experiment, one key variable, repeated prompting, was used to assess consistency across multiple iterations of the same prompt. The same prompt was issued five times. Every iteration started as a new session, ensuring no memory or context carried over between prompts. The temperature parameter for this experiment was set at the default value (1.0).

Experiment 2 tested the factual accuracy of LLM prompt responses. In this experiment, one key variable, the temperature setting of the model, was used to control for the randomness and creativity of the model's output. We used two different temperature settings in GPT-4o-mini to determine whether there were any differences between them. One was the default temperature setting of 1.0, which typically allows for more diverse responses; the other was the most reduced temperature setting of 0.1, to elicit more faithful outputs. Additionally, we conducted a set of consistency experiments with repeated prompting at the reduced temperature (0.1).

3.3 API Pipeline

To streamline the experiment, an automated pipeline using the OpenAI API(Application Programming Interface) was implemented. The pipeline facilitates repeated prompting and response collection systematically.

Step 1: API Integration with RAG — We employed the OpenAI API to prompt the model against a file-based database containing the four Exhibits. This setup qualifies as RAG because the API retrieves relevant document segments before generating a response, ensuring that all outputs are grounded in the source materials. The advantage of this approach is twofold: first, it minimizes hallucinations, enhancing the reliability of responses; second, it allows us to focus specifically on the content of the Exhibits without noises from external sources.

Step 2: Automated Queries to API via Python Pipeline — We extended the API pipeline with a Python-based automation script to handle repeated prompts efficiently. This script performs the following tasks:

- Sends API requests with predefined parameters (temperature settings and prompts);
- Records responses systematically, associating each with its respective temperature setting and iteration count;
- Facilitates subsequent analysis by exporting responses to a structured format (.json) for comparison and consistency assessment.

3.4 Prompting Strategy and Prompt Tuning

Under our initial prompting strategy, we prompted the GPT API assistant to perform two tasks: (1) extract relevant warrant information (Documents, Literature, Concepts, and Concept Schemes) from the Exhibits based on our provenance-based conceptual model; and (2) translate the extracted information into Web Ontology Language (OWL) syntax following the model's ontology structure.

After several preliminary trials, we found that providing the model's OWL file directly in the system instructions could improve the assistant's accuracy in the second task. However, there were challenges in the first task in accurately extracting relevant concepts. The assistant struggled

to identify and classify the subclasses consistently, which introduced inaccuracies in the subsequent OWL translation step. Recognizing that reliable extraction is a prerequisite to successful ontology construction, we decided to focus exclusively on the first task.

Thus, our final prompt instructs the assistant to only extract warrant information without proceeding to OWL translation. This change was intended to improve the extraction accuracy before reintroducing the assistant to proceed with the second task. We adopted modified chain-of-thought and few-shot prompting strategies (Li et al, 2024; Ma et al, 2023) in the system's instruction of RAG to provide clear definitions and examples of what we aimed to extract.

The following sections display the final system instruction and user prompt used consistently across all four exhibits. The instructions for the second task on translating everything into OWL, though not discussed further in the main text, are included in Appendix A.

System Instruction:

You are an intelligent and honest agent who provides exactly what is presented in the input text. Be as thorough and exhaustive as you can. You have two tasks. The first task is to extract the Warrant and its subclasses from the documents, and the second task is to structure these based on ProvKOS into ontology-compliant format (e.g., RDF triples).

The first task: The "Warrant" class is to represent all types of sources used during the decision-making on the vocabulary changes in knowledge organization systems (KOS). When KOS changes are suggested during the editorial discussions, the editorial documents frequently include references to external resources. These external resources include citing editorial documents, referencing literature about a topic in discussion, or connecting to other relevant concepts and concept schemes. The subclasses of Warrant are Document, Literature, Concept, and Concept Schemes.

Document refers to the editorial records itself.

Literature means citable publications (e.g., scholarly articles, web links, news reports); these should be either cited in the main texts or appear in the reference section.

Concept means a controlled vocabulary or subject from external Knowledge Organization Systems (e.g., ICD, Homosaurus, Getty thesaurus, LCSH). It should not contain the vocabulary from the KOS in the editing process, for example, if we are investigating DDC, then concepts mentioned in the Document that are from DDC should not be listed as a Concept here.

Concept schemes means other Knowledge Organization Systems (e.g., ICD, Homosaurus, Getty thesaurus, LCSH). It should not contain the KOS in the editing process, for example, if we are investigating DDC, then the Concept scheme should not include DDC. Some notable common concept schemes are: DSM-5, ICD-11, LCSH, MeSH, Homosaurus.

Here is an example of the input file and the expected output:

Input: Exhibit 155-Drag.pdf

Output:

- Document: Exhibit 155-Drag

-Literature:

Blazucki, Sarah, and Jeff McMillan, eds. "NLGJA Stylebook on LGBTQ Terminology."

Stylebook on LGBTQ+ issues, March 2023. https://www.nlgja.org/stylebook/.

"Cross-Dressing." Wikipedia. Wikimedia Foundation, March 24, 2023.

https://en.wikipedia.org/wiki/Cross-dressing.

"Cross-Gender Acting." Wikipedia. Wikimedia Foundation, November 8, 2022.

https://en.wikipedia.org/wiki/Cross-gender_acting.

"Homosaurus Vocabulary Terms." Homosaurus Vocabulary Site. Accessed March 29, 2023.

https://homosaurus.org/v3.

-Concept:

Drag performance

Drag shows

Performing arts

Theater

Drag

-Concept scheme:

LCSH

Homosaurus

User instruction (prompt):

Please read [Exhibit_Name.pdf], and complete the first task given in the instructions. Do not move forward to the 2nd task.

