Knowledge Organization

Knowl. Organ. **2026**; 53(6): 50577
https://doi.org/10.31083/KO50577

Review

The Role of KOS in AI-Supported Semantic Integration: Insights From Good Practice Examples

Marcia Lei Zeng[1,*]

[1]School of Information, Kent State University, Kent, OH 44240, USA
*Correspondence: mzeng@kent.edu (Marcia Lei Zeng)

Academic Editor: Natália Tognoli
Submitted: 1 February 2026 Revised: 24 July 2026 Accepted: 5 August 2026 Published: 29 October 2026



Marcia Lei Zeng is a Professor of Information Science at Kent State University (USA). She holds a Ph.D. from the School of Computing and Information at the University of Pittsburgh (USA). Her research interests include knowledge organization systems, metadata, smart data and big data, semantic technologies, and digital humanities, with over 150 research papers and six books published. Her projects have received funding from the NSF, IMLS, OCLC, Fulbright, etc. She is the recipient of the 2024 ASIS&T Award of Merit, the highest honor presented by the Association. She has chaired and served on committees, working groups, and executive boards including IFLA, SLA, ASIS&T, NISO, ISO, DCMI, ISKO, iSchools, and W3C. Currently, she is chairing the DCMI Education Committee, while serving as a member of the DCMI Governing Board and ISKO Board of Directors.

Abstract

Advanced artificial intelligence (AI)-supported semantic integration tools can perform semantic analysis, content categorization, entity and relationship extraction, identity matching, and real-time processing and thus can be highly useful for processing news and formal publications, exploring cultural heritage (CH) resources, or analyzing digitized historical documents. Despite the unparalleled power of AI-based discovery services over vast data sets, robust Knowledge Organization Systems/Services/Structures (KOS) and reliable metadata are crucial for any data service aiming to provide accurate, high-quality, and interoperable information access. This paper analyzes relevant cases and examines two information domains exhibiting these characteristics: (1) machine-readable text-based resources related to health and healthcare and (2) non-machine-readable heterogeneous CH resources. In both contexts, KOS play important and well-defined roles in improving information discovery services based on AI-supported semantic integration, which is the process of interrelating information from diverse sources, enhancing identification quality, and enabling the effective organization and exchange of knowledge across scripts, languages, and cultures.

Keywords: knowledge organization systems; health informatics; cultural heritage resource; AI-supported application; semantic integration

1. Introduction

The past decades have witnessed the rise of advanced information processing technologies, including automatic categorization and indexing based on natural language processing (NLP) and semantic connection with linked data. Artificial intelligence (AI)-supported tools are capable of semantic analysis, content categorization, entity and relationship extraction, identity matching, and real-time processing and have therefore become very helpful in processing news and formal publications, exploring diverse resource types, and analyzing digitized documents. Moreover, these tools enable effective communication supported by structured data, interoperable mind maps and concept maps, semantic reasoning, and path inference, in addition to providing question and answer (Q&A) and content summary services. AI has also gained importance in cultural heritage (CH) exploration, transforming the way we interact with historical artefacts while opening new avenues for research, improving accessibility, and enhancing preservation efforts. However, the implementation of AI-based tools faces challenges when cross-linguistic or cross-cultural processing is involved while establishing data sovereignty becomes even more important (Berners-Lee, 2025; Wang et al., 2026).

Knowledge Organization Systems/Services/Structures (KOS), such as classifications, gazetteers, lexical databases, ontologies, taxonomies, and thesauri, model the underlying semantic structure of a domain (Busch and Zeng, 2025; Hill et al., 1999; Tudhope and Koch, 2004). Created by subject experts as standalone terminology tools and embodied as Web-based services, KOSs[1] ([1]In the present article the plural form KOSs is used when specific knowledge organization systems are referred, following Mazzocchi, 2018, while the references may use KOS in both the singular and the plural sense.) have

Copyright: © 2026 The Author(s). Published by IMR Press.
This is an open access article under the CC BY 4.0 license.

Publisher's Note: IMR Press stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

been designed to support the organization of knowledge and information in order to make their management and retrieval easier, even though they are characterized by different specific structures and functions, varied ways of relating to technology, and are used in a plurality of contexts by diverse communities (Mazzocchi, 2018). KOS can facilitate metadata-based information retrieval and resource discovery, acting as semantic road maps and enabling a common orientation by indexers and future users (human or machine) (Tudhope and Koch, 2004). Over the past two decades, semantic technologies and associated linked data standards have led to an improved and more functional data presentation of traditional KOSs created as non-digital. The new standardized KOS data models and formats can capture semantic relationships and make them processable and more aligned with the principles of formal ontologies (Zeng and Mayr, 2019). These advances have enabled the creation of new KOS vocabularies as native linked data systems ready for implementation in more innovative AI-supported applications.

This paper explores practical examples of both machine-readable text-based resources and non-machine-readable heterogeneous resources with significant outcomes. It discusses the important role of KOS in AI-supported semantic integration, which is the process of interrelating information from diverse sources (especially for eliminating ambiguity), enhancing identification quality, and enabling the effective organization and exchange of knowledge across scripts, languages, and cultures. It showcases KOS applications in two resource discovery scenarios, namely (1) machine-readable textual resources related to health and healthcare and (2) non-machine-readable heterogeneous CH resources.

2. KOS Applications for Machine-Readable Textual Resources

2.1 Medical Text Indexer: From Automated to Next-Generation Health Informatics

Health informatics and biomedical research in general have experienced exponential data and information growth due to the high level of automation and development of computer-supported diagnostics, treatment, and patient care. In healthcare workflows, data and information are exchanged among many parties, including medical professionals, medical technicians, paramedics, clinical laboratory technologists and technicians, government agencies, international organizations, academic researchers, medical center administrators, patients and families, and insurance providers. In the case of a pandemic such as COVID-19 (coronavirus disease 2019), public health laboratories need to rapidly test the new pathogen. The challenge is dramatically exacerbated when reports are communicated across languages, regions, and cultures, especially with statistics generated based on different contextual information and gathered from different sources (Zeng et al., 2020).

In healthcare information systems, data normalization is fundamental for subsequent data analysis, information management, and decision-making. Timely information exchange is essential for health-related situation reports, treatment, and in-depth research, and requires suitable actions. Standardized health KOSs have always played an important role in healthcare information systems, facilitating data normalization (Hong and Zeng, 2022; Patel et al., 2000).

A notable example is MEDLINE, a primary component of PubMed. PubMed is a literature database containing references to articles in life science journals with a focus on biomedicine and is developed and maintained by the United States National Library of Medicine (NLM).[2] ([2]PubMed website: https://pubmed.ncbi.nlm.nih.gov/) In MEDLINE, records are indexed using NLM Medical Subject Headings (MeSH).

MEDLINE is remarkably effective.[3] ([3]See "Frequently Asked Questions about Indexing for MEDLINE" https://www.nlm.nih.gov/bsd/indexfaq.html) In 2021, the average time for a complete article to be indexed by humans was 145 days, whereas in 2025, automatically indexed article citations were generally completed within one day of receipt and appeared as indexed for MEDLINE in PubMed the following day. Notably, these values do not account for the time required for bibliographic data review. Given that roughly one-third of automatically indexed articles currently receive human input, human indexers continue to play an important role in quality assurance.

- In 2019, MEDLINE began testing the full automation of indexing (with human review of selected citations) for a smaller number of journals. MeSH indexing for MEDLINE was completely performed by human indexers up to 2011.
- Medical Text Indexer – Automated (MTIA) is a complex system based on a dictionary of MeSH terms, synonyms, and other trigger phrases, with rules created and refined by humans over many years.
- In April 2022, all journals indexed for MEDLINE were indexed automatically, with human review and editing as appropriate.
- MTIA was finely tuned to favor precision over recall when evaluated against human indexing.
- In April 2024, MEDLINE released Medical Text Indexer – NeXt Generation (MTIX) algorithm (National Library of Medicine, 2024).[4] ([4]https://www.nlm.nih.gov/pubs/techbull/ma24/ma24_mtix.html)
- MTIX was trained on millions of MEDLINE entries published between 2007 and 2022 and uses a machine learning model known as a neural network (a type of AI). Once trained, MTIX can apply the acquired knowledge to new citations, determining MeSH terms that are statistically likely to be appropriate for their indexing.
- Compared with MTIA, MTIX maintains a similar level of precision but makes large gains in recall,

correctly applying 50% more terms and thus achieving markedly more comprehensive indexing (National Library of Medicine, 2024).

The important role of MeSH is obvious. For example, after the World Health Organization (WHO) officially named the pandemic-responsible disease as COVID-19 on February 11, 2020 (World Health Organization, 2020), MeSH rapidly provided a Supplementary Concept Record[5] ([5]Supplementary Concept Records (SCRs) are terms in a separate thesaurus from MeSH and are updated daily. https://www.nlm.nih.gov/oet/ed/pubmed/mesh/mod01/03-400.html) featuring "COVID-19" as the preferred lexical label with a unique ID and other entry terms to facilitate subject access to the related information resources.[6] ([6]https://meshb.nlm.nih.gov/record/ui?ui=D000086382) As observed for "Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2)",[7] ([7]https://meshb.nlm.nih.gov/record/ui?ui=D000086402) in which case 10 entry terms, including those with place names, are available, MeSH follows WHO's Best Practices for Naming of New Human Infectious Diseases, ensuring that the name does not refer to a geographical location, an animal, or an individual or group of people while being pronounceable and related to the disease (World Health Organization, 2015).

2.2 Theoretical and Practical Advances in Knowledge Representation

The innovation brought by MTIX is its use of neural networks, which prompts us to analyze and understand the theoretical and application advances in knowledge representation in this subsection.

Neural networks are complex systems of interconnected nodes that pass information through multiple layers to find connections and extract meaning from data (International Organization for Standardization, 2024). The AI world had long been divided into two: the "logic" camp and the "neural net" camp, in which the "neural nets" is mimicked the structure of biological brains (Berners-Lee, 2025, Ch.16). Neural networks, including traditional Artificial Neural Networks (ANNs) and more advanced Spiking Neural Networks (SNNs), form a part of machine learning that provides a basis for deep learning and are inspired by and mimic the human brain and communication between neurons (Qamar and Zardari, 2023). AI tools of this kind, particularly convolutional neural networks, are transforming healthcare by enhancing predictive, diagnostic, and decision-making capabilities (Kourounis et al., 2023).

