Using LLMs for Semantic Alignment: A Study on Archival Metadata Description

Abstract

The advantages of aligning custom data schemas with standardised ontologies within their respective knowledge domain have long since been proven in practice. Sharing a common structural representation by mapping concepts and relationships between the schemas is essential to ensure data interoperability (especially on a semantic level), integration, reuse, and the ability to leverage machine-processable and advanced-search capabilities. Archival institutions preserve, manage, and provide access to large amounts of diverse cultural and historical data, demonstrating a high potential to be active contributors to a global knowledge network, should archival data be transformed and offered as linked (open) data. Based on the expert-validated dataset of the alignment (mapping) of the Swedish National Archives schema to the Records-in-Contexts (RiC-O) ontology, the purpose of this study is two-fold. First, to examine whether it is possible to automatically and effectively extend one case (Sweden) to other archival institutions and align new custom schemas to RiC-O, given an expert-curated dataset of this domain. Secondly, using the aforementioned dataset and one more of a few human-evaluated examples of mapping to other cultural heritage ontologies as input, to examine whether an LLM (e.g., GPT-4o) can recommend meaningful alignments for enhanced metadata description to more ontologies within the same domain (CH and archives), but also across other domains. The experiments reveal several challenges and shortcomings of the LLM prompting approach for these tasks, but also possible opportunities to leverage towards this direction.

Keywords *Archival data, archives, cultural heritage, semantic interoperability, ontology alignment, linked open data, digital archives, LLMs, GPT-4*

1. Introduction

The advantages of aligning custom metadata schemas with standardised ontologies within their respective knowledge domain have long since been emphasised (Linked Data - W3C Wiki, n.d.). Sharing a common structural representation by mapping concepts and relationships between the schemas is essential to ensure data interoperability (especially on a semantic level), integration, reuse, and the ability to leverage machine-processable and advanced search capabilities. Archival institutions preserve, manage, and provide access to large amounts of diverse cultural and historical data, demonstrating a high potential to be active contributors to a global knowledge network, should archival data be transformed and offered as linked (open) data. Digital archival collections are growing larger as mass digitisation takes place, accompanied by an equally increasing volume of born-digital archives over the past two decades (Hawkins, 2021a). However, mass digitisation often lacks the equivalent progress in digital humanities requirements for structured, discoverable, and interoperable data, bringing forth the fundamental necessity for Linked Data practices for this purpose (Hawkins, 2021a).

The Swedish National Archives (Riksarkivet, n.d.) (in Swedish: Riksarkivet), is the responsible institution for Sweden's documentary and historical heritage preservation, and one of the oldest archival institutions globally, tracing its foundation back in 1618. The archives cover a wide variety of documents, ranging from state, military, legal, genealogical, and regional documents, to royal charters, maps, letters and medieval scripts and books. Millions of archival records

(amounting over 75 km of physical artifacts) are available and, although a small percentage of archives is digitised, due to the immense size of the archival collection, the digital archives still amount to over 100 million digital artifacts, available (some might fall under classification restrictions) through the Swedish National Archives' portal and search function. The digitisation process is still ongoing, aiming to make archives more accessible digitally, while taking into consideration the adoption and sharing of new methods and technologies for archival data management.

This research uses the outcome of the conceptual mapping between the Records-in-Context Ontology (RIC-O) and the schema in use at the Swedish National Archives as a basis for further experimentation on the potential of introducing automated work streams in the mapping process. The process of schema alignment may vary greatly from case to case and depending on the schema of both the institution and the ontology (or ontologies involved), some of them might reach too high of a level of complexity for human actors to process. Apart from the time constraints, human actors can easily get entangled in the complexity of a high interconnectedness between classes and attributes, leading to an incomplete or insufficient result. In fact, in many cases it could act as a deterring factor for even attempting this process. Existing tools and documentation might help in this regard, however, for most of them to function effectively and provide valuable mapping recommendations, still a lot of effort needs to be devoted by a human actor to input proper information, consider the peculiarities of each schema and case, and guide the tools towards the desired direction. Several ontology matching and alignment tools exist, such as AML, LogMap, or other semi-automated approaches, however, recent advances in Large Language Models (LLMs) bring forth new possibilities for ontology alignment tasks that rely on contextual reasoning. The goal of the present study is not to replace existing ontology alignment

frameworks and tools, but to test whether LLMs could work complementarily to existing approaches (specifically for archival metadata alignment and ontology reuse), as well as identify strengths and limitations in a human-in-the-loop setting. So, the purpose of this study, apart from the alignment of a custom metadata schema to a standardised one, is to leverage a local case study towards experimentation which can potentially bring forth value in similar cases, as well as showcase how the results and produced data can further expand this research beyond the boundaries of a single case scenario. The experiment of this study is two-fold; the expert-validated mapping between the Swedish National Archives and Records-in-Contexts (RiC-O) is used as prior knowledge on GPT-4o in order to: i) similarly perform the mapping given another custom archival institution schema and, ii) assist in recommending alignments with other schemas in Cultural Heritage but also other domains, such as ones available in Linked Open Vocabularies (LOV), in order to foster re-use and enhance semantic granularity for metadata description, wherever feasible.