[Exhibit_Name.pdf] is the DDC editorial exhibits. In the document, DDC is commonly mentioned in the form of class number and human readable labels such as '613.04 Personal health of people by gender, sex, or age group'.

3.5 Data Analysis

For both Experiment 1 and 2, precision, recall, and F1-score were calculated across all iterations. For instance, in Table 1, given the ground truth data in the Concept Scheme for the Exhibit on Gender Dysphoria {Homosaurus, ICD-11, MeSH} and low temperature model output {ICD-11, Homosaurus, DSM-5}, two true positives (TP = 2) were correctly identified by the model {Homosaurus, ICD-11}. There was one false positive (FP = 1), where the model incorrectly predicted DSM-5 as relevant, and one false negative (FN = 1), where the model failed to identify MeSH.

Table 1. Example model outputs of one exhibit in comparison with the ground truth data.

Category	Ground Truth	LLM (Default Temp 1.0)	LLM (Low Temp 0.1)
Document	EPC 144-S61.1 Gender dysphoria	EPC 144-S61.1 Gender Dysphoria	EPC 144-S61.1 Gender dysphoria
Literature	1. Anderson 2022 2. WHO's ICD-11 3. Darcy 2022 4. Homosaurus Vocabulary Terms 5. Mayo Clinic 2022 6. Rodríguez et al. 2018	1. Anderson, Danyon et al. 2. Darcy, Andrea M. 3. Mayo Clinic 4. Homosaurus Vocabulary Terms 5. Rodríguez et al.	1. Anderson, Danyon et al. 2. Darcy, Andrea M. 3. Mayo Clinic 4. Homosaurus Vocabulary Terms 5. Rodríguez et al.
Concept	1. Gender dysphoria 2. Sexual health 3. Gender Identity 4. Transsexualism 5. Sexual Dysfunctions 6. Sexual Behavior 7. Transgender persons	1. Gender dysphoria 2. Depressive Disorder 3. Transgender identity 4. Intersexuality	1. Gender dysphoria 2. Depressive disorder 3. Dysphoria 4. Transgender identity 5. Intersexuality
Concept Scheme	1. Homosaurus 2. ICD-11 3. MeSH	1. DDC (Dewey Decimal Classification) 2. ICD-11 3. Homosaurus Vocabulary	1. ICD-11 2. Homosaurus 3. DSM-5

Notes: LLM, Large Language Model; EPC, Editorial Policy Committee; DDC, Dewey Decimal Classification; DSM, Diagnostic and Statistical Manual of Mental Disorders; ICD, International Classification of Disease; MeSH, Medical Subject Headings; WHO, World Health Organization.

To calculate precision, recall, and F1-Score, we used the following formula:

$$Precision = \frac{TP}{TP + FP} = \frac{2}{2+1} = \frac{2}{3} \approx 0.667 \quad (1)$$

$$Recall = \frac{TP}{TP + FN} = \frac{2}{2+1} = \frac{2}{3} \approx 0.667 \quad (2)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} = 2 \times \frac{0.667 \times 0.667}{0.667 + 0.667}$$
$$= 2 \times \frac{0.445}{1.334} \approx 0.667 \quad (3)$$

Since the prompt was repeated in all exhibits five times, the average precision, recall, and F1-score were calculated across the four categories (Document, Literature, Concept, Concept Scheme) for the model's outputs. For example, if the above case were repeated five times and yielded precisions of 0.667, 0.8, 0.667, 0.54, and 0.8 for Concept Scheme, the model's output achieved an average precision of 0.695.

4. Findings

4.1 Large Language Models can be Leveraged in Obtaining Warrant Information

Our first research question asks: In what ways can LLMs be leveraged to obtain warrant information from KOS editorial records? In our experiments, we have been able to successfully obtain warrant information from editorial documents, using LLMs. The warrant information extracted by LLMs was compared against the ground truth data, and the results were satisfactory. In the section below, we discuss our findings on how ChatGPT, as an example of LLMs, can successfully extract some categories of warrant information, though other categories remain challenging even for human annotators. We also discuss how temperature settings can be adjusted to maintain factual accuracy of prompt answers.

4.2 Consistency is Maintained in Several Categories of Warrant Information

To address our sub-research question RQ1.1 "How can consistency be maintained when obtaining warrant information using LLMs?", Fig. 1 presents the results of our experiment on the repeatability of prompt answers. The table, which provides all performance scores is provided in Appendix B. Fig. 1 demonstrates the five iterations for all four exhibits and the model's performance in each category (Document, Literature, Concept, Concept Scheme) compared with the ground truth data.

In the Document category, in every iteration across all Exhibits, ChatGPT performed extremely well in terms of consistency, repeating the exact same answers in all instances. Despite lower recall, resulting in a lower overall F1-score, the five iterations for Exhibit 1 remained consistent. For Literature, consistency was maintained in three Exhibits. Though the recall in the three cases did not equal 1, every iteration of these three exhibits produced the same answers. For Exhibit 4, consistency was not achieved; only

two of five iterations produced the same answers. For Concept Scheme, consistency was achieved perfectly in Exhibit 3. In Exhibit 2 and Exhibit 4, four of five iterations produced the same answers with ChatGPT. For Exhibit 1, two pairs of iterations produced the same answers. For Concept, there was a lack of consistency across iterations for all four exhibits. There was some limited consistency for Concept in Exhibit 2; otherwise, the Concept category has proven to be the most challenging in terms of consistency.