As a neural network, MEDLINE's MTIX outperforms MTIA by understanding more complex concept representations. As MTIX makes determinations based on many features and not just trigger words, it can recognize abstract ideas not literally stated in the text, predict (based on the abstract) some MeSH concepts present in the full text, and avoid contextual errors when encountering metaphors and idioms. An example given by MEDLINE is the avoidance of the index term "elephants" if the text being indexed contains an idiom "the elephant in the room".[8] ([8]See "MTIX Performance" section in: https://www.nlm.nih.gov/pubs/techbull/ma24/ma24_mtix.html)

Yan et al. (2022) reported similar progress in indexing automation based on their survey of automated International Classification of Diseases (ICD) coding and observed three development stages: Stage 1, Rule-based; Stage 2, Traditional Machine Learning-based; and Stage 3, Neural Network-based.

The emergence of large language models (LLMs) has opened new possibilities and reinvigorated enthusiasm for AI-based solutions among indexing experts but prompted an invitation for caution: "LLM reasoning methods, which are based on finding useful data in a vast repository, produce a kind of abduction or educated guessing. But abduction alone cannot determine the accuracy of the guesses. To evaluate the guesses, some generative AI systems are hybrids that also use symbolic AI methods for detecting and avoiding errors" (Sowa and Majumdar, 2025). These insights carry special weight as they come from John Sowa, the inventor of conceptual graphs and computer design in the digital age.[9] ([9]https://en.wikipedia.org/wiki/John_F._Sowa) His book Knowledge Representation: Logical, Philosophical and Computational Foundations (Sowa, 1999) introduces knowledge representation as a multidisciplinary subject that applies theories and techniques from three other fields.

(1) Logic provides the formal structure and rules of inference.

(2) Ontology defines the kinds of things that exist in the application domain.

(3) Computation supports the applications that distinguish knowledge representation from pure philosophy.

A very important and timely webinar, "Evaluating and Reasoning with and about GPT", took place at the 2023 Ontology Summit. This webinar was given by John Sowa and Aran Majumdar, who have both taken advantage of the full range of technologies enabled by 60+ years of AI, computer science, and computational linguistics development. By demonstrating Permion OSIRIS – Open Source Intelligence and Relationship Inference System (Sowa and Majumdar, 2023), the webinar provided an opportunity to learn about the strengths and limitations of generative pre-trained transformer (GPT) and LLMs using a real application. The following discussion provided explanations and demonstrated features showing the ways humans achieve cognition from perception.

■ To achieve high precision, Permion uses an integrated combination of the best generative and symbolic methods. Reasoning is performed in a cognitive cycle of abduction, deduction, testing, induction, and repetition, using common logic with extensions for metalanguage, modality, and higher-order logic.

■ To handle vagueness and approximations, Permion uses neural networks and mathematical statistics and probability methods. The cognitive cycle begins with a basic ontology, which is dynamically developed, revised, and extended.

■ In addition to normal text analysis, the analysis of a document can be observed immediately in multiple outcomes provided by Permion, with the sections of Scaffolding Net, Proto-Ontology, Path Inference, and Summary (including Abstract, Keywords, Question and Answer, Synsemantic Words, and Signals).[10] ([10]Watch the demo at https://www.youtube.com/watch?v=6K6F_zsQ264)

Since memory is a fundamental AI system component underpinning LLM-based agents, Du et al. (2025) developed a unified framework for memory taxonomy, operations, and applications in AI systems and categorized memory representations as parametric and contextual. Contextual structured memory is an explicit memory organized into predefined interpretable formats or schemata, e.g., knowledge graphs, relational tables, or ontologies. The above-mentioned research introduced six fundamental memory operations, namely consolidation, updating, indexing, forgetting, retrieval, and compression, and mapped them to the most relevant research topics across long-term, long-context, parametric modification, and multi-source memory. This work also provided a structured and dynamic perspective on research, benchmark datasets, and tools related to memory in AI, clarifying the functional interaction in LLM-based agents while outlining promising directions for future research.

2.3 Progress in KOS Usage in Healthcare

Health information is at the forefront of KOS development, embracing linked data standards to the navigation of vocabularies such as MeSH and Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT). The advanced systems and services listed below demonstrate the transition of KOSs from traditional indexing resources to machine-processable tools.

2.3.1 Embracing Digital Technologies and Unifying Health KOSs

Widely used and standard health KOSs are required to support essential health services and closely follow developments in health issue treatment in the rapidly advancing digital age. Since a dramatic revision may require content, structure, coding system, and use guideline adjustments, such KOSs should ensure semantic interoperability, data reusability, and community agreement.

International Classification of Diseases (ICD)[11] ([11]https://www.who.int/standards/classifications/classification-of-diseases) has been representing innovative features in its revisions for more than 100 years. The most recent 11th revision of ICD (ICD-11) has been in the process of implementation since 2022, almost three decades after the release of the 10th revision (ICD-10). The former revision is exceptional in the way it utilizes and combines digital technologies and KOS approaches typically used for classifications, thesauri, ontologies, etc. Several advantageous features of the new revision include a framework that combines classification and terminology (Fig. 1), provides concept Uniform Resource Identifiers (URIs), and is delivered with an implementation package (Zeng and Hong, 2022). The case of ICD-11 shows that KOS vocabularies can successfully and creatively utilize digital technologies with multiple KOS management approaches and are desirable in all domains, including health information.

To determine whether an entity qualifies to become a unique category/class, ICD-11 applies a set of measures specified within a smart coding algorithm capable of interpreting more than 1.6 million terms.[12] ([12]https://www.who.int/news/item/11-02-2022-icd-11-2022-release) The provision of categories grouping more specific terms helps manage multiple access points and identify specific diseases, which enables the statistical presentation of data for broader groups and the attainment of useful and understandable information (World Health Organization, 2022, Section 1.2.1). Thus, ICD-11 is entirely digital and holds more than 120,000 codable terms aligned with ~17,000 unique classification codes.

For example, a search for COVID-19-related entries returns a significant number of deprecated terms with country names originating from media news, which lead to formal classification entries (Fig. 1). This reflects the World Health Organization (2021) policy for labeling the key variants of the virus causing COVID-19. In May 2021, these labels used letters of the Greek alphabet rather than the names of places where these variants were detected to avoid discrimination and semantic conflicts.

2.3.2 Enhancing Interoperability

Owing to the realities of everyday healthcare, various KOS standards are needed for different processes. Fig. 2 illustrates the usage of these standards in electronic health records and provides those most commonly used in health information exchange.

As shown in Fig. 2, ICD and SNOMED CT are mainly used to normalize diagnostic processes and health problem lists. Other standards have been developed to encode medical lab tests and clinical observations [Logical Observation Identifier Names and Codes (LOINC)], medical procedures [Current Procedural Terminology (CPT)], and standardized nomenclature for clinical drugs (RxNorm), playing a crucial role in healthcare information systems by supporting data normalization and semantic linking/sharing.

The KOSs listed above are based on practices in the USA. Nealy 200 biomedical KOS vocabularies from around the world are included in the Unified Medical Language System (UMLS) Metathesaurus—a large, multi-

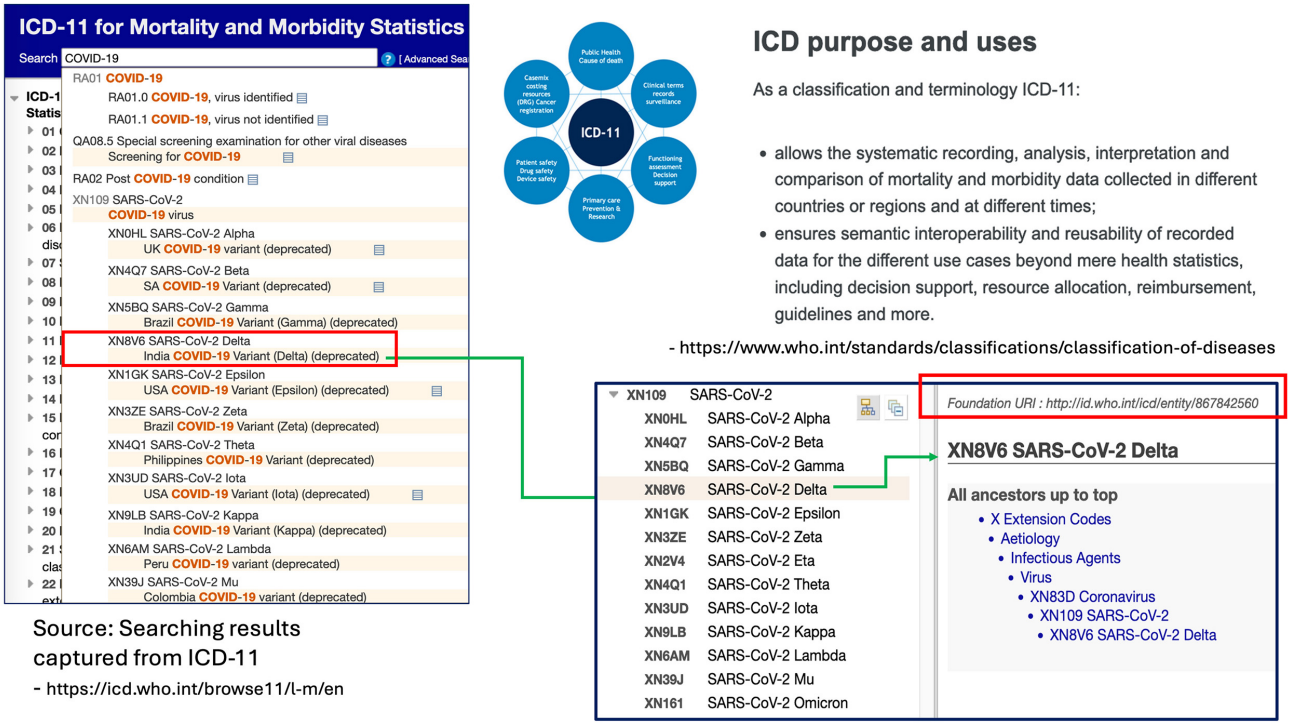


Fig. 1. Search results obtained using ICD-11. A search for COVID-19-related entries returns a significant number of deprecated terms (left), which lead to formal classification entries (right). Images captured and marked based on the results of a search through ICD-11 Browser https://icd.who.int/browse11/l-m/en, provided by the World Health Organization (WHO) for International Classification of Diseases (ICD) https://www.who.int/standards/classifications/classification-of-diseases. SARS-CoV-2, Severe Acute Respiratory Syndrome Coronavirus 2.