This document is structured as follows: Section 2 (Background) includes the research basis and motivation of this study, as well as introductory information about the Swedish National Archives structural schema. Section 3 (Method) presents the methodological steps followed for the two-fold experiment, and Section 4 (Results) is where the outcomes of the experiments are presented and analysed. Section 5 (Discussion) includes the analysis and insights gained from the results, their potential use and re-use by practitioners of the field, and, finally, Section 6 refers to the limitations present in this work, along with future directions it can follow.

The present study is a practical application and continuation of the work presented at the Extended Semantic Web Conference - ESWC2025: “Empowering Knowledge Through Semantics: From Knowledge Graphs to Neurosemantics” (Maratsi, Haskiya, et al., 2025).

2. Background

Digital data is a core resource for digital humanities; however, digitalised archival data also needs to be integrated, interoperable, and interrogable (Bikakis, 2021) and align with the FAIR guiding principles of data publication (Findable, Accessible, Interoperable and Re-usable) (GO FAIR initiative, 2022). Linked data allows for the facilitation of FAIR data implementation, creating Archival Linked Data (ALD), which is machine-readable, contextual, and can be analysed using digital humanities (and beyond) methods for research and engagement, but despite the traction, linked data remains under-examined in an archival data context (Hawkins, 2021). Linked archival data shepherds numerous benefits; improvements in knowledge discovery, information retrieval, revealing unknown relationships across archival collections (even cross-domain navigation between cultural and non-cultural heritage data sources) (Gracy, 2014, 2017), improving data quality, and enabling semantic search queries (SPARQL) (Hawkins, 2021b), are but a few of them, enabling digital humanists and users to more easily understand and use archival data in context (Hawkins, 2021a).

The quest for making archival records' metadata available as Archival Linked Data (ALD) involves standardisation and alignment with archival descriptive standards and ontologies, such as the Records in Contexts Ontology (International Council on Archives Records in Contexts Ontology (ICA RiC-O) Version 1.0.2, n.d.), unification with the global network of OpenGLAM for GLAM (Galleries, Libraries, Archives and Museums) (OpenGLAM.org, n.d.) or Europeana's Linked Data web service (Overview for "linked-open-data" | Europeana PRO, n.d.). Linked Open Data use in cultural heritage paves the way towards application also in the archival domain, with RDF and ontological approaches being sine qua non for semantic interoperability

and better knowledge management for archival data (Mazzini, Ricci, 2011). RDF (Resource Description Framework) is the basis for Linked Data structure and expressed in triples: a semantic unit consisting of three components: subject -> predicate -> object. Archival resources may follow international standards, such as EAD (Encoded Archival Description), an XML standard for the encoding of finding aids for use in an online environment (Encoded Archival Description (EAD) | Society of American Archivists, n.d.), the web indexer Schema.org model, or more specialised models such as LODE (Linking Open Descriptions of Events) (Dobreski, Park, et al., 2020).

However, to achieve contextual archival description in a semantically valuable manner, it is necessary to identify resources by the means of dereferenceable URIs, standardised descriptions and relations format, and linked descriptions to other information resources to the largest possible extent (Mazzini, Ricci, 2011). In general, the hierarchical structure of archival documents, according to the ISAD(G) - General International Standard Archival Description (ICA, 2025) follows the pattern shown in Fig. 1.

At the same time, archives usually contain diverse types of data and materials (letters, maps, paintings, books, photographs, sound recordings, etc.), a characteristic introducing challenges in archival description and the ability to sufficiently capture the richness and depth of each individual case, considering that existing ontological approaches and schemas differ in scope and descriptive granularity (Dobreski, Park, et al., 2020). For instance, archival descriptions often offer less granularity compared to linked data vocabularies (e.g., vocabularies available in the Linked Open Vocabularies repository (Linked Open Vocabularies (LOV), n.d.) or similar.). Modeling approaches and choosing among standards to form an appropriate metadata schema differ according to each institution’s individual needs and data infrastructure. Chen (2019)

explored various methods of semantic enrichment for art-related archival resources and chose the Europeana Data Model (EDM) as the core data model design, while also following additive approaches such as direct reuse of other external vocabularies, local links to other data sources, introduction of contextual classes, and the utilisation of named entity extraction. Earlier studies, such as the research by Bountouri and Gergatsoulis (2011) demonstrate different approaches in representing archival hierarchies and instead mapping the corresponding EAD (Encoded Archival Description) elements to the CIDOC Conceptual Reference Model (CIDOC CRM) for cultural heritage metadata integration. CIDOC CRM and additional ontologies were also encompassed in an archival context of a linked open data model for the Portuguese Archives in which case, due to the lack of deployment testing of the newest Records in Contexts Conceptual Model (RiC-CM) by ICA (International Council on Archives), it was deemed more suitable to turn to CIDOC-CRM's long maturity for element representation (Koch et al., 2023). On the other hand, RiC-CM's most recent models (e.g., RiC-O), apart from the archival intrinsic structure, also feature a larger collection of properties to describe archival relations. RiC-O's adaptability to various archival contexts and connectivity to other cultural heritage domains is also showcased by the several semantic modeling projects it is used in, including projects and initiatives led by the National Archives of France (Ministère de la Culture de France, Service interministériel des Archives de France, n.d.).