Overall, consistency was maintained for the Document and Literature categories of the warrant information. Performance on Concept Scheme was still marginally acceptable, but consistency in the Concept category was notably poor. In our observation, one of the keys to producing consistent outputs with ChatGPT could be the number of items extracted. The Document category is usually straightforward, with only 1–2 items; hence, the consistency performance may have been perfect due to this factor. The number of items to be extracted varies in the Literature category, ranging from 3 to 17 items. In Exhibits 1, 2, and 3, consistent performance in Literature was very stable because there are only 5 to 6 items. For Exhibit 4, there were 17 references in total to be extracted, and this may explain the variability in the five iterations of Exhibit 4.

As for Concept and Concept Scheme, we suspect several factors: first, despite taking the RAG approach to making ChatGPT more faithful in its generated answers, ChatGPT is likely not as familiar with KOS-specific ideas like Concepts and Concept Schemes; second, Concept Scheme sometimes functions similarly to the Literature category, and ChatGPT mistakenly regards some concept schemes as literature in certain outputs; and last, a stricter definition of Concept was used in our warrant information extraction task, and ChatGPT may have skipped that definition in some of the iterations. Some vocabulary placed in quotation marks for emphasis in the Exhibits, but not actual KOS concepts, may have been incorrectly identified as Concepts by ChatGPT.

4.3 Temperature Setting Will Impact the Factual Accuracy of Warrant Information

Our sub-question RQ1.2 asks, "How does the temperature setting of an LLM affect the factual accuracy of extracted warrant information?". To address this, we tested the variation across temperature settings in ChatGPT, finding that chats with a lower temperature setting (temperature at 0.1) performed consistently better than chats at the default temperature (temperature = 1.0) in all categories of warrant information (Table 2). In all categories, the low temperature setting chats had better precision, recall, and F1-scores. Except for the Concept category, all categories consistently achieved an average precision above 84%, average recall above 75%, and an average F1-score above 78% in the low temperature setting. Even in the default temperature setting, ChatGPT achieved an average precision above 73%, average recall above 65%, and an average F1-score above 72% in all categories except Concept.

Notably, precision in both temperature settings is 100% in the Document and Literature categories. We attribute this to the tendency of LLMs to over-compensate prompt responses and include more matches—even at the expense of sacrificing recall (i.e., potentially omitting relevant but harder-to-identify items). For the Concept category, though the overall performances in both temperature settings are not ideal, there is a ~10% increase in performance when switching to the low temperature. This suggests that a lower temperature setting is more ideal for faithfully extracting all categories of warrant information.

Overall, there are noticeable differences in performance across different temperature settings. This means that at lower temperatures, ChatGPT is unlikely to generate content deviant from the ground truth data. For instance, at the higher temperature setting, ChatGPT identifies two Concept schemes mentioned in Exhibit 4, while the low temperature faithfully matches the ground truth data, determining only one concept scheme in that specific Exhibit.

5. Recommended Process for Incorporating LLMs in Extracting Provenance Information

Our second research question asks, "What recommendations can be made for using LLMs in documenting warrant information in KOS?" Our results indicate that while ChatGPT performs well in extracting two subclasses, Document and Literature, its performance is limited in the Concept and Concept Scheme categories. Given these limitations, we propose a semi-automated approach that integrates LLMs with additional verification mechanisms (Fig. 2).

For example, we propose implementing an extra verification layer that cross-checks extracted concepts against the original concept scheme as well as external authoritative concept schemes (e.g., MeSH or LCSH), as illustrated by the process in Fig. 2, which verifies (d) LLMs' output against (a) KOSs. By doing so, a system can automatically validate extracted concepts by linking directly to KOSs (e.g., linking to the MeSH database) to confirm the existence of a concept. This approach would ensure that only concepts recognized in established schemes are recorded, reducing the likelihood of inaccurate or false detection of concepts by LLMs. The variability in the Concept category shows ChatGPT's limitation in following defined instructions and its tendency to include a broader set of concepts, including terms that are not 'controlled' or standardized (e.g., any terms placed in quotation marks might be misidentified as Concepts). A linkage to existing KOSs could further serve as a filter, so that the output warrant information adheres to standardized vocabularies.

Another recommendation is to include humans in the process to assess the relevance and correctness of extracted concepts. Echoing existing literature, we found that Chat-