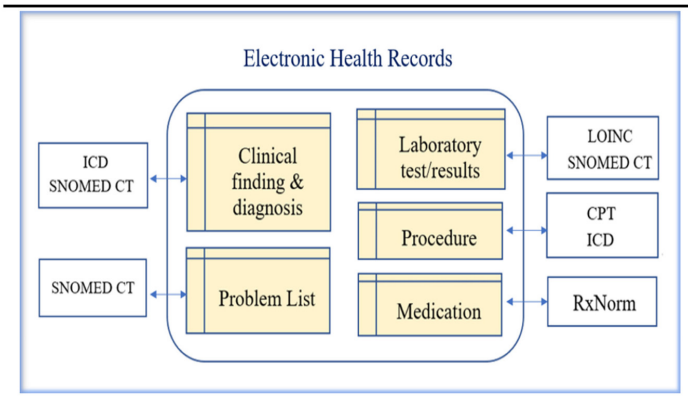
Use of KOSs in electronic health records	Commonly used KOS standards in health information exchange
	**ICD** *International Classification of Diseases* | World Health Organization **SNOMED CT** *Systematized Nomenclature of Medicine--Clinical Terms* | International Health Terminology Standards Development Organization **LOINC** *Logical Observation Identifiers Names and Codes* | LOINC.org **CPT** *Current Procedural Terminology* | American Medical Association **RxNorm**, a normalized naming system for generic and branded drugs | United States National Library of Medicine

Fig. 2. Main uses of different standardized health KOSs in electronic health records according to practices in the USA. Image (left) source: Hong and Zeng, 2022. KOS, Knowledge Organization Systems/Services/Structures.

purpose, and multilingual biomedical vocabulary database hosted by NLM.[13] ([13]https://uts.nlm.nih.gov/uts/umls/home)

■ The UMLS[14] ([14]https://www.nlm.nih.gov/research/umls/index.html) integrates and distributes key terminology, classification, and coding standards, and associated resources to promote the creation of more effective and in-

teroperable biomedical information systems and services, including electronic health records.

■ The UMLS Metathesaurus also identifies useful relationships between concepts and preserves the meanings, concept names, and relationships from each vocabulary. The 2025 Metathesaurus contains ~3.49 million concepts and 17.39 million unique concept names from 190 source vocabularies in 29 languages.

BioPortal[15] ([15]https://bioportal.bioontology.org/) is the world's most comprehensive repository of biomedical ontologies, providing numerous KOSs in addition to those listed in Fig. 2. The high number of ontologies (1260 as of January 2026) indicates the need for standardized KOS vocabularies in diverse circumjacencies and communities.

The Artificial Intelligence Ontology[16] ([16]https://bioportal.bioontology.org/ontologies/AIO) released on BioPortal in 2024 aims to address the rapidly evolving AI landscape by providing a comprehensive framework that encompasses the technical and ethical aspects of AI technologies (Joachimiak et al., 2024).

■ The ontology is structured around multiple top-level branches, including networks, layers, functions, LLMs, preprocessing, and bias, and is designed to support the modular composition of AI methods and facilitate a deeper understanding of deep learning architectures and ethical considerations in AI.

■ Twelve language model classes are present under "Language Model", including one "Large Language Model" class which has nearly 50 subclasses.[17] ([17]While the terms may overlap in usage, each refers to a specific technology or application.)

■ Over 100 sub-classes are present at multiple levels under the "Bias" main class.

Both Metathesaurus and BioPortal enable cross-vocabulary mapping that facilitates knowledge and semantically interoperable data throughout knowledge sources developed independently to meet diverse needs.

OntoPortal Alliance[18] ([18]https://ontoportal.org/) is a rapidly growing consortium that includes more than ten special portals in addition to BioPortal and features various domains and countries/languages.[19] ([19]See more reported in Networked Knowledge Organization Systems (NKOS) 2023 Workshop: https://nkos.dublincore.org/2023NKOSworkshop/NKOS2023.html). For example, SIFR BioPortal and MedPortal are repositories for French and Chinese biomedical terminologies and ontologies, respectively. Techno Portal—a repository for the technology science domain—encompasses nearly 30 ontologies, e.g., Software Ontology, Qualitative Research Ontology, Failure Mode and Effect Analysis, and Anomaly Ontology. The primary mission of OntoPortal installation is hosting and serving ontologies and semantic artefacts.

Schema.org, a collaborative project of major search engine owners, aims to provide shared vocabulary formats (schemata) under one namespace to help webmasters describe and make websites easier to understand and index by search engines.

■ In early 2020, Schema.org rapidly responded to the COVID-19 pandemic by releasing versions 7.0 through 7.04 within a month. The new versions addressed medical and health-related properties as well as the urgent need for structuring public and official information regarding COVID-19.[20] ([20]Schema.org Releases, from section 7.0, 2020-03-17 to 7.04, 2020-04-16: https://schema.org/docs/releases.html)

■ To support AI applications and help address situations regarding the responsible use of generative AI, Schema.org added a digital source type vocabulary based on International Press Telecommunications Council (IPTC) definitions in early 2024. This consists of a property digitalSourceType, an enumeration type IPTCDigitalSourceEnumeration, and a set of 14 coded values such as TrainedAlgorithmicMediaDigitalSource mapping directly to the IPTC list.[21] ([21]Schema.org Releases, section 24.0, 2024-01-09: https://schema.org/docs/releases.html)

3. KOS Applications for Non-Machine-Readable and Heterogeneous Resources

The global information resource environment is rich in nondigital and digital nontextual resources, e.g., musical recordings, images, objects or documents, and datasets in various formats hosted on different types of carriers. Even when textual and digitized, some of these resources do not easily lend themselves to standard machine processing because of their language, scripts, or historical and cultural determinants. Organization, discovery, and access to these resource collections have been traditionally managed through descriptive metadata. Such heterogeneous and diverse information resources are prevalent in the CH domain and largely found in libraries, archives, and museums (LAMs) and cultural resource repositories. The challenges and promises of using AI-assisted automatic processing in this context and the potential role of KOS in improving the quality of AI applications are very important and necessitate better understanding.

3.1 Understanding Diverse CH Resources

Memory institutions and information services are increasingly embracing new information technologies to meet the demands of the changing digital age. In the 21st century, many digital collections were born in LAMs, while numerous digitization projects have successfully embraced emerging semantic technologies and contributed to the realization of the Semantic Web idea. The corresponding data are abundant and of different types and quality, as the information-bearing objects (textual or nontextual, digitized or nondigitized) have diverse formats. The page from Turning the Pages Online[22] ([22]Archive page: https://wayback.archive-it.org/7867/20190220143708/ht



Fig. 3. Page from Turning the Pages Online for Edwin Smith's Surgical Papyrus. Image captured in 2018. Source: NLM.

tps://ceb.nlm.nih.gov/proj/ttp/smith_home.html) in Fig. 3 shows an output from an NLM project from the 2010s and is part of Surgical Papyrus, the oldest surviving surgical text written in Egyptian hieratic script around the 17th century BCE and discovered by the American archaeologist Edwin Smith in Egypt in the 1860s.[23] ([23]https://wayback.archive-it.org/7867/20190220142516/https:/ceb.nlm.nih.gov/proj/ttp/books.htm) Fig. 3 illustrates the level of complexity and challenges presented by this type of CH resource.

The digitization of such textual resources includes data formatting, structuring (datafication), and information contextualization. Further innovations are driven by the challenges in data management, data discovery, and the need to share/inspire research and showcase cultural wealth. Treating data only as files for preservation purposes can limit the impact of new research ideas and findings. Embracing AI is now considered necessary for LAMs to maintain their relevance and transform the way humanity interacts with the past.

Selected CH projects where diverse resources are processed using AI-enabled applications are introduced below and listed in Appendix A. Some of the notable efforts listed here are explored in the next section with a focus on knowledge organization (KO) approaches.

Digitized textual resources may be very diverse (Stokes et al., 2021), e.g., documents can be written in different languages and scripts and be read from left to right, right to left, top to bottom, or bottom to top. Similarly, the document may have a carrier based on different materials (e.g., paper, parchment, stone, clay, and wood) and be written using different scribing tools (e.g., pen, brush, needle, and chisel). Furthermore, handwritten annotations, especially in rare and old printed books, have to be preserved during digitization.[24] ([24]https://training.iiif.io/iiif-5-day-workshop/day-four/iiif-and-ai.html) Some exceptional and challenging examples of textual resources are presented below, and methodologies on which these services are based are discussed in subsequent sections.

- The 20th Century Press Archives (PM20)[25] ([25]https://pm20.zbw.eu/about.en.html) is a large digitized newspaper archive covering 1908–2005, containing materials published in Germany and worldwide, and providing millions of press clippings and other materials regarding persons, or-

ganizations, wares, events, and general subjects extracted from more than 1500 sources.

■ eScriptorium[26] ([26]https://escriptorium.inria.fr/) is an advanced service for the processing of printed documents and handwritten texts using machine learning techniques. This service is available as a web application offering a workspace to manage the various steps of transcription tasks.