As far as the involvement of Large Language Models in semantics and ontology engineering is concerned, the symbiosis of humans and machines as a new domain to explore was brought up by Doumanas et al. (2024), who performed LLM-based ontology engineering but at the same time observed that human contribution and involvement notably enhanced the process and the results. A similar statement was supported by Osman et al. (2024), who emphasise the benefits

and generalisability of having a well-designed, semi-supervised approach in ontology matching, after experimenting with automatic approaches and observing that fully automated solutions are still unreliable. Hoseini et al. (2024) utilised the natural language processing capabilities of LLMs to label and model data semantically in the context of data spaces, while Pan et al. (2025) presented an LLM-based retrieval-augmented generation (RAG) approach for automatic generation of competency questions in ontology engineering. Pan et al. (2025) also observed that adding domain knowledge to their RAG process improved LLM performance in this task. Cigliano and Fallucchi (2025) identified the potential of an intersection among open and linked data, ontologies, and LLMs, and support the statement that this combination may revolutionise how value from data and structured information can be derived. The potential of utilising LLMs for domain-specific (Cultural Heritage) but also cross-domain recommendations for alignment and evaluation was also studied by Maratsi et al., (2024b), and the potential of a proposed methodological framework combining automated means (LLMs), ontological foundation, and graph theory metrics for improving semantic interoperability and interdisciplinary discoverability of data as an enhanced semantic search capability, was presented by Maratsi et al. (2025c). Other approaches such as AML for general ontology matching, LogMap for logic-based matching, studies on cultural heritage ontology matching and integration, such as the work by Rinaldi et al., (2025) as well as approaches combining LLMs with knowledge retrieval such as KROMA by Nguyen et al., (2026) do exist, on the other hand, the focus of the present study is to establish a first benchmark for prompt-based alignment. Unlike AML or LogMap which primarily rely on logical or structural matching or KROMA which combines external retrieval with LLMs, this study intentionally targets the capabilities and limitations of prompting alone.

Apart from that, following the latest advancements in LLM-enabled processes for semantic interoperability and ontology mapping, and taking into consideration the observed value of human-in-the-loop in such a process, the present work aims to combine human and machine expertise for ontology alignment and metadata enrichment in the archival and cultural heritage domain using LLMs (GPT-4o) by first exemplifying the mapping based on an a priori human-validated manual mapping process. The goal of this venture is to pave the way towards technologically enabled semantic alignment, not only within the same domain, but also enabling cross-domain ontology and vocabulary re-use (such as vocabularies available at the Linked Open Vocabularies repository).

3. Method

3.1 Prior Knowledge - Groundwork

The groundwork for this study lies and builds on the conceptual alignment between the Swedish National Archives (abbreviated RA from "Riksarkivet") and Records-in-Contexts (RiC-O) performed by Maratsi et al. (2025a). The expert-validated dataset which resulted from this process is used to anchor a measurement of "truth" for the expected results and lead the LLM throughout the prompting interaction.

Before presenting the methodological process followed for the experiments of this study, a brief reference to the basics of the method followed for the prior semantic alignment between RA and RiC-O is made. As described by Maratsi et al. (2025a), to facilitate the conceptual mapping, the first step was to create a general taxonomy of archival data organisation in the Swedish National Archives, a mental model to help visualise the concepts schematically. The RA taxonomy consists of the main archival document hierarchical structure, very closely following the

ISAD(G) structure (see Fig. 1), with the differentiation, however, that RA includes the concept of “Volume (in Swedish: Volym)” which refers to a box of documents or items and is part of an archival Series (in Swedish: Serie), which is part of an Archival fond. Apart from the archive hierarchical structure, the taxonomy includes the types of archival resources (as these are organised in the RA search function), the archival institution responsible for the archive, the archivist (responsible for the insertion of archival entries), the archive contributors, and the places (and types of places) the archives derive from. For the archival record description there are several metadata groups, such as basic metadata (reference code, date, dimensions, access rules, archival institution etc.), metadata related to the content of the record, control and actions metadata (latest modification, source year, source, etc.), accessibility metadata, relevant records, and other (e.g., notes). The names of the items of the RA schema derive from their database schema (ARKIS) and they are denoted in both Swedish and English (the original fields are in the Swedish language).

The overall methodological process followed for the mapping is shown in Fig. 2. It is initiated by extracting all Classes, Data Type Properties and Object Properties of the Records-in-Contexts ontology (RiC-O), as well as the main Classes and Attributes of the RA schema. The classes and properties of RA were documented accompanied by a field description in English and Swedish. The list of fields from the RA schema involved in the mapping is not exhaustive but contains the necessary concepts to describe archival records and the most important entities and related actions to put them in context. The mapping matrix (set of spreadsheets) is then prepared, organised in a Class-to-Class mapping and an Object Properties-Relations mapping, the latter including Domain and Range to express the directive (or symmetric if this is the case) relations.

The mapping between RA and RiC-O was performed by first identifying which RiC-O classes align well with the RA classes, so a mapping on Class level. After the Class level mapping, the identification of all related Object Properties and Relations was performed, keeping the relations in RiC-O that bear meaning for RA and can be reused to express context. This process leads to the creation of a set of class mappings and a set of object properties and relations mapping, the latter following an RDF triple structure in principle, e.g.,: **Document** *includesOrIncluded* in **Series**, **ReferenceCode** *isOrWasIdentifierOf* **ArchiveFond**, or **Agent** *isOrWasManagerOf* **ArchiveFond**, where the first part of the triple is the Domain of the relation, the second part is the relation which connects the subject and the object of the triple, and the third part is the Range, so the passive part of the triple. The results of the mapping process described were evaluated by a group of experts during an organised Workshop session at the Swedish National Archives headquarters, where the produced mappings were shared with the group of experts in order to assess their suitability, level of alignment, and integrate feedback in the loop.