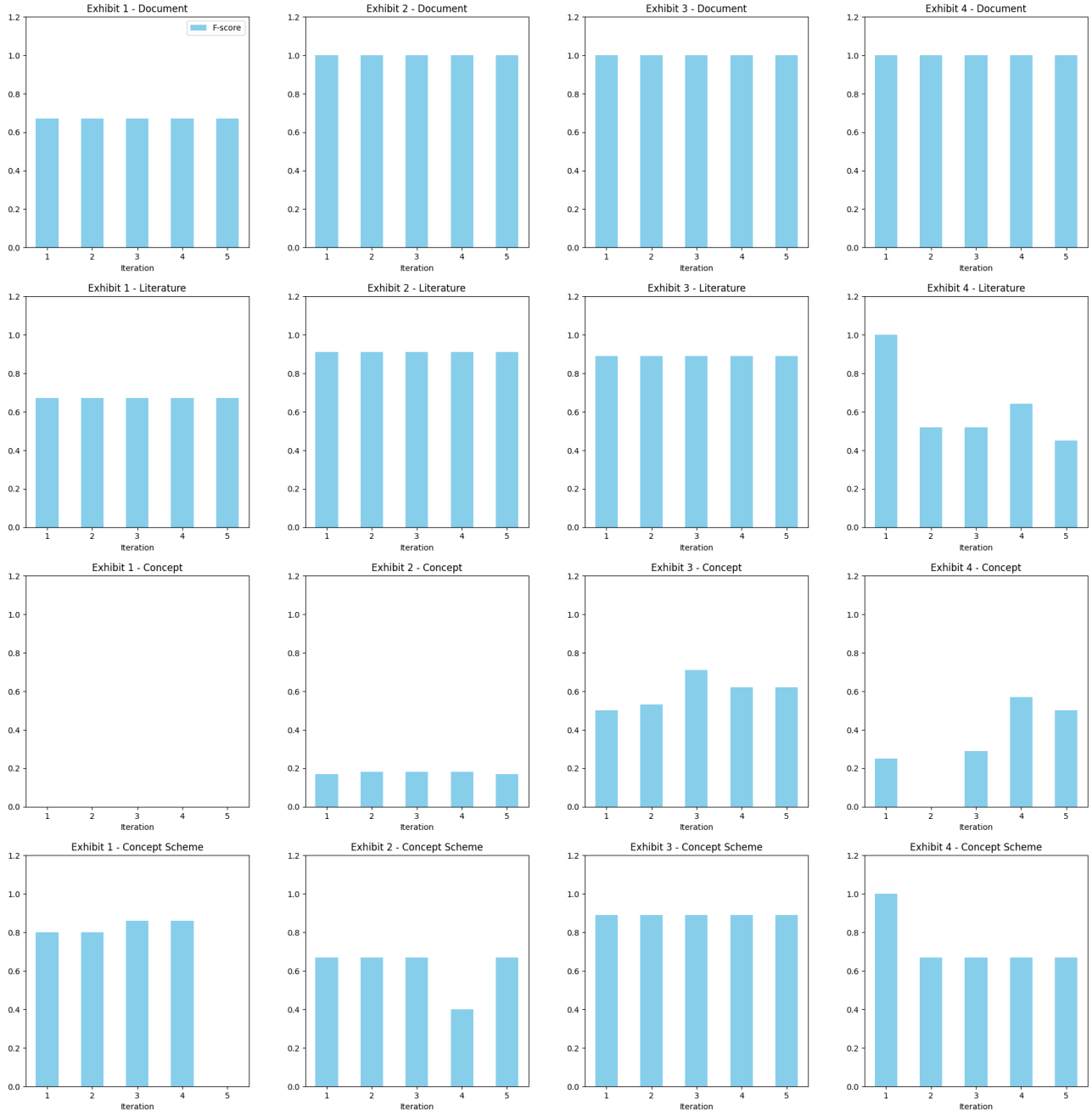

Fig. 1. The performance in terms of F1-scores of the five iterations on all four Warrant categories (Document, Literature, Concept, Concept Scheme) of all four exhibits.

GPT can complement human efforts but cannot completely replace the nuanced reasoning required for concept identification. The high variability in Concept extraction in this study further demonstrates the need for a human-in-the-loop to mitigate the risks of false positives (terms incorrectly identified as controlled vocabulary). In Fig. 2, there are two segments that depict the process channeling down from (a) KOS to the (b) Editorial Team. For retroactively extracting warrant information from past editorial records (the top grey box in Fig. 2), we recommend following a process similar to the example case in our study, that is, taking the pipeline of (a) → (b) → (c) → (d) → (a) → (b) → (e). This pipeline involves the editorial team at the outset, when they 'document' changes in editorial records, and at the end, verifying the output generated by LLMs through external KOS verification. We also recommend that for future processes in documenting KOS changes (the bottom grey box in Fig. 2), the editorial team bypass LLMs altogether and directly inscribe structured provenance information instead of unstructured editorial records (Fig. 2, (a) → (b) → (e)).

Finally, we recommend setting the temperature parameter lower for tasks that require truthful representations

Table 2. Comparison of performance in terms of average scores (Precision, Recall, F-1) under two temperature setting.

	Average (default temp)	Average (low temp)
Document		
Precision	1.000	1.000
Recall	0.875	0.875
F1-Score	0.918	0.918
Literature		
Precision	1.000	1.000
Recall	0.656	**0.719**
F1-Score	0.774	**0.828**
Concept		
Precision	0.375	**0.475**
Recall	0.269	**0.497**
F1-Score	0.288	**0.427**
Concept Scheme		
Precision	0.734	**0.846**
Recall	0.768	**0.752**
F1-Score	0.727	**0.785**

Boldfaced scores indicate better performance between the two.