■ Wooden Slip Character Dictionary (WCD)[27] ([27]https://wcd-ihp.ascdc.sinica.edu.tw/woodslip/index.php) is a database of Han Dynasty (206 BCE–220 CE) wooden slips from Juyan, providing access to and full digital interpretation of characters written on ancient Chinese wooden slips. WCD is a part of a union catalog for historical Chinese characters on wooden slips and other written materials from East Asia and around the world.[28] ([28]Historical text database integrated retrieval system https://wcd-ihp.ascdc.sinica.edu.tw/union/)

■ KuroNet character recognition (AI OCR)[29] ([29]https://codh.rois.ac.jp/kuronet/) is an open-source AI-powered platform for recognizing the characters of the premodern Japanese Kuzushiji script, which can be fluently read by only several thousand people (~0.01% of the Japanese population).[30] ([30]https://codh.rois.ac.jp/miwo/index.html.en)

Many kinds of information-bearing CH entities, once digitized, remain nontextual and non-machine-readable and present challenges in data processing. Compared with text-based unstructured data and two-dimensional materials, three-dimensional objects pose greater challenges. Several AI-assisted innovative projects involving the processing of symbols, artwork images, sounds, scents, museum objects, 3D+, rock art, subtractive notations, ancient coins, etc. have been reported. These advances make applications for promoting CH more abundant and practical. Readers are advised to refer to Appendix A "Diverse Cultural Heritage Entities Explored in the New Age of AI".

3.2 Digitization → Datafication → Contextualization Enabled by KOS

The effective processing and diverse nature of the CH resources described above and listed in Appendix A illustrate the speed with which new advanced tools are made available for digitizing and datafying, i.e., rendering digitized objects machine-readable. However, this progress is accompanied by concerns regarding the vastness, vagueness, uncertainty, inconsistency, and ambiguity of the resources to be handled. Beyond digitization and datafication, information resource contextualization is even more challenging given that the materials in question are usually special, involving image-based, media-based, and other non-written-language representations.

Taking place-related situations as an example, many possible issues arise from the outset (Zeng and Mayr, 2019).

■ Place names change over time (e.g., "Bombay" and "Mumbai")

■ Alternative names (e.g., "New York City", "NYC", and multilingual labels)

■ Same names associated with multiple locations (e.g., various "St. Petersburg")

■ Unidentifiable places (e.g., places referred to in a creative work but not identified)

■ Unnamed places

■ Cartographic versus geographical placement

■ Incorrect or inconsistent feature typing/categorizing results

A KOS vocabulary such as a name authority file, a terminology list, a taxonomy or classification scheme, a thesaurus, or an ontology is always needed to ensure semantic consistency. The use of KOSs, in the above-mentioned cases, eliminates ambiguity, enhances identification quality, and enables the effective connection of CH entities across scripts, languages, and cultures.

3.2.1 Applying KOS Methods for Special Resources

Below, the methods and approaches used in KOS applications are discussed in more detail with selected advanced projects as examples.

The 20th Century Press Archives (PM20)[31] ([31]https://pm20.zbw.eu/about.en.html) is managed by ZBW – Leibniz Information Centre for Economics (formerly German National Library of Economics), which published the STW Thesaurus for Economics (STW), the world's most comprehensive bilingual thesaurus for representing and searching for economics-related content,[32] ([32]https://www.zbw.eu/en/about-us/information-organisation/stw-thesaurus-for-economics) and continuously updates it to reflect the latest changes in economics terminology. After this large newspaper archive was digitized in 2019–2021, PM20 donated data to Wikidata. The data contains 5266 person folders and 9200 folders of country/-subject archive metadata with digitized documents linked to Wikidata.[33] ([33]https://www.wikidata.org/wiki/Wikidata:WikiProject_20th_Century_Press_Archives) The efforts continued and also contributed to Wikidata with its Press Archives. In 2025, ZBW published an introduction to the STW, together with its editorial guidelines, to present the editorial processes behind the STW in a comprehensible manner and increase transparency for the specialist community and cooperation partners in the increasingly dynamic environment of digitalization, AI, and open science. As stated by its 2025 annual review, "We do not merely organize information. We create the conditions for reliable, transparent, and increasingly digital science" (ZBW, 2025).

Wikidata is a complementary tool for authority control in the Semantic Web stage (Bianchini et al., 2021). Using the user-friendly query templates provided by PM20, large amounts of data can be obtained

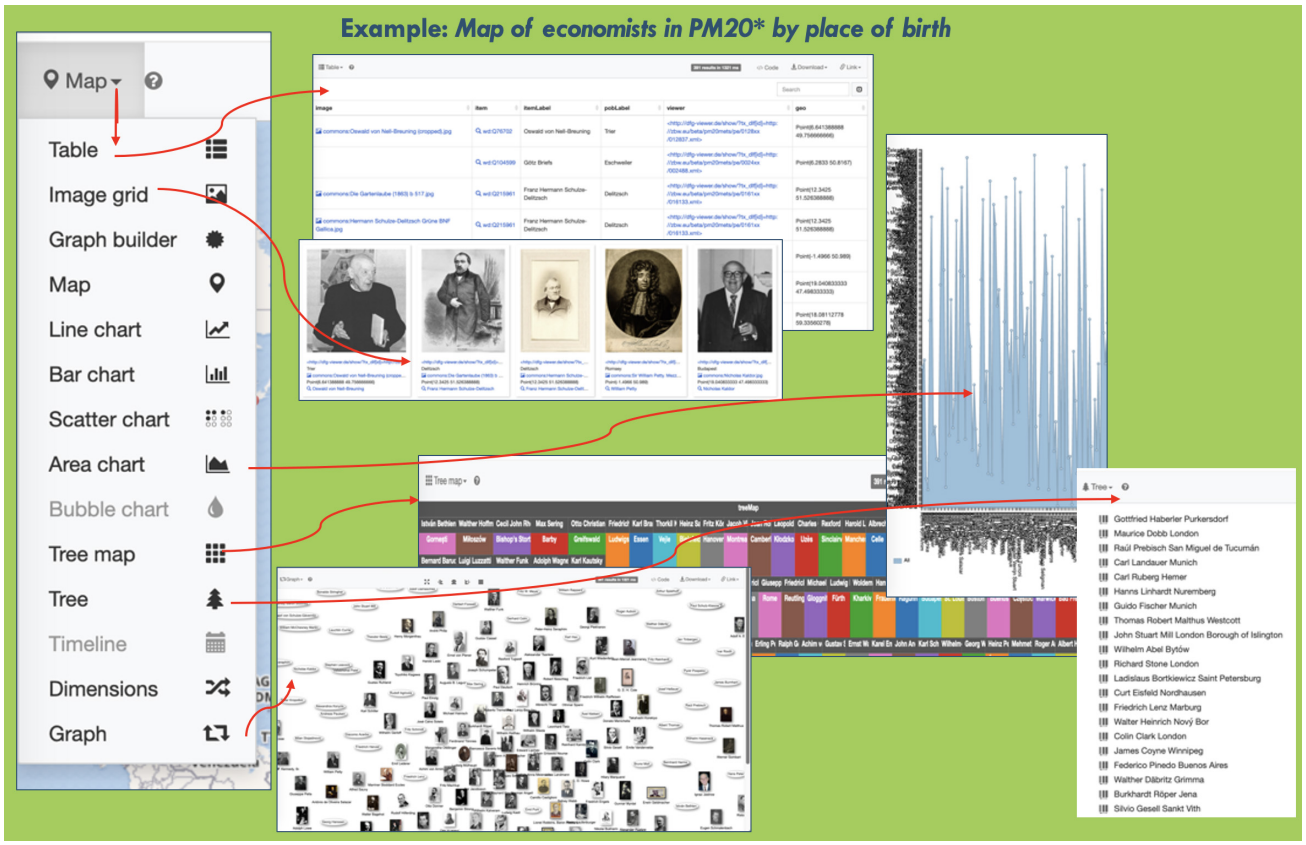


Fig. 4. Examples of data collected using the templates provided by PM20 and operated on the Wikidata query service. Images captured and marked based on the search results provided by Wikidata. *The 20th Century Press Archives (PM20).

with a variety of presentations provided. Taking an example for Map of economists in PM20 by place of birth,[34] ([34]https://www.wikidata.org/wiki/Wikidata:WikiProject_20th_Century_Press_Archives/Use_cases) Wikidata provides multiple relevant datasets, presented as tables, image grids, graph builders, maps, line charts, bar charts, scatter charts, area charts, tree maps, trees, dimensions, and graphs[35] ([35]https://query.wikidata.org/) (Fig. 4).

eScriptorium[36] ([36]https://escriptorium.inria.fr/) offers a workspace for managing the various steps of transcription operations, hosted and led by a team at Université Paris Sciences et Lettres, and provides a digital text production pipeline for print and handwritten texts using machine learning techniques. The "Segment with eScriptorium" part contains a special instruction on "Using the Ontology feature and labeling lines and regions", which allows users to define the types of lines and regions and associate them when processing manuscripts. Readers are advised to refer to the website and Stokes et al. (2021).

Wooden slips are the traditional writing medium of the Early imperial China, especially the Han Dynasty period (206 BCE–220 CE). The Wooden Slip Character Dictionary (WCD) project[37] ([37]https://wcd-ihp.ascdc.sinica.edu.tw/woodslip/index.php) focuses on imaged texts on ancient wooden slips and explores the use of ontologies and the International Image Interoperability Framework (IIIF) from the perspective of interoperability. Studies on text interpretation and slip restoration (Juyan Han Wooden Slips project) integrated the demands of wooden slip research and discerned 10 functional primitives for these scholarly activities. Wooden Slip Research ontologies composed of interrelated Wooden Slips, Chinese Characters, and Annotation ontologies were proposed to support digital wooden slip research and enable the cross-institutional search for Chinese characters (Fig. 5) (Chen, 2021; Chen and Lu-Yen, 2022).