3.2 LLM (GPT-4o and GPT-4.5) Experiments

The two-fold experiment of this study includes the following two cases/scenarios where GPT-4o (and GPT-4.5 for the first one) was asked to perform.

i) Given the full, expert-validated dataset of the mapping between RA and RiC-O, and a custom archival institution schema, can the LLM perform the mapping between RiC-O and the new schema effectively?

ii) Given the full, expert-validated dataset of the mapping between RA and RiC-O and a few human-evaluated examples of recommended mapping between some RA elements and elements from other standard schemas (e.g., Linked Open Vocabularies - LOV), can the LLM

output meaningful recommendations to more vocabularies, according to a concept-in-word context?

The experiment set-up for the two scenarios is shown on Fig. 3.

In the first case, the aim is to test the performance of the model in undertaking the role of a human expert in order to semantically align a given, custom archival institution schema with RiC-O without any prior knowledge (zero-shot) and with having the full expert-validated mapping (RA and RiC-O) as an example to anchor the process. In the second case, the aim is to test the ability of the model to identify relevant concepts in other schemas within proximal domains (e.g., Cultural Heritage, Archives) but also across other domains (e.g., Geospatial information, Arts, etc.) and recommend semantically interesting cases for re-use in metadata description. Apart from aligning part of the RA schema (ARKIS) with the Records-in-Contexts Ontology (RiC-O), other ontologies in the Cultural Heritage (CH) domain could potentially be reused to enhance representation expressivity in cases where RiC-O classes do not live up to the same level of granularity. For the most generic descriptions, the Swedish National Archives already reuses fields from schema.org, such as schema:ArchiveComponent, schema:Creator, or schema:Identifier, however, for deeper level of granularity there is currently no reuse from other ontologies in the Archival or Cultural Heritage domain apart from RiC-O. RiC-O is an ontology which emphasises and offers a variety of relations and attributes for archival record description, but it does not primarily focus on a rich representation of an artifact itself. Other ontologies, for instance the Context Description Ontology (ArCo network) (cdesc), the Denotative Description Ontology (ArCo network) (ddesc), and the LinkedGeoData ontology (lgdo) could lend some of their classes to the RA schema. All these ontologies are also available in the Linked Open

Vocabularies (LOV) repository. In other words, the model is expected to make recommendations on LOV re-use for concepts which can be more accurately expressed by borrowing concepts from other knowledge domain schemas, thus potentially facilitating an ontological multi-domain re-use.

Prompt preparation

The 1st experiment was performed on OpenAI's GTP-4o and GPT-4.5. The 2nd experiment was performed on GPT-4o. During the 1st experiment, the model was asked to conduct the mapping between a custom archival institution schema and RiC-O with and without prior knowledge. During the 2nd experiment, the model was asked to consider a dataset of validated mapping examples between some RA elements and ones borrowed from the Linked Open Vocabularies (LOV) ecosystem in order to similarly suggest more cases for re-use both within the CH domain and beyond. First the model was asked to make further LOV recommendations on the sample given (Appendix 3) and then extend the recommendations in a similar way to the full RA dataset (Appendix 1).

Zero-shot trial

For the first experiment, the model was initially asked if it is familiar with RiC-O and whether, given a custom data element set from an archival institution, it can map the concepts and make it compliant with RiC-O. In this case, no more contextual information was provided. This task was performed on both GPT-4o and GPT-4.5.

Informed trial

Following the zero-shot trial, the model was then asked to perform the same process, given prior

knowledge to consider. In this case, the full, expert-validated mapping between RA and RiC-O was given as input, including both class–to-class and properties mapping. This task was performed on both GPT-4o and GPT-4.5.

Evaluation technique

The outputs of experiment 1 were initially human-evaluated by the three authors, since they have manually performed this process before and have the experience necessary for a preliminary (not case-specific) judgement. More specifically, the outputs were evaluated by two experts with prior experience in semantic alignment processes, and one expert with prior experience in archival metadata semantic alignment. The initial evaluation was conducted independently, and conflicting cases were then resolved during a post-evaluation discussion. Each mapping recommendation was classified as True Positive (TP), False Positive (FP), False Negative (FN), or True Negative (TN), according to predefined evaluation criteria and examining both the semantic relevance of the proposed mapping and its conformity with RiC-O modelling practices. The commonalities two archival metadata schemas may share do not necessarily imply the same mapping but rather serve as intuition offered to the LLM to perform the task better. So, the previously provided alignment (RA to RiC-O) serves as a truth measure while the human evaluation, which is performed upon the produced results of the LLM, is based mainly on personal human domain expertise. The metrics used to evaluate the outputs were accuracy, precision, recall, and F1 score, which were separately calculated for each experiment set-up (for the zero-shot and informed trial of experiment 1. In practice, how this task was performed was to add one more column in the output files, denoting for each mapping if it is acceptable or not. From the resulting confusion matrix, accuracy, precision, recall, and F1 score were calculated according to their formulas (Fig. 4).