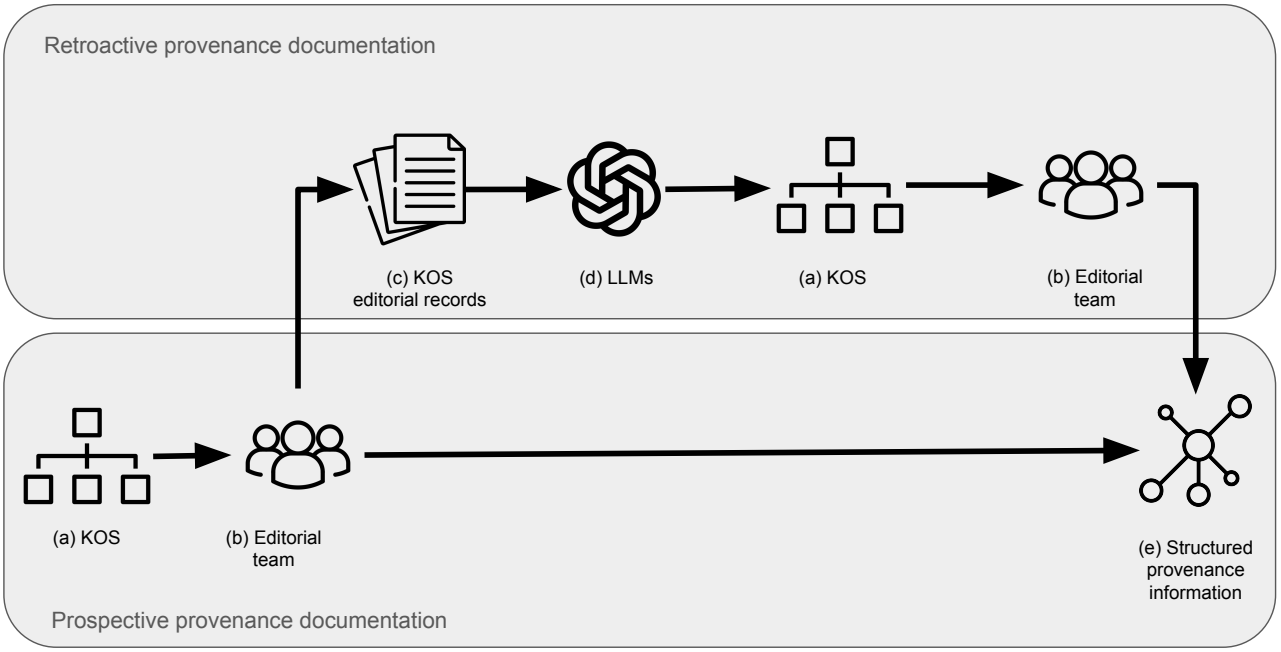

Fig. 2. A recommended process for incorporating LLMs in extracting provenance information based on our experiment findings. LLMs, Large Language Models; KOS, Knowledge Organization System.

of prompt answers. Lower temperature settings reduce the risk of false answers and can enhance the consistency of responses, making them ideal for extracting warrant information and, by extension, structured provenance information, which requires precise adherence to the source KOSs.

6. Conclusion

In this study, we experimented with the use of LLMs to obtain warrant information from KOS editorial records. We find that LLMs, specifically ChatGPT, can be used to obtain warrant information. Consistency and factual accuracy are maintained in most categories of warrant; however, they are lacking when it comes to extracting concepts from editorial records.

We further recommend a semi-automated process implementing (1) an LLM-driven extraction pipeline that identifies and classifies warrant information according to our provenance-based conceptual model; (2) a verification step with a human-in-the-loop to validate whether extracted warrants align with existing KOSs; (3) another human-in-

the-loop process involving editors to confirm the accuracy of the validated suggestions, focusing on review rather than initial extraction. By reducing manual effort, this semi-automated process can streamline the provenance documentation of KOSs and make documentation more scalable and efficient.

To adopt the provenance-based conceptual model for prospective warrant documentation, we recommend that the KOS editorial process incorporate the documentation of structured information in the first place. Editorial guidelines should outline explicit warrant statements that cite the specific class (Document, Literature, Concept, or Concept Scheme) and provide a reference to the source. Moreover, warrants classified under the Concept or Concept Scheme categories must be validated against recognized vocabularies, such as DDC, MeSH, or LCSH. For the Literature category, designing templates for a brief justification of changes and references or links to supporting documents can further assist KOS editors.

Key directions for future research, informed by the limitations of the current study, include: (1) Model selection and Performance: This study used GPT-4o-mini due to computational constraints. Future work will systematically compare the performance of more advanced models (e.g., GPT-4o and GPT-4.5) on provenance reasoning tasks. A comparative evaluation with other metrics (e.g., BLEU, ROUGE) to establish benchmarks for LLM performance on KO tasks will also be considered; (2) Expansion beyond warrant information: This study focused solely on warrant information. To make this process more generalizable and applicable to all types of provenance information, we envision experimenting with LLMs on other types of provenance information such as changes and agents; (3) Adopting the pipeline in practice: We propose leveraging LLMs in KO practices. In subsequent work, we hope to investigate the challenges in implementing such practices for prospective warrant documentation through a qualitative study examining the costs, benefits, barriers, and resistance faced by editorial teams and practitioners; (4) Structured provenance information in ontologies: We have devised a prompt for LLMs to format the extracted provenance information into an ontology-compliant format (Appendix A). Future work will refine these prompts and apply them to a real-world KOS example. By centering editorial practices around systematic provenance documentation, changes in knowledge organization systems will not only be well-recorded but also more easily—and even automatically—captured.

Abbreviations

API, Application Programming Interface; AI, Artificial Intelligence; DDC, Dewey Decimal Classification; DSM, Diagnostic and Statistical Manual of Mental Disorders; EPC, Editorial Policy Committee; GenAI, Generative AI; GPT, Generative Pre-trained Transformer; ICD, International Classification of Disease; KO, Knowledge Organization; KOS, Knowledge Organization System; KR, Knowledge Representation; LLMs, Large Language Models; LCSH, Library of Congress Subject Headings; ML, Machine Learning; MeSH, Medical Subject Headings; NLP, Natural Language Processing; OWL, Web Ontology Language; RDF, Resource Description Framework; RAG, Retrieval Augmented Generation; SKOS, Simple Knowledge Organization Systems.

Availability of Data and Materials

The exhibits, prompts, API python script and the results from our data analysis are publicly available on our online repository: https://doi.org/10.5281/zenodo.15242754.

Author Contributions

YYC and IC contributed equally to the conceptualization, data curation, formal analysis, investigation, methodology, validation, visualization, and writing of this study. Both authors have read and approved the final version of this manuscript. Both authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

Acknowledgment

Not applicable.