AI Explorer[38] ([38]https://ai.harvardartmuseums.org/) is one of the most interesting LAM-related projects, assisted by multiple AI tools (five computer vision services and nine LLMs, as of July 2026).[39] ([39]https://ai.harvardartmuseums.org/statistics) This project harnesses AI and has gathered over 84,969,000 machine-generated descriptions and tags for more than 370,600 images from across the collections hosted by and originally documented at the Harvard Art Museums. This approach aims to enrich the ways the collection is categorized and described—beyond what the curators can provide on their own—and build and study a dataset revealing how commercial AI services operate.[40] ([40]https://ai.harvardartmuseums.org/about) Annotation types generally describe what the annotation contains or depicts, including the analyzed results for the category (a broad categorization of image contents) and tags



Fig. 5. Examples of Juyan Han Wooden Slips contained in the project. Image source: Slide #3, Chen and Lu-Yen, 2022.

(a term or set of terms describing all or part of the image), in addition to description, face, and text identification. As stated by the project, "Named entities are extracted using Compromise, a JavaScript NLP library. It's not exact, but pretty close."[41] ([41]https://ai.harvardartmuseums.org/statistics)

The provision of category search enables the exploration of auto-categorized images in 12 unique groups for thousands of images, allowing for an effective service with impressive outcomes. However, the question of how a machine performs categorization remains unanswered. Some category search results contain wrongly categorized images, e.g., various images of humans are categorized as pets by AI tools.[42] ([42]https://ai.harvardartmuseums.org/category/Pets%20animals) This situation highlights the important issues addressed by KOSs, as discussed in the next section.

3.2.2 Enhancing KOS Functions to Deepen CH Resource Connections

3.2.2.1 Incorporation of Representative Visual Resources in a Thesaurus. Representative visual resources are visual aids (images, thumbnails, and illustrations) used in descriptive metadata and KOSs to stand for, explain, or contextualize concepts or entities and can also be used to embody complex concepts, objects, or datasets. This solution is, in particular, relevant for CH resources because of their temporal, cultural, and linguistic determinants. The representative visual resources in a thesaurus and other KOS vocabularies can bring other meaningful components for enhancing the semantic consistency, communication, and FAIRness[43] ([43]https://www.go-fair.org/fair-principles/) of open CH resources in cross-cultural, cross-language, and cross-time period situations.

Existing KOS vocabularies for CH are enhancing the functions of visual aid approaches to tap their full potential. For instance, certain vocabulary records in the Art and Architecture Thesaurus (AAT) have been linked to images in Getty collections and elsewhere since 2010. For the concept rhyta (drinking vessels),[44] ([44]http://vocab.getty.edu/page/aat/300198841) over 20 terms in different cultures, languages, and time periods are provided and marked with speech flags, which indicate the category into which the term would be placed relative to its normal function in a grammatical context. Additionally, six representative images are provided (Fig. 6).

3.2.2.2 Using a Thesaurus to Connect Digitized CH Resources. The growing diversity of open CH resources provided by LAMs on the Web necessitates the use of resources beyond alphanumeric to enhance semantic consistency in a globally diverse community.

Chinese Iconography Thesaurus,[45] ([45]https://chineseiconography.org/) initiated in the Victoria and Albert Museum in the United Kingdom and first released in 2019, is freely accessible and brings together sinology, art history, and information studies to create the first alternative classification scheme. This thesaurus is designed with a focus on Chinese visual culture and serves as a comple-

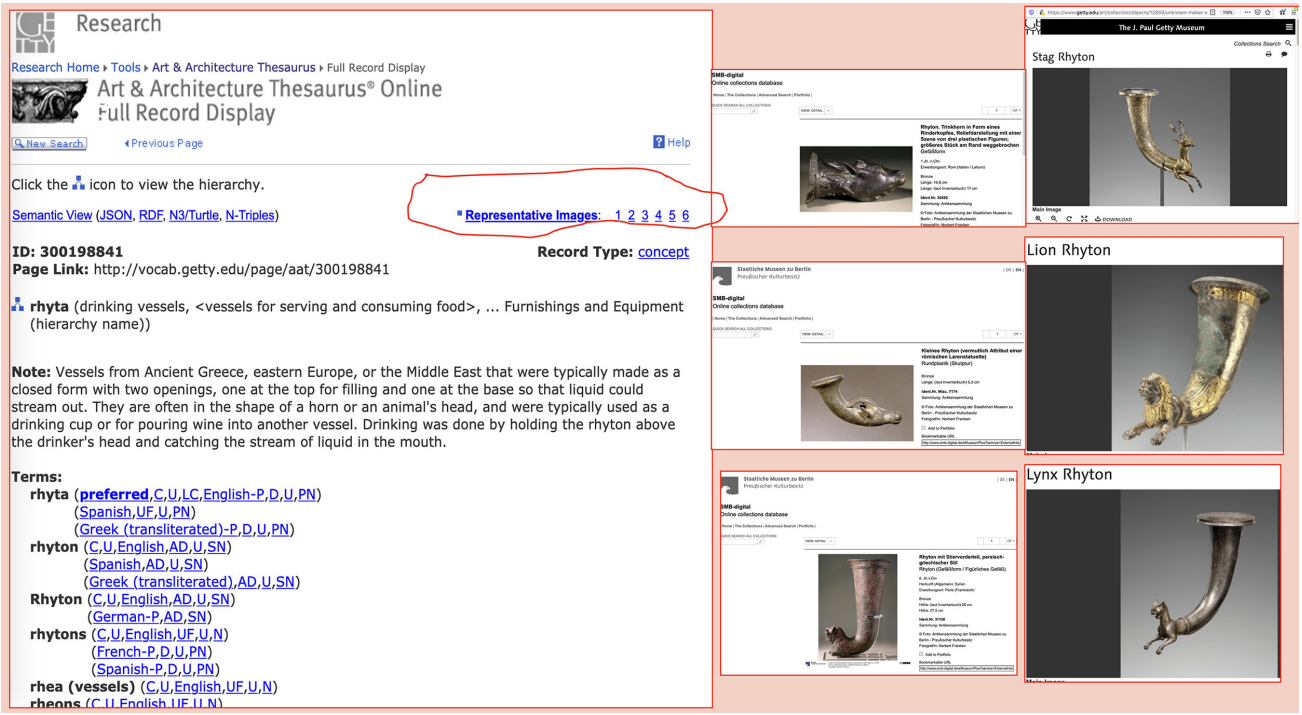


Fig. 6. Representative images from external collections provided for an entity in AAT. Image source: Slide 3 in Zeng (2024). This figure, with a human-readable full record for rhyta (http://vocab.getty.edu/page/aat/300198841), contains information from the J. Paul Getty Trust, Getty Research Institute, and AAT, which is made available under the Open Data Commons Attribution License (ODC-By). AAT, Art and Architecture Thesaurus.

ment to the ICONCLASS[46] ([46]https://iconclass.org/) system for Western Art. The material forms include paintings, sculptures, prints, book illustrations, ceramics, and other types of applied arts produced in China from 700 to 1900.[47] ([47]https://www.vam.ac.uk/research/projects/chinese-iconography-thesaurus-cit) The system is available in a user-friendly online version and as a Simple Knowledge Organization Systems (SKOS) Linked Open Data backed by a SPARQL[48] ([48]SPARQL, pronounced "sparkle", a recursive acronym for SPARQL Protocol and Resource Description Framework (RDF) Query Language, is the standard query language and protocol for Linked Open Data and RDF databases, made by the RDF Data Access Working Group (DAWG) of the World Wide Web Consortium. https://en.wikipedia.org/wiki/SPARQL) endpoint and a complementary IIIF-based image archive, illustrating the application of a classification system. As of July 2026, 17,419 terms are included, encompassing 12,893 artworks from five well-known museums in Europe, America, and Asia. Readers are advised to consult Figs. 7,8 and refer to slides 61–65 in Zeng (2025).

3.2.3 Advanced KOS Vocabulary Services

The abovementioned case of Chinese Iconography Thesaurus illustrates the level of expertise and amount of time required to create a high-quality KOS. Then, how could any special project start making its own KOS? The following section provides examples of KOS-based and aggregate (registry/repository) services that support the complete KOS lifecycle, including vocabulary creation, transformation, mediation, migration, exchange, integration, and aggregation.

Typically, KOS-based services can support indexing, searching, and the exploration of an extended vocabulary and act as a knowledge resource in their own right. For any KOS, a good vocabulary-based service allows its components to connect with external real resources while ensuring that associated datasets are semantically linked.

The advancement of semantic technologies and associated standards enables the improvement of KOS functionalities and underlying tools supporting vocabulary creation, management, and sharing. The most important advantage of semantic technology is that it allows KOSs shared in this environment to be accessed, queried, and processed directly by computer programs, making every user able to generate controlled vocabularies suiting specific requirements (Zeng and Mayr, 2019). For instance, some advanced KOS services enable the creation of special microthesauri in different domains, languages, translations, etc., and other contextualization-needed value vocabularies. Value vocabularies such as controlled lists of terms, concepts, or codes used to populate specific metadata fields are important for ensuring metadata consistency and findability. Advanced services can promptly provide the most urgently needed vo-