The confusion matrix values in the experiment case are as follows:

- **TP:** A correct suggestion was given by the LLM
- **FP:** An incorrect suggestion was given by the LLM
- **TN:** No suggestion was given and human experts did not provide any suggestions either
- **FN:** No suggestion was given by the LLM, but there is a match in reality

Apart from the semantic meaning of the mapping rows, the LLM-generated suggestion was also tested for "hallucinations", e.g., ensuring that the suggested concept truly exists and is part of the ontology or vocabulary, and not superficially manufactured to fit the match. For experiment 2, due to a more flexible nature of this task (recommending appropriate vocabulary fields within and across the CH domain for each given data element), the human evaluation of the output was performed by distinguishing the cases where the LLM proposed acceptable recommendations, and the cases for which it proposed non-acceptable recommendations (in which there is the distinction of True Negative and False Negative cases). Once more, the recommendations were all checked for "hallucinations", ensuring that each of them exists and belongs to the schema of a valid Linked Open Vocabulary.

In the next Section, the results of the described processes are presented. The purpose of the experiments is not to replace human involvement in such a process, but rather to augment it in a more time-efficient and effective way and help fine-tune domain-specific applications by providing insights and areas for improvement in this regard in the form of informed set-up and methodology. The datasets used as groundwork in the prompting interaction are shown in the Appendices, where Appendices 1 and 2 include the full RA - RiC-O mapping, and Appendix 3

includes the validated mapping examples (sample) between RA and other schemas in LOV. Appendix 4 includes the prompting template used to interact with GPT-4o in both scenarios and GPT-4.5 in the 1st scenario. The following Section includes the key prompting interaction results and detailed process outcomes.

4. Results

The results of the two-fold experiment using the prompt interface of GPT-4o and GPT-4.5 and the human evaluation for each respectively are presented in this Section.

4.1 Experiment 1 - ALD: From one Archive to more

As described earlier, the first experiment concerns the mapping between Records-in-Contexts (RiC-O) and a given, custom data element list from an archival institution. The rationale is to first conduct the mapping without prior knowledge (zero-shot) and then ask the model to repeat the procedure by initially providing the full, expert-validated mapping between RA and RiC-O for context (RA 1 and RA 2). This trial was attempted at both GPT-4o and GPT-4.5. The prompt interaction and the analysis were relatively fast, taking a few seconds to go through the input given in tabular form. The input tables were prepared accordingly prior to the experiments in order to present the information in a well-organised and clear form. In some cases, some intermediate clarifications were required but overall, the model seemed to clearly grasp the instructions given. The output seemed promising at first sight, but several shortcomings were revealed during the evaluation process. During the zero-shot trial in both GPT-4o and GTP-4.5, the model was provided only with the custom archival institution data elements list that were to be included in the mapping process to RiC-O. The full results of this process, including human

evaluation, are available in Appendix 6. During the second phase of the experiment, the full, expert-validated mapping between RiC-O and RA was provided as well, to check whether the results of the mapping would be improved. The full matrix of this process can be found in Appendix 7.

Table 1 shows an excerpt of the model output for the case with GPT-4o and considering the mapping with RA. The columns of the output include the custom archival institution schema element and their short description, a column with the initial mapping the model proposed, a column with Notes and Comments that the model assigned, and a column with the model's own revised mapping based on the validated mapping as benchmark. The last column Evaluation (Human) is the extra column added by the human evaluators (authors) during the evaluation process. The evaluation result follows the rules described in the Method Section; TP for a correct suggestion, FP for an incorrect suggestion, FN for no given suggestion when one exists, and TN for no given suggestion when one does not exist.

Following the manual, human evaluation, the confusion matrix for each use case was calculated. The use cases are as follows:

- **GPT-4o (zero-shot):** No prior knowledge mapping to RiC-O
- **GPT-4o (RA 1)**: Starting from zero-shot and then directly inputting the RA and ask it to repeat
- **GPT-4o (RA 2)**: Start directly with RA- took a few extra steps to get the final mapping
- **GPT-4.5 (zero-shot):** No prior knowledge mapping to RiC-O
- **GPT-4.5 (RA as validation):** In this case the mapping to RA was used as a strict

validation anchor for the model

The results of the confusion matrix calculation for all case scenarios are shown in Table 2.

At this point, having calculated the True Positive, True Negative, False Positive, and False Negative cases, the metrics for each use case were calculated and are presented in Table 3. The maximum accuracy (65%) was achieved with GPT-4o in the second informed trial with RA. The last column shows the percentage of hallucinations per trial, meaning the number of times out of the total that the model proposed a superficial (non-existent) class or property for the mapping.

Similarly, the precision, recall, and F1-score percentages are presented, giving a better overview of the output score distribution and cases. The highest accuracy (65%) was reached by GPT-4o in the second trial with informed input from the RA mapping to RiC-O (RA 2), while the lowest was GPT-4.5's 31.34% when it used the RA-RiC-O mapping as validator for its own mapping. This locked the model's flexible potential to choose an appropriate mapping and forcefully tried to impose the input mapping on the new examples regardless of better, existing matches, which led to the lowest accuracy, although for the same reason the hallucination cases dropped drastically. All the average accuracies of this experiment had a hallucination percentage ranging between 26-31.8%.