Funding

This research received no external funding.

Conflict of Interest

The authors declare no conflict of interest.

Declaration of AI and AI-Assisted Technologies in the Writing Process

ChatGPT is the core focus of this research in terms of studying its applicability towards KO-tasks, therefore was used throughout this research. During the preparation of this manuscript, the authors used ChatGPT-4 to check grammar. The authors reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Appendix

Appendix A. Prompt for the second task.

Here we illustrate an example of the prompt we have used for the second task on structuring the warrant information in an ontology-compliant format. The full prompt is available in our online repository: https://doi.org/10.5281/zenodo.15242754.

The second task: Based on the warrant information you extracted from the first task, structure them based on the provKOS.owl file. Use the prefix provKOS. for classes and properties native to ProvKOS. ontology only. Then print

them.

For example, the Warrant information of ‘EPC 144-S61.1 Gender dysphoria.pdf’ can be represented like:

<rdf:RDF xmlns=“https://w3id.org/def/ ProvKOS#”
xml:base=“https://w3id.org/def/ ProvKOS”
xmlns:dc=“http://purl.org/dc/elements/1.1/”
xmlns:owl=“http://www.w3.org/2002/07/owl#”
xmlns:rdf=“http://www.w3.org/1999/02/22-rdf-syntax-ns#”
xmlns:xml=“http://www.w3.org/XML/1998/namespace”
xmlns:xsd=“http://www.w3.org/2001/XMLSchema#”
xmlns:prov=“http://www.w3.org/ns/prov#”
xmlns:rdfs=“http://www.w3.org/2000/01/rdf-schema#”
xmlns:skos=“http://www.w3.org/2004/02/skos/core#”
xmlns:skos-xl=“http://www.w3.org/2008/05/skos-xl#”>
<!– Document –>
<owl:NamedIndividual rdf:about=“http://example.org#EPC_144-S61.1”>
<rdf:type rdf:resource=“https://w3id.org/def/ ProvKOS #Document”/>
<prov:wasGeneratedBy rdf:resource=‘‘http://example.org/deweyeditorialactivity/23/349292”/><!--Placeholderactivityforgeneration-->
<citerdf:resource=‘‘http://example.org#Anderson_Gender_Dysphoria_Article”/>
<citerdf:resource=‘‘http://example.org#Darcy_What_Is_Dysphoria”/>
<citerdf:resource=‘‘http://example.org#MayoClinic_Gender_Dysphoria”/>
<citerdf:resource=‘‘http://example.org#Homosaurus_Vocabulary”/>
</owl:NamedIndividual>
<!– Concepts –>
<owl:NamedIndividual rdf:about=“http://example.org#Gender_Dysphoria”>
<rdf:typerdf:resource=‘‘http://www.w3.org/2004/02/skos/core#Concept”/>
<rdfs:label>Gender Dysphoria</rdfs:label>
<skos:inScheme rdf:resource=“http://example.org#ICD-11”/> <!– Link to concept scheme –>
</owl:NamedIndividual>

Appendix B. Full Results of the Data Analysis: Performance of five iterations on four categories of all four exhibits.

Document				Literature				Concept				Concept Scheme			
Exhibit 1				Exhibit 1				Exhibit 1				Exhibit 1			
Iteration	precision	recall	f-score	Iteration	precision	recall	f-score	Iteration	precision	recall	f-score	Iteration	precision	recall	f-score
1	1	0.5	0.67	1	1	0.5	0.67	1	0.20	0.50	0.29	1	1	0.67	0.80
2	1	0.5	0.67	2	1	0.5	0.67	2	0	0	0	2	1	0.67	0.80
3	1	0.5	0.67	3	1	0.5	0.67	3	0	0	0	3	0.75	1	0.86
4	1	0.5	0.67	4	1	0.5	0.67	4	0	0	0	4	0.75	1	0.86
5	1	0.5	0.67	5	1	0.5	0.67	5	0	0	0	5	0	0	0
Exhibit 2				Exhibit 2				Exhibit 2				Exhibit 2			
Iteration	precision	recall	f-score	Iteration	precision	recall	f-score	Iteration	precision	recall	f-score	Iteration	precision	recall	f-score
1	1	1	1	1	1	0.83	0.91	1	0.20	0.14	0.17	1	0.67	0.67	0.67
2	1	1	1	2	1	0.83	0.91	2	0.25	0.14	0.18	2	0.67	0.67	0.67
3	1	1	1	3	1	0.83	0.91	3	0.25	0.14	0.18	3	0.67	0.67	0.67
4	1	1	1	4	1	0.83	0.91	4	0.25	0.14	0.18	4	0.50	0.33	0.40
5	1	1	1	5	1	0.83	0.91	5	0.20	0.14	0.17	5	0.67	0.67	0.67
Exhibit 3				Exhibit 3				Exhibit 3				Exhibit 3			
Iteration	precision	recall	f-score	Iteration	precision	recall	f-score	Iteration	precision	recall	f-score	Iteration	precision	recall	f-score
1	1	1	1	1	1	0.80	0.89	1	0.80	0.36	0.50	1	1	0.80	0.89
2	1	1	1	2	1	0.80	0.89	2	1	0.36	0.53	2	1	0.80	0.89
3	1	1	1	3	1	0.80	0.89	3	1	0.55	0.71	3	1	0.80	0.89
4	1	1	1	4	1	0.80	0.89	4	1	0.45	0.62	4	1	0.80	0.89
5	1	1	1	5	1	0.80	0.89	5	1	0.45	0.62	5	1	0.80	0.89
Exhibit 4				Exhibit 4				Exhibit 4				Exhibit 4			
Iteration	precision	recall	f-score	Iteration	precision	recall	f-score	Iteration	precision	recall	f-score	Iteration	precision	recall	f-score
1	1	1	1	1	1	1	1	1	0.20	0.33	0.25	1	1	1	1
2	1	1	1	2	1	0.35	0.52	2	0	0	0	2	0.50	1	0.67
3	1	1	1	3	1	0.35	0.52	3	0.25	0.33	0.29	3	0.50	1	0.67
4	1	1	1	4	1	0.47	0.64	4	0.50	0.67	0.57	4	0.50	1	0.67
5	1	1	1	5	1	0.29	0.45	5	0.40	0.67	0.50	5	0.50	1	0.67