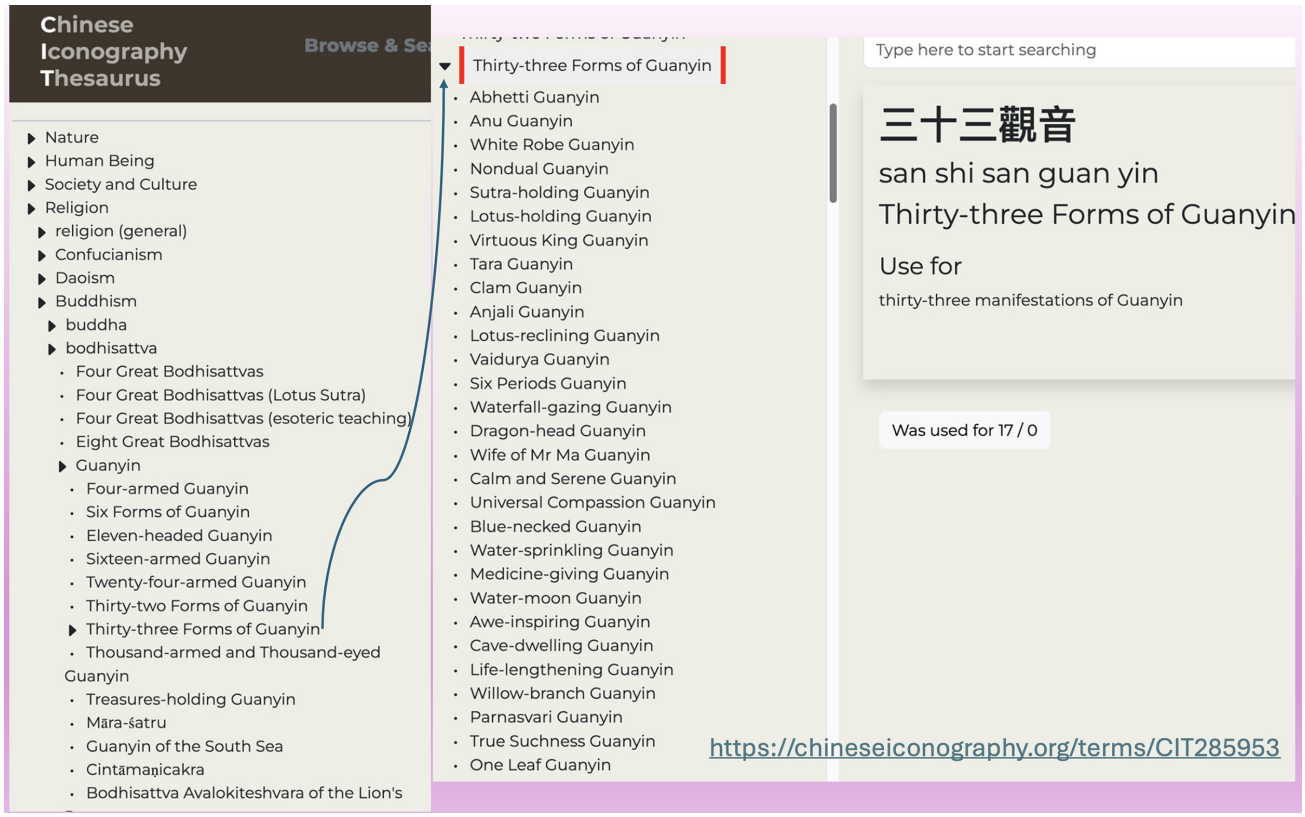


Fig. 7. A special exploration of the Chinese Iconography Thesaurus. © Victoria and Albert Museum, London. Images captured and marked 2025-10-8.

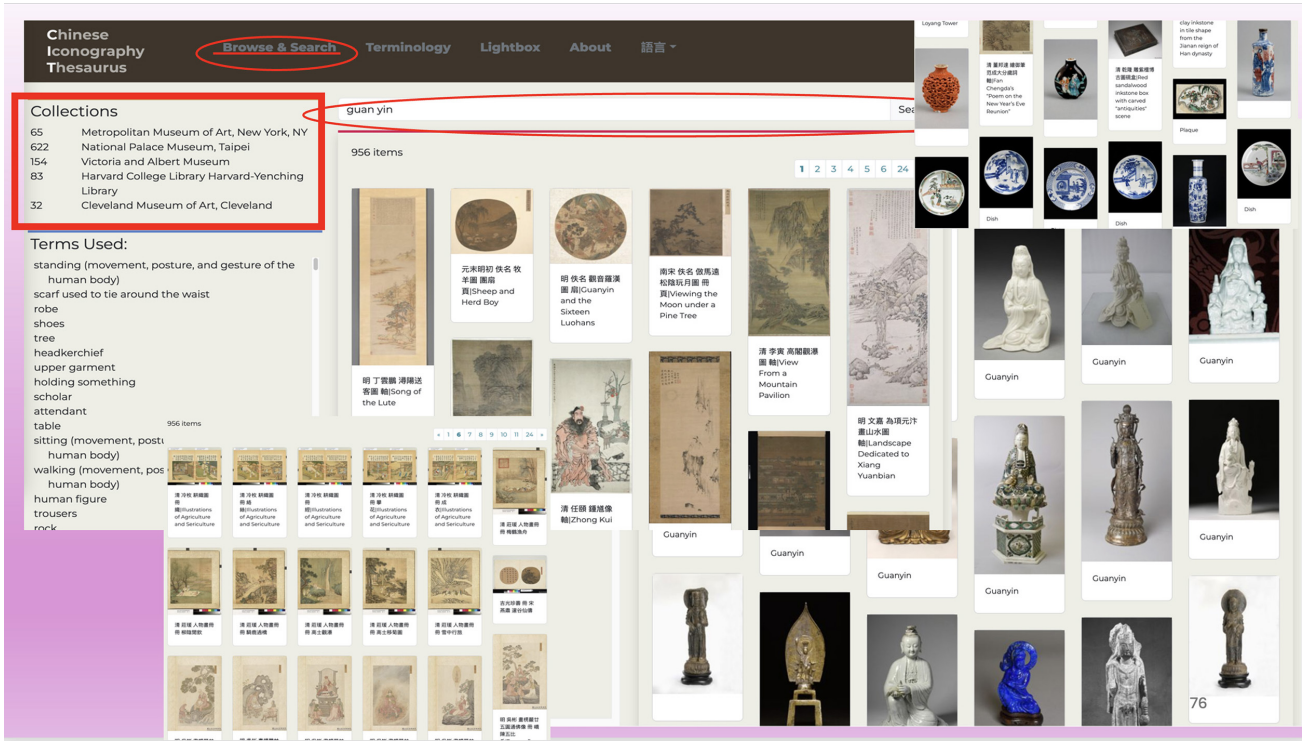


Fig. 8. Results of searching the Chinese Iconography Thesaurus. © Victoria and Albert Museum, London. Images captured and marked 2025-10-8.

cabulary solution. Some of these functionalities are well illustrated by the following KOS services, allowing vocabularies to be directly accessed and queried by computer programs.

■ UNESCO Thesaurus, available in English, French, Spanish, Russian, and Arabic, provides a SPARQL query form.[49] ([49]https://vocabularies.unesco.org/sparql-form/) The query templates include "Get all concepts in English and French, with synonyms, notes, broader, narrower, and related (with IDs)", "List all microthesauri" (which results in over 100 microthesauri, including more than 10 in "Culture"), etc.

■ The Food and Agriculture Organization (FAO)'s AGROVOC covers over 40 languages. Each entry is associated with a full list of exactly matching concepts from at least 10 well-known KOSs (e.g., EuroVoc, ZBW Thesaurus for Economics) and DBpedia (see example for the entry for AI).[50] ([50]https://agrovoc.fao.org/browse/agrovoc/en/page/c_27064) In addition to the "Search & browse" option, it provides a programmatic access interface, i.e., a set of REST APIs,[51] ([51]A REST API is an application programming interface (API) that fits the design principles of the architectural style of representational state transfer (REST), a style used to connect distributed hypermedia systems.) allowing computer programs to access its vocabulary data, and a SPARQL endpoint service to provide predefined query templates.[52] ([52]https://agrovoc.fao.org/sparql or access from https://agrovoc.fao.org/ — SPARQL endpoint) For example, one can "Select all labels for the concept 'water' in different languages", "Select the countries and regions with their M49 code and with prefLabel in English", etc.

■ Getty Vocabularies LOD[53] ([53]https://vocab.getty.edu/) provides a SPARQL endpoint for multiple Getty vocabularies, AAT, Getty Thesaurus of Geographic Names (TGN), and Union List of Artist Names, through both a data dump and a SPARQL endpoint, plus a comprehensive list of nearly 100 query templates and documentation.[54] ([54]See the list of Sample Queries at https://vocab.getty.edu/queries) The contents are directly linked to the website of the vocabularies. There are multiple ways to generate microthesauri for subjects and obtain geographical ones with various contexts. For example, one can find places and generate datasets based on different focuses, in addition to the data of places by or within a coordinate bounding box. Specific artist entries (e.g., Native American painters, ancient artists, or groups by nationality) can be collected easily. Using this service, one can, for instance, rapidly collect data for certain types of places (e.g., caves, UNESCO Heritage sites, etc.) on or around the Silk Road. Following these places and using TGN-provided geo-coordinates, one can obtain over 100 locations on the map (through TGN's API) and launch a search for various museums' collections through them (Zeng et al., 2017).

Reliable aggregate KOS repositories and registries empower the reuse and interoperability of existing KOS while facilitating the development and ensuring the quality of new ones. Aggregate portal services can support terminology searching, sharing, mapping, and visualization, and the analysis of large repositories of ontologies, vocabularies, terminologies, and annotations. Readers are advised to refer to "Appendix 2. A list of vocabulary repositories and registries" in Zeng (2018).

In addition to the portals presented above, for CH, the Linked Conservation Data project provides a conservation controlled vocabulary service[55] ([55]https://www.ligatus.org.uk/lcd/controlled-vocabularies) with more than 50 listed vocabularies. In addition to AAT and multiple Forum on Information Standards in Heritage thesauri, several illustrated glossaries and a direct Visual Glossary – Australian Institute for the Conservation of Cultural Material[56] ([56]https://aiccm.org.au/conservation/visual-glossary/) are available.

Basel Register of Thesauri, Ontologies & Classifications (BARTOC)[57] ([57]https://bartoc.org/) is a terminology registry facilitating the use of KOSs. Currently, it has over one thousand KOS metadata entries, including active, inactive, or historical vocabularies. Notable examples are as follows.

Classification of Artificial Intelligence, Artificial Intelligence Glossary, and others for AI;
over 100 entries dedicated to CHs, including

- cross-languages, such as

– Multilingual Inventory of Cultural Heritage in Europe (updated 2024), available in 16 languages,
– Glossary of Intellectual Property Terms: English–Korean–Chinese, Korean–English–Chinese;

- specific languages or cultures, such as

– Korean History Thesaurus,
– Classification of the Morrison Collection in Toyo Bunko,
– Thot – Thesauri & Ontology for Documenting Ancient Egyptian Resources, etc.

Furthermore, BARTOC includes the metadata of over 140 other registries, which provide access to full terminologies either via an API (terminology service) or by other means (terminology repository).