4.2 Experiment 2 - LOV Ontological Multi- (or Inter-) Relation

The second experiment concerns the ability of the model to make recommendations for re-use from the Linked Open Vocabularies (LOV) ecosystem, considering both adjacent domains (e.g., CH, Archives), but also other domains, if these are useful for enriched metadata description.

The input given to GPT-4o for this task were both the RA to RiC-O mapping, and a table of 15 human-validated examples of RA data elements and their extended mapping to several LOV vocabularies (shown in Appendix 3). The model was then asked to extend the mapping for the rest of the elements in the same way, given the full RA schema list. An excerpt of the output mapping by GPT-4o is shown in Table 4. The table includes the Swedish National Archives (RA) element list, description, and the equivalent mapping to RiC-O, followed by two columns generated by GPT-4o, which show its provided recommendations for LOV re-use both in the CH and other domains respectively. The last 3 columns of the table were added afterwards, and include the manual, human evaluation of these results. The full Experiment 2 results, including human evaluation, are available in Appendix 8.

As described earlier, the output was checked for hallucinations or random correspondence, and the output recommendations were judged as acceptable or non-valid ones. The model did not provide output for several examples, denoted as “no recommendations”, which were separated in True Negative and False Negative cases. The reason for this is that we would not like to forcefully receive recommendations if truly there are no good matches, so true negative cases support the model’s overall performance. It is meaningless to judge the result in the same way as Experiment 1 as, depending on the context, the choice of vocabulary can change and has many levels of flexibility. However, what is important is that the mapping recommendations make sense and can be directly borrowed and trusted. The percentages of the evaluation of the results are presented in Table 5.

The total cases were 69, out of which 37 cases were given no recommendations, and 32 cases were given acceptable recommendations. The 32 acceptable recommendation cases are divided into cases where both acceptable and non-acceptable recommendations were given (71.8%), cases where exclusively acceptable recommendations were given (31.25 %), and cases where exclusively non-acceptable recommendations were given (25%). Similarly, the 37 cases where no recommendations were provided by the model are divided into TN cases (43.24%) and FN cases (56.75%).

Overall, the model recommended the re-use of 21 vocabularies, apart from RiC-O. The different domains in which the suggested vocabularies belong are shown in Fig. 5, including the vocabulary names in the label. These graphs were generated by the model, right after the analysis, at the request of the authors. As can be seen in the Figure, apart from the Cultural Heritage and Archival domain, we have recommendations from Web Metadata, Libraries and Bibliography, Geodata, Provenance and Process, Media, and Organisations.

Similarly, Fig. 6 shows the number of occurrences in the recommendations generated per vocabulary. For instance, classes and properties from cidoc-crm (crm) were suggested about 65 times, elements from bibo about 10 times, elements from the Europeana data model (edm) about 55 times, and ones from geo about 8 times.

The visualisation of Fig. 7 shows how each element of the RA schema is related to each recommended LOV, capturing also the most basic hierarchical relations among them. For instance, the class “Arkivinstitution” (Archival institution) is mapped to the LOV classes foaf:Organization, edm:Agent, org:Organization, gr:BusinessEntity, and schema:Organization.

Similarly, “Fotografi” (Photograph) is mapped to vcard:Photo, media:Image, edm:ProvidedCHO, crm:E38_Image, and schema:ImageObject. “Serie” (Archive series) is mapped to bibo:Collection, schema:Collection, and cdesc: ArchivalSeries, while “Skapad” (Created) is mapped to schema:dateCreated, time:TemporalEntity, dcterms:date, crm:P4_has_time-span. “Koordinater” (Coordinates) are aligned with geo:lat, geo:long, wgs84:pos:lat, wgs84:pos:long, and “Arkivhistorik” (Archive history) is aligned with skos:note, bio:Event, prov:Activity, and schema:Event.

In addition, a specific mapping to DCAT was requested, showing how the current schema could aim for compliance. A graph on how dcat as an overarching, higher-level metadata descriptor corresponds to RA is shown in Fig. 8, where (where meaningful) some RA elements are expressed in dcat terms.
In blue colour are the RA elements, and in green and yellow the dcat and dct terms respectively. Dcat:Dataset, dcat:Distribution, dcat:accessURL, dct:rights, dct: description, dct:publisher, dct:issued, dct:title, dct:identifier, dct:spatial, dct:modified and dct:temporal are used in this representation. The granularity level of this correspondence pertains to metadata which allows the data to be described appropriately through a catalog or data portal which uses a DCAT Application Profile (DCAT-AP).

5. Discussion

The purpose of this study was to test to what extent a user can talk an LLM, such as GPT, into performing (or assisting) the process of semantic alignment between a custom and a standardised schema of a given domain (in this case Archives), and to what extent can an LLM provide useful recommendations for vocabulary re-use (in this case LOV).

5.1 Experiment 1: ALD - From one Archive to more

As far as the alignment of a given custom archival institution schema to RiC-O is concerned, there is understandably not one single approach that is correct. The manual mapping process is an iterative and most often time-consuming process, which requires expert feedback and careful evaluation and alignment to each context. The process of manually mapping elements between two schemas remains highly subjective and is dependent on several factors which might affect the outcome. This is not necessarily a negative consequence as schema alignment needs to also be flexible and easily adaptable to each case scenario, but it introduces several parameters which need to be carefully assessed and considered. The organisation's mission, structural focus, legacy systems (if these exist), and infrastructure, are all factors which can force the mapping process to deviate significantly across applications of the same standard schema in various contexts or occasions. Different mapping approaches might work equally well for their intended purpose, provided that they are carefully corresponding to the institution's needs and infrastructural organisation. In this light, Experiment 1's intention is not to blindly trust an LLM to produce a well-tailored semantic alignment, but rather to make the process more time-efficient and, if possible, effective, if it is given enough context to analyse.