References

Adolpho K, Billey A, Johnson J, Kizzie J, Rawson KJ, Roles, C, et al. Homosaurus Documentation and Implementation. 2023. Available at: https://docs.google.com/document/d/1PgwSKGHnr4dokazFRbkoqQuYWbw8c_0LWYw0L2rkyAw/edit?tab=t.0 (Accessed: 11 February 2025).

Aljanabi M, Ghazi M, Ali AH, Abed SA. ChatGpt: open possibilities. Iraqi Journal for Computer Science and Mathematics. 2023; 4: 7. https://doi.org/10.52866/ijcsm.2023.01.01.0018

Bagchi M. Towards knowledge organization ecosystem (KOE). Cataloging & Classification Quarterly. 2021; 59: 740–756. https://doi.org/10.1080/01639374.2021.1998282

Bail CA. Can Generative AI improve social science?. Proceedings of the National Academy of Sciences. 2024; 121: e2314021121. https://doi.org/10.1073/pnas.2314021121

Beghtol C. Semantic validity: concepts of warrant in bibliographic classification systems. Library Resources & Technical Services. 1986; 30: 109–125.

Berners-Lee T, Hendler J. Publishing on the semantic web. Nature. 2001; 410: 1023–1024. https://doi.org/10.1038/35074206

Buçinca Z, Malaya MB, Gajos KZ. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction. 2021; 5: 1–21. https://doi.org/10.1145/3449287

Campbell J. 'Gulf of America' arrives on Google Maps. 2025. Available at: https://edition.cnn.com/2025/02/11/business/trump-gulf-of-america-google-maps-hnk-intl/index.html (Accessed: 11 February 2025).

Cheng YY, Choi I, Bettivia R, Lee WC, Watson BM. Knowledge Organization Systems and Provenance: Experiences and Challenges. Proceedings of the Association for Information Science and Technology. 2024; 61: 741-744.

Choi I, Cheng YY. A conceptual model for tracking the provenance of activities in knowledge organization systems. Journal of documentation. 2025; 81: 147-167.

Cremonez Rosa P, Barizon Filho AL, Torrão Valentim R, Tognoli N. Datafication, Artificial Intelligence and Images: The Dominant Paradigm in the Representation of Knowledge in Images. Knowledge Organization. 2024; 51: 117–126. https://doi.org/10.5771/0943-7444-2024-2-117

Dobreski B. Common usage as warrant in bibliographic description. Journal of Documentation. 2020; 76: 49–66. https://doi.org/10.1108/JD-05-2019-0094

Giunchiglia F, Dutta B, Maltese V. From knowledge organization to knowledge representation. Knowledge Organization. 2014; 41: 44–56. https://doi.org/10.5771/0943-7444-2014-1-44

Goodrich B, Rao V, Liu PJ, Saleh M. Assessing the factual accuracy of generated text. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 166–175). 2019.

Greenberg J, Zhao X, Monselise M, Grabus S, Boone J. Knowledge organization systems: A network for ai with helping interdisciplinary vocabulary engineering. Cataloging & Classification Quarterly. 2021; 59: 720–739. https://doi.org/10.1080/01639374.2021.1995918

Hartig O, Zhao J. Publishing and consuming provenance metadata on the web of linked data. Provenance and Annotation of Data and Processes. Springer. 2010; 6378: 78–90. https://doi.org/10.1007/978-3-642-17819-1_10

Higgins M. Totally Invisible: Asian American Representation in the Dewey Decimal Classification, 1876-1996. Knowledge Organization. 2016; 43: 609–621. https://doi.org/10.5771/0943-7444-2016-8-609

Homosaurus. "Homosaurus releases.". n.d. Available at: https://homosaurus.org/releases (Accessed: 11 February 2025).

Isaac A. Summers E. SKOS primer (W3C Working Group Note, 18 August 2009). World Wide Web Consortium (W3C). 2009. Available at: https://www.w3.org/TR/2009/NOTE-skos-primer-20090818/ (Accessed: 11 February 2025)

Jiang X, Tian Y, Hua F, Xu C, Wang Y, Guo J. A survey on large language model hallucination via a creativity perspective. arXiv preprint arXiv:2402. 2024; 06647. https://doi.org/10.48550/arXiv.2402.06647

Laban P, Kryściński W, Agarwal D, Fabbri AR, Xiong C, Joty S, et al. Llms as factual reasoners: Insights from existing benchmarks and beyond. arXiv. 2023. https://doi.org/10.48550/arXiv.2305.14540 (preprint)

Lassila O, Hendler J, Berners-Lee T. The semantic web. Scientific American. 2001; 284: 34–43.