"DE-BIAS – Detecting and cur(at)ing harmful language in cultural heritage collections" is a project cofunded under the Digital Europe Programme (DIGITAL) of the European Union,[58] ([58]https://pro.europeana.eu/project/de-bias) aiming to make CH metadata more inclusive.

■ DE-BIAS Vocabulary,[59] ([59]https://pro.europeana.eu/page/the-de-bias-vocabulary) covering 700 words across five languages, was established to support the AI-powered DE-BIAS Tool[60] ([60]https://pro.europeana.eu/page/the-de-bias-tool) (created 19 December 2024), which detects outdated and potentially harmful language in the descriptions of CH collections and thus allows professionals

to review these descriptions with regard to potential bias. This vocabulary mainly focuses on migration and colonial history, gender and sexual identity, and ethnicity and ethnoreligious identity, with some terms relating to other minority communities.

■ With input from underrepresented groups and research into over 100 existing glossaries and (academic) publications, the terms in this vocabulary are accompanied by contextual information and, where appropriate, suggestions for reflection and alternative wording. The vocabulary cites the sources for the description of each term to increase their visibility and support their automated use for bias detection in CH databases.

■ The presentation of the vocabulary (a living document which continues to be updated) is also very distinctive, featuring five language-specific parts that are generated from the knowledge graph and available in PDF format, with a machine-readable knowledge graph formulated in SKOS.[61] ([61]See the demo in https://pro.europeana.eu/files/Europeana_Professional/Projects/de_bias_vocabulary/DE-BIAS_vocabulary_presentation_slides.pdf)

4. Discussion and Concluding Remarks

Despite the unparalleled power of AI-based discovery services over vast data sets, having a robust KOS and reliable metadata is crucial for any data service aiming to provide accurate, high-quality, and interoperable information access (Slavic et al., 2025). This availability is critical for all information domains where precision and accuracy are of the essence and information may be ambiguous and open to interpretation. This paper looks into two information domains exhibiting these characteristics, namely health and CH, and analyzes examples where AI-based discovery tools can be deployed and resources—primarily textual—lend themselves to easy automatic processing. Moreover, CH information resources often cannot be processed by programs and thus require information discovery to be based on metadata. In both contexts, the role of KOS and the potential for improving information discovery services are obvious.

The most innovative approaches and greatest potentials are observed for the healthcare domain, which contains abundant digitally born and well-structured information resources. This paper illustrates the advantages of using AI-empowered data and information services with tools featuring advanced machine learning methods beyond LLMs. One can also deploy numerous specialized KOSs directly supporting clinical practice and connected to advanced AI-supported KOS services. In the health domain, these vocabularies are not only processed for automatic indexing but also automatically managed, updated, and semantically linked. The role of human indexers in these applications is primarily reserved for quality control and enhancement.

In the CH domain, two categories of information resources and related services are typically encountered, namely (i) textual and machine-readable/processable data; and (ii) diverse nontextual CH objects which need to move from digitization to datafication and further to contextualization empowered by advanced data and information services and tools. Being more reliant on metadata and semantic disambiguation, this domain features abundant KOS vocabularies and services. Many of these vocabularies are enhanced and supported through semantic technology, are available for direct processing, and can be utilized in advanced KOS services while being further improved through AI-based solutions. Thus, AI opens new avenues for research, improves accessibility, and enhances preservation efforts. The role of humans in validating and ensuring KOS quality in this domain remains strong because of the high level of ambiguity caused by the temporal, language, and cultural characteristics of CH resources.

In addition to the two resource discovery scenarios focused on herein, other noticeable actions have been reported. As highlighted in the Online Computer Library Center (OCLC) 2025 report, notable improvements with new AI features were added to WorldShare Record Manager[62] ([62]https://www.oclc.org/en/worldshare-record-manager.html) and Connexion[63] ([63]https://www.oclc.org/en/connexion.html) cataloging applications. These timely actions help catalogers by automatically suggesting classification numbers and subject headings, saving time while improving accuracy, and providing AI-generated suggestions for Dewey Decimal Classification numbers, Library of Congress Classification numbers, and Library of Congress Subject Headings entries as records are created or edited. The system draws on WorldCat data—built from hundreds of millions of library records—to make these intelligent recommendations (OCLC, 2025). Similar to MEDLINE expectations in relation to MTIX, OCLC suggests that human indexers will continue to play a significant role in quality assurance for AI-generated metadata. The OCLC strategy ensures that catalogers stay in full control of the indexing process. Each library can customize settings to decide which suggestions to see, and catalogers can accept or ignore them. Their choices help improve future recommendations, making the system smarter.[64] ([64]https://www.oclc.org/en/news/releases/2025/20251208-ai-recordmanager-connexion.html)

Many supplementary individual KOSs address the perceived lack of diversity and equity and provide unbiased views with respect to persons, peoples, cultures, and religions, to ensure that KOSs become more multilingual, multicultural, and inclusive. Ethically, culturally, and politically prompted actions on KOS terminology are reported in numerous publications. Below, an illustrative selection of terminology supplements, recommendations, and new vocabularies is provided.

■ Diversity, Equity and Inclusion Principles for Custom Taxonomies (Mizota, 2023)

■ Getty Vocabularies: Issues Regarding Diversity and Inclusion and Terminology for Expressing Points of View (Harpring, 2022, 2025)

■ Reimagine Descriptive Workflows: A Community-informed Agenda for Reparative and Inclusive Descriptive Practice, OCLC (Frick and Proffitt, 2022)

■ Recommendations for Describing Gender Identity[65] ([65]https://guides.library.yale.edu/ld.php?content_id=80267736) and A Gender Glossary,[66] ([66]https://reflections.yale.edu/article/sex-gender-power-reckoning/gender-glossary) Yale University Library

■ Homosaurus[67] ([67]https://homosaurus.org/), an international linked data vocabulary of LGBTQ+ (abbreviation for lesbian, gay, bisexual, transgender, queer, and other identities) terms. Designed to enhance broad subject term vocabularies, this robust cutting-edge thesaurus increases the discoverability of LGBTQ+ resources and information.

Overall, it is important to have KOS in AI-supported semantic integration, especially when it comes to eliminating ambiguity, enhancing identification quality, and enabling effective CH entity connections across scripts, languages, and cultures. There are endless possibilities.

Availability of Data and Materials

Not applicable.

Author Contributions

MLZ designed the study, conducted the research, and wrote the manuscript. The author read and approved the final manuscript and is accountable for all aspects of the work.

Acknowledgment

The author would like to thank Dr. Aida Slavic and ISKO-UK for their encouragement and providing their valuable feedback.

Funding

This research received no external funding.

Conflicts of Interest

The author declares no conflicts of interest. Marcia Lei Zeng is serving as one of the Editorial Board members and Guest Editors of this journal. We declare that Marcia Lei Zeng had no involvement in the peer review of this review and has no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to Natália Tognoli.

Appendix

Appendix A. Diverse Cultural Heritage Entities Explored in the New Age of AI

1. Text

- eScriptorium

  https://escriptorium.inria.fr/

2. Images of Artworks

- AI Explorer, Harvard Art Museums

  https://ai.harvardartmuseums.org/

- AI Collection Toolsets, Cleveland Museum of Art

  https://www.clevelandart.org/digital-innovations

3. Sound

- Project Gutenberg Open Audiobook Collection

  https://marhamilresearch4.blob.core.windows.net/gutenberg-public/Website/index.html

- Wildlife media archive, Cornell Univ. Macaulay Library

  https://www.macaulaylibrary.org/ & "Machine Learning at Macaulay Library"

  https://www.macaulaylibrary.org/machine-learning/

4. Scents

- Odeuropa project

  https://odeuropa.eu/

5. Museum Specimens

- Read: "Global push to put natural history collections online gets major U.K. boost" - London's Natural History Museum gets nearly $200 million to scan country's collections

  https://www.science.org/content/article/global-push-put-natural-history-collections-online-gets-major-u-k-boost. — Science, Vol 384, Issue 6691. April 2024

6. 3D+

- EUreka3D - European Union's REKonstructed content in 3D

  https://pro.europeana.eu/project/eureka3d-european-union-s-rekonstructed-content-in-3d

- 5DCulture - Deploying and demonstrating a 3D cultural heritage space

  https://pro.europeana.eu/project/5dculture-deploying-and-demonstrating-a-3d-cultural-heritage-space

- DE-BIAS - Detecting and cur(at)ing harmful language in cultural heritage collections

  https://pro.europeana.eu/project/de-bias

- AI4Culture - Developing a platform which offers access to AI-related resources (including openly labelled datasets for training and testing AI models), alongside a set of deployable and reusable tools and capacity building materials

  https://www.dataspace-culturalheritage.eu/en/projects/ai4culture

7. Rock Art

- Read: "Artificial intelligence helps catalogue Aboriginal rock art in Far North Queensland"

  https://www.abc.net.au/news/2023-09-15/artificial-intelligence-rock-art-far-north-queensland-indigenous/102851802

- Read: "Artificial intelligence is helping researchers identify and analyse some of the world's oldest art"

https://www.theartnewspaper.com/2023/03/21/artificial-intelligence-ancient-rock-art-australia-andrea-jalandoni

8. Subtractive Notations

- Guqin Subtractive Notation (shows the fingering techniques for Playing Guqin, one of the oldest traditional Chinese musical instruments).

Read: "The ontology enhanced multimodal large language models for the knowledge organization and representation of multi-modal cultural memory resources"

https://nkos.dublincore.org/2024NKOSworkshops/Xia-NKOS2024_upload.pdf & Read: "A Knowledge-enhanced Multi-modal Large Language Model for Chinese Guqin Subtractive Notation Interpretation" https://www.taylorfrancis.com/reader/download/a5d0dffc-d0e8-436f-8a63-f87d959c6721/chapter/pdf

9. Unicode Symbols

- Read: "AI Enabled Text-to-Speech Synthesis for Unicode Language"

https://sciresol.s3.us-east-2.amazonaws.com/IJST/Articles/2024/Issue-42/IJST-2024-2645.pdf

10. Coins

- Online Coins of the Roman Empire (OCRE).

https://numismatics.org/ocre/

- Read: "From coin to data: the impact of object detection on digital numismatics"

https://doi.org/10.1093/llc/fqaf046

- CoinSampo. Finnish Numismatic Public Finds 2013–2023

https://rahasampo.ldf.fi/en/

11. Multiple and Unified all Kinds of Resources

- The Sampo Model and Semantic Portals for Digital Humanities

https://seco.cs.aalto.fi/applications/sampo/

For more cases and research reports, consult Digital Humanities and Intelligent Computing, a biennial book series, London: Routledge. Open sources are available online.