While from the results of the experiment it becomes obvious that, at least through prompting interaction, GPT is not ready to undertake this task as well as one would think, it is still an action worth investigating its limits and maximum performance boundaries should some guidelines allow for better results with time. It is noteworthy that there was no significant improvement from zero-shot to informed trial (with the RA-RiC-O mapping as input). The quite high percentage of hallucinations was reduced, but not to an extent that might strongly imply causality. GPT-4.5 generated considerably less hallucinations than GPT-4o but scored less in

accuracy and all other metrics. A noticeable tendency was also that at zero-shot learning, the model tended to only use properties for the mapping (and making up quite some of them), but after the informed trial, it started (correctly so) identifying direct class-to-class mapping as well. What surprisingly did not work well either was that even when the RA-RiC-O mapping was used as an anchor of validation for the model, the model did not change its mind to give a better match alternative. In addition, in both set-ups, there were 0 True Negatives (TN), showing that GPT-4o (same as for Experiment 2), was rather reluctant to not give any result, preferring a wrong answer instead. If one aims to classify the produced errors, most of them fall into one of the following error types: structural errors, hallucinated elements, semantic drift, and unidentified or missing matches. Structural errors refer to wrong domain-range relation, for instance, Place of Deposit mapped to rico:Place instead of institution-based relation rico:heldBy. Hallucinated elements refer to the LLM superficially producing a result to satisfy the case, as reported in Table 1, for example rico:hasGender or rico:hasAccessRestriction. Semantic drift refers to the cases where the LLM selects a concept which could be related but not of the equivalent level of granularity (e.g., Document mapped to CreativeWork instead of a more precise archival concept). Lastly, unidentified or missing matches refer to cases where a real, suitable match was omitted by the LLM (false negative). These failures seem to or could stem from ambiguity in short field descriptions, the tendency of a general-purpose LLM to overgeneralise semantically related concepts, or the tendency to forcefully generate an answer (even if this is wrong) rather than abstaining.

Apart from the failures and pitfalls noticed, experiment 1 shows that at least from the viewpoint of saving time in the alignment process, the model can in very little time (few seconds) give over 60% good fits, which means the human might need to go through the output for validation and

context-specific needs satisfaction but already having something to start from. Human intervention cannot be eliminated and the results with high hallucination percentage are not to be trusted blindly but rather allow for a first drafting of the mapping, providing also some explanation for each selection. Fine-tuning this process in an even more controlled environment could potentially bring forth better results. Compared to established ontology matching systems such as AML or LogMap, the accuracy and results of this experiment suggest that prompt-based approaches are not yet ready to become standalone replacements, but rather that their value and their role lies in supporting human experts in this process.

5.2 Experiment 2: LOV Ontological Multi- (or Inter-) Relation

The initial alignment of the Swedish National Archives schema to Records-in-Contexts Ontology (Maratsi, Haskiya, et al., 2025). aimed to bring forth the advantages associated with standardisation and the transition to Linked Data and Archival Linked Data. RiC-O's focus concept on archival resource description and relations to express actions performed by core entities in an archival setting make it a suitable candidate for standardised schema alignment. However, as far as describing the schema on more granularity levels is concerned, an alignment plan involving other ontologies in the archival or Cultural Heritage domain could be considered in order to achieve improved expressivity for the variety of different types of archival records offered by an archival institution (e.g., photographs, maps, paintings, video recordings, films, and other) which could require a more tailored approach, and this is the motivation behind Experiment 2. In a real scenario, the concerns around using multiple ontologies to describe one's schema, such as possible lack of ontology maintenance, resources, documentation, or technical support, are not to be neglected, however, to avoid customisation and keep it to a minimum, reuse of standard and well-maintained schemas is an ever-common practice.

It is worth mentioning that Experiment 2 had next-to-no hallucination cases. All recommendations from the vocabularies actually exist and are legitimate, meaning that, at least in the context of this small-scale study, the results for this experiment could be trusted from that aspect. However, the model produced repetitive results in some cases, possibly due to bias from the given input. The repetition for some cases is justifiable due to the lack of granularity of the given concepts, but other times the result is very poor when it is obvious that a match exists, the latter denoted as False Negatives (FN) during evaluation. For other cases and concepts, finding a direct match is not meaningful (e.g., mapping to more general purpose vocabularies), however, what was observed is that when the model does not find a match, in many cases, it simply repeats the same pattern as one of the previous examples, even if it is blatantly wrong, instead of not giving any results (True Negative -TN).

Overall, experiment 2 could be used to identify potential cross-domain use and give a good (also graphical) overview of the overlapping domains and intersections in a trusted way. The visuals generated are intuitive and worth consulting for a good overview of the output. The rest of the process is not yet very promising in the quality of results; even though in this experiment there are almost no hallucinations, which is good, the otherwise not so good performance is on one side expected as the model is faced with the challenging task of deciding a proper match and granularity level for cases it does not have so much contextual information (except a few words description and title). On the other hand, all those False Negative (FN) cases could possibly be populated with maybe generic but correct recommendations, seeing as they clearly exist but the model did not find any.