Li C, Sugimoto S. Provenance description of metadata application profiles for long-term maintenance of metadata schemas. Journal of Documentation. 2018; 74: 36-61. https://doi.org/10.1108/JD-03-2017-0042

Li M, Zhou H, Yang H, Zhang R. RT: a Retrieving and Chain-of-Thought framework for few-shot medical named entity recognition. Journal of the American Medical Informatics Association. 2024; 31: 1929–1938. https://doi.org/10.1093/jamia/ocae095

Liang W, Zhang Y, Cao H, Wang B, Ding DY, Yang X, et al. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. NEJM AI. 2024; 1: 2400196. https://ai.nejm.org/doi/abs/10.1056/AIoa2400196

Library of Congress. "Process for adding and revising Library of Congress subject headings.". n.d. Available at: https://www.loc.gov/aba/cataloging/subject/lcsh-process.html (Accessed: 11 February 2025).

Littletree S, Metoyer CA. Knowledge organization from an indigenous perspective: The Mashantucket Pequot thesaurus of American Indian terminology project. Cataloging & Classification Quarterly. 2015; 53: 640–657. https://doi.org/10.1080/01639374.2015.1010113

Lodi G, Maccioni A, Scannapieco M, Scanu M, Tosco L. Publishing official classifications in linked open data. In Proceedings of the 2nd International Workshop on Semantic Statistics

(SemStats 2014) at ISWC. 2014; 1–12.

Lucassen T, Schraagen JM. Factual accuracy and trust in information: The role of expertise. Journal of the American Society for Information Science and Technology. 2011; 62: 1232-1242.https://doi.org/10.1002/asi.21545

Ma X, Li J, Zhang M. Chain of thought with explicit evidence reasoning for few-shot relation extraction. arXiv. 2023. https://doi.org/10.48550/arXiv.2311.05922 (preprint)

Markovic M, Edwards P, Corsar D. Sc-prov: A provenance vocabulary for social computation, In Provenance and Annotation of Data and Processes: 5th International Provenance and Annotation Workshop. Cologne, Germany. 2014.

Martínez-Ávila D, Budd JM. Epistemic warrant for categorizational activities and the development of controlled vocabularies. Journal of Documentation. 2017; 73: 700-715.https://doi.org/10.1108/JD-10-2016-0129

McCarthy J, Minsky ML, Rochester N, Shannon CE. A proposal for the dartmouth summer research project on artificial intelligence. AI Magazine. 2006; 27: 12. https://doi.org/10.1609/aimag.v27i4.1904

OCLC. “Dewey Editorial Policy Committee.”. n.d. Available at: https://www.oclc.org/en/dewey/resources/epc.html (Accessed: 11 February 2025).

Patwardhan A, Vaidya V, Kundu A. (2024, October). Automated Consistency Analysis of LLMs, In 2024 IEEE 6th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA). Washington, DC, USA. 2024.

Peeperkorn M, Kouwenhoven T, Brown D, Jordanous A. Is temperature the creativity parameter of large language models?. arXiv preprint arXiv:2405.2024; 00492. https://doi.org/10.48550/arXiv.2405.00492

Qin J. Knowledge Organization and Representation under the AI Lens. Journal of Data and Information Science. 2020; 5: 3–17. https://doi.org/10.2478/jdis-2020-0002

Renze, M. The effect of sampling temperature on problem solving in large language models, In Findings of the Association for Computational Linguistics: EMNLP 2024. Miami, Florida, USA. 2024

Roitblat HL, Kershaw A, Oot P. Document categorization in legal electronic discovery: computer classification vs. manual review. Journal of the American Society for Information Science and Technology. 2010; 61: 70–80. https://doi.org/10.1002/asi.21233

Smiraglia RP, Cai X. Tracking the evolution of clustering, machine learning, automatic indexing and automatic classification in knowledge organization. Knowledge Organization. 2017; 44: 215–233. https://doi.org/10.5771/0943-7444-2017-3-215

Sun K, Yu J, Li J, Hou L. Exploring sequence-to-sequence taxonomy expansion via language model probing. Expert Systems with Applications. 2024; 239: 122321. https://doi.org/10.1016/j.eswa.2023.122321

Tennis JT. The strange case of eugenics: A subject’s ontogeny in a long-lived classification scheme and the question of collocative integrity. Journal of the American Society for Information Science and Technology. 2012; 63: 1350–1359. https://doi.org/10.1002/asi.22686

Thomson C, Reiter E, Sundararajan B. Evaluating factual accuracy in complex data-to-text. Computer Speech & Language. 2023; 80: 101482. https://doi.org/10.1016/j.csl.2023.101482

Turner H, Bruegeman N, Moriarty PJ. Provenance and historical warrants: histories of cataloguing at the Museum of Anthropology. Journal of Documentation. 2024; 80: 1419-1441. https://doi.org/10.1108/JD-02-2024-0037

Wang C, Liu X, Yue Y, Tang X, Zhang T, Jiayang C, et al. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. arXiv. 2023. https://doi.org/10.48550/arXiv.2310.07521 (preprint)

Westin F. Comparing Feature Engineering Techniques for the Time Period Categorisation of Novels. Knowledge Organization. 2024; 51: 330–339. https://doi.org/10.5771/0943-7444-2024-5-330

Zhao Y, Zhang R, Li W, Li L. Assessing and understanding creativity in large language models. Machine Intelligence Research. 2025; 22: 417-436.https://doi.org/10.1007/s11633-025-1546-4

Zhang Y, Ives Z, Roth D. “Who said it, and Why?” Provenance for Natural Language Claims, In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online. 2020.

Zhou R, Chen L, Yu K. Is LLM a reliable reviewer? A comprehensive evaluation of LLM on automatic paper reviewing tasks. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 9340–9351). 2024.