References

Berners-Lee T. Intelligent machines. In This is for everyone: The unfinished story of the World Wide Web. Farrar, Straus and Giroux: New York. 2025.

Bianchini C, Bargioni S, Pellizzari di San Girolamo CC. Beyond VIAF: Wikidata as a complementary tool for authority control in libraries. Information Technology and Libraries. 2021; 40. https://doi.org/10.6017/ital.v40i2.12959

Busch J, Zeng ML. Special Issue of KO (Knowledge Organization): KOS in AI and AI in KOS. Knowledge Organization. 2025; 52. https://doi.org/10.31083/ko47822

Chen SJ. Design of interoperability in digital humanities: A case study of the interpretation and restoration of the Han Dynasty Wooden Slips from Edsen-Gol. Journal of Educational Media & Library Sciences. 2021; 58: 193–235. https://doi.org/10.6120/JoEMLS.202107_58(2).0007.RS.AM (In Chinese)

Chen SJ, Lu-Yen LU. The application of IIIF and LOD in digital humanities: A case study of the Dictionary of Wooden Slips. In The 14th Semantic Web in Libraries Conference (SWIB 2022). Online. 2022.

Du Y, Huang W, Zheng D, Wang Z, Montella S, Lapata M, et al. Rethinking memory in AI: Taxonomy, operations, topics, and future directions. arXiv preprint. 2025. Available at: https://arxiv.org/pdf/2505.00675 (Accessed: 2 February 2026).

Frick RL, Proffitt M. Reimagine descriptive workflows: a community-informed agenda for reparative and inclusive descriptive practice. OCLC Research: Dublin, OH. 2022. https://doi.org/10.25333/wd4b-bs51

Harpring P. Getty Vocabularies: Issues regarding diversity and inclusion. AAT & Unbiased Terminology workshop. Los Angeles, CA. USA. 2022. Available at: https://youtu.be/AlznfxKwI6Y (Accessed: 2 February 2026).

Harpring P. Getty Vocabularies: Unbiased terminology and diversity. Getty Vocabulary Program. Los Angeles, CA. USA. 2025. Available at: https://www.getty.edu/research/tools/vocabularies/Vocabs_unbiased_terminology.pdf (Accessed: 2 February 2026).

Hill LL, Hodge G, Busch J. Networked knowledge organization systems (NKOS). Proceedings of the Fourth ACM Conference on Digital Libraries. 1999; 263. https://doi.org/10.1145/313238.322885

Hong Y, Zeng ML. International Classification of Diseases (ICD). Knowledge Organization. 2022; 49: 496–528. https://doi.org/10.5771/0943-7444-2022-7-496

International Organization for Standardization. What is artificial intelligence (AI)? ISO Insights & news. 2024. Available at: https://www.iso.org/artificial-intelligence/what-is-ai (Accessed: 2 February 2026).

Joachimiak MP, Miller MA, Caufield JH, Ly R, Harris NL, Tritt A, et al. The Artificial Intelligence Ontology: LLM-Assisted Construction of AI Concept Hierarchies. Applied Ontology. 2024; 19: 408–418. https://doi.org/10.1177/15705838241304103

Kourounis G, Elmahmudi AA, Thomson B, Hunter J, Ugail H, Wilson C. Computer image analysis with artificial intelligence: a practical introduction to convolutional neural networks for medical professionals. Postgraduate Medical Journal. 2023; 99: 1287–1294. https://doi.org/10.1093/postmj/qgad095

Mazzocchi F. Knowledge Organization System (KOS): An Introductory Critical Account. Knowledge Organization. 2018; 45: 54–78. https://doi.org/10.5771/0943-7444-2018-1-54

Mizota S. Diversity, equity and inclusion principles for custom taxonomies. Journal of Digital Media Management. 2023; 12: 46–52.

National Library of Medicine. MTIX: the next-generation algorithm for automated indexing of MEDLINE. NLM Technical Bulletin. 2024: e4.

OCLC. OCLC introduces new AI tools to make cataloging faster and smarter. OCLC News. 2025. Available at: https://www.oclc.org/en/news/releases/2025/20251208-ai-recordmanager-connexion.html (Accessed: 2 February 2026).

Patel VL, Kushniruk AW, Yang S, Yale JF. Impact of a computer-based patient record system on data collection, knowledge organization, and reasoning. Journal of the American Medical Informatics Association : JAMIA. 2000; 7: 569–585. https://doi.org/10.1136/jamia.2000.0070569

Qamar R, Zardari BA. Artificial Neural Networks: An Overview. Mesopotamian Journal of Computer Science. 2023; 2023: 124–133. https://doi.org/10.58496/mjcsc/2023/015

Slavic A, Golub K, Haynes D, Gnoli C, Salaba A. Subject metadata in library discovery systems: A survey report. IFLA Metadata Newsletter. 2025; 11: 34–37. Available at: https://repository.ifla.org/handle/20.500.14598/6984

Sowa JF. Knowledge representation: logical, philosophical and computational foundations. Brooks/Cole Publishing Co: Pacific Grove, CA. 1999.

Sowa JF, Majumdar A. Evaluating and reasoning with and about GPT. Ontology Summit, online. 2023. Available at: https://youtu.be/6K6F_zsQ264 (Accessed: 2 February 2026).

Sowa JF, Majumdar A. Generating ontologies. Ontology Summit 2025, Track 2. 2025. Available at: https://ontologforum.org/index.php/ConferenceCall_2025_02_26 (Accessed: 2 February 2026).

Stokes PA, Kiessling B, Ezra DS, Tissot R, Gargem EH. The EScriptorium VRE for manuscript cultures. Classics@ Journal. 2021; 18.

Tudhope D, Koch T. New applications of knowledge organization systems: introduction to a special issue. Journal of Digital Information. 2004; 4.

Wang X, Zeng ML, Gao J, Zhao K. AI and Smart Data for Cultural Heritage–Global Achievements and China's Innovations: Preface. Routledge: London. 2026. https://doi.org/10.4324/9781003666530

World Health Organization. ICD-11 Reference Guide. 2022. Available at: https://icdcdn.who.int/icd11referenceguide/en/html/index.html (Accessed: 2 February 2026).

World Health Organization. Naming the coronavirus disease (COVID-19) and the virus that causes it. 2020. Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it (Accessed: 2 February 2026).

World Health Organization. WHO announces simple, easy-to-say labels for SARS-CoV-2 Variants of Interest and Concern. 2021. Available at: https://www.who.int/news/item/31-05-2021-who-announces-simple-easy-to-say-labels-for-sars-cov-2-variants-of-interest-and-concern (Accessed: 2 February 2026).

World Health Organization. WHO best practices for naming of new human infectious diseases. 2015. Available at: https://www.who.int/publications/i/item/WHO-HSE-FOS-15.1 (Accessed: 2 February 2026).

Yan C, Fu X, Liu X, Zhang Y, Gao Y, Wu J, et al. A survey of automated International Classification of Diseases coding: development, challenges, and applications. Intelligent Medicine. 2022; 2: 161–173. https://doi.org/10.1016/j.imed.2022.03.003

ZBW-Leibniz Information Centre for Economics. Annual Reviews by the ZBW: ZBW-Highlights from 2025. 2025. Available at: https://www.zbw.eu/fileadmin/pdf/ueber-uns/2025/jb-2025-e.pdf (Accessed: 2 August 2026).

Zeng ML, Hong Y. Challenges, opportunities, and approaches in a health KOS vocabulary's revision – Insights from the 11th revision of the International Classification of Diseases (ICD-11). NKOS Workshop 2022 online. 2022. Available at: https://nkos.dublincore.org/2022NKOSworkshop/Zeng&Hong-ICD11-2022NKOS-PPT.pdf (Accessed: 2 February 2026).

Zeng ML, Hong Y, Clunis J, He S, Coladangelo LP. Implications of Knowledge Organization Systems for Health Information Exchange and Communication during the COVID-19 Pandemic. Data and Information Management. 2020; 4: 148–170. https://doi.org/10.2478/dim-2020-0009

Zeng ML. Incorporating representative visual resources in a thesaurus. 18th ISKO International Conference, Wuhan, China. 2024. Available at: https://nkos.dublincore.org/2024NKOSworkshops/MZdemos-visualReps.pdf (Accessed: 2 February 2026).

Zeng ML. Interoperability: Appendix 2. A list of vocabulary repositories and registries. ISKO Encyclopedia of Knowledge Organization. 2018. Available at: https://www.isko.org/cyclo/interoperability (Accessed: 2 August 2026).

Zeng ML, Mayr P. Knowledge Organization Systems (KOS) in the Semantic Web: a multi-dimensional review. International Journal on Digital Libraries. 2019; 20: 209–230. https://doi.org/10.1007/s00799-018-0241-2

Zeng ML. The role of KOS in AI-supported semantic integration: disambiguation, identification, linking. 8th ISKO UK biennial conference, online. 2025. Available at: https://www.iskouk.org/resources/Documents/2025Conference/MZeng-ISKOUK2025Conf.pdf (Accessed: 2 February 2026).

Zeng ML, Ward J, Garcia G, Hu T. Extending exhibitions to historical journeys through data [in the Semantic Web]. Professional Forum, Museums and the Web Conference. Cleveland, OH. USA. 2017. Available at: https://mw17.mwconf.org/proposal/extending-exhibitions-to-historical-journeys-through-data-in-the-semantic-web/index.html (Accessed: 2 February 2026).