5.3 Limitations and Future Directions

Some limitations of these experiments are due to limitations introduced by the model itself, and others fall under the responsibility of methodological dos and don'ts. Model limitations can be introduced due to short context windows, not allowing the model to sufficiently combine knowledge and input given all at the same time, without "forgetting" prior information, thus failing to consider all desired parameters to reach a decision. This entails the risk of requiring more correction steps during the prompting interaction, risking in its turn the soundness of the process, should one enter a loop of repetitive errors and propagation of misunderstanding. The input should balance clear instructions, well- structured input, and contextual information. As mentioned previously, the black-box effect is considerably high, causing insecurity when it comes to trusting the output. If a human is to spend as much time validating the output as it would take to perform the task themselves, the assistance loses any value. However, some cautious optimism might be allowed by the results, at least considering that the evaluation process for both experiments took very little time compared to the manual mapping process of Maratsi, Haskiya, et al., (2025) or similar.

In addition, methodological and experiment limitations include the flexibility in human interpretation of the results, which might affect the evaluation of the output as well. Semantic alignment in real-case applications is a process very highly dependent on case-specific context and mapping recommendations should be considered auxiliary and not written in stone. Moreover, different approaches in prompting techniques can most likely produce different results, the divergence of which is not easy to measure in this environment. Regardless of how cautious and methodologically sound the process might seem, the model's nature will always

introduce entropy in the output, making it extremely challenging to set strict boundaries which allow for both improved performance and minimum intervention.

For the reasons described, future directions of this study may involve local, domain specific LLM training on clean data in order to repeat the experiments and see the difference. The potential and the outskirt limits of mere prompting interaction for this purpose is perhaps not exhausted and further experimentation may provide improved insights and guidelines to boost the process's performance. Moreover, the present study evaluates a limited set of prompt formulations, indicating that different prompting strategies, role-based prompting, or retrieval-augmented prompting could produce different results, so prompt sensitivity remains a subject that should be further investigated in future work. Using and assessing alternative prompt formulations would be an important aspect to test, including the trial of several other models to compare the results of, e.g., LLAMA 3. Zero-shot, few-shot, or chain-of-thought prompting strategies could also be tested and compared. GPT-4o was used at the time since it constitutes a common choice among both technically and non-technically skilled groups of people, it is easily accessible through the user interface, and constitutes a common point of reference, so while it might not constitute the most suitable LLM to perform the given task (others could possible perform better, e.g., LLAMA), it was still deemed as a worthwhile model for the experiments for the aforementioned reasons. However, to define a highly configurable environment with more trusted output, training and fine-tuning a model in context and domain-specific setting would be a valuable pathway to walk and experiment with. For instance, Retrieval-Augmented Generation (RAG) techniques could be utilised to enhance the output and provide more tailored results.

Finally, this study experiments solely with prompt-based LLM reasoning so another possible research direction to be taken into consideration concerns hybrid ontology matching approaches,

such as combining LLMs with already established ontology matching techniques, or reasoning engines to try and reduce hallucinations and overall improve the quality of the produced mapping. Future evaluations could also integrate graph-based similarity metrics in addition to the presented confusion matrix ones of this study. For instance, graph edit distance, neighbourhood similarity, or alignment consistency scores could provide additional insight into the mapping quality.

6. Conclusion

Some common pitfalls for all set-ups were identified. The output during the trials in some cases showed some predictability but many times the black box effect takes over. In addition, it is impossible to go one step back and receive exactly the same results as before. Even though this is expected from a highly probabilistic model, in cases with strict benchmarking and guidance, the result would be expected to be a bit more coherent and predictable. This is not a problem if it only concerns the flexibility of vocabulary use or selection of mapping according to case-specific examples, and if the performance shows signs of normalised behaviour. However, there is still high uncertainty about trusting that “this is the best the model can do” or “this is the worst the model can do”. This form of consistency could allow for a better quantification of the results and standardisation of the process to improve performance, and, regardless of the ability or not to reach a high accuracy level, it could produce much more consistently trusted outputs, maximising what this method can potentially offer. The findings of this study suggest that current prompt-based LLM approaches for semantic alignment are best viewed as complementary human-assistance mechanisms, rather than automated, standalone practices.

Hybrid architectures combining LLMs, ontology matching systems, or graph-based and reasoning frameworks remain domains to be further investigated and evaluated.

References

A global network on sharing cultural heritage. (n.d.). OpenGLAM. https://openglam.org/

Bikakis, A., Hyvönen, E., Jean, S., Markhoff, B., & Mosca, A. (2021). Editorial: Special issue on Semantic Web for cultural heritage. Semantic Web, 12(2), 163–167. https://doi.org/10.3233/sw-210425

Bountouri, L., & Gergatsoulis, M. (2011). The semantic mapping of archival metadata to the CIDOC CRM ontology. Journal of Archival Organization, 9(3–4), 174–207. https://doi.org/10.1080/15332748.2011.650124

Chen, S. (2019). Semantic enrichment of linked archival materials. Knowledge Organization, 46(7), 530–547. https://doi.org/10.5771/0943-7444-2019-7-530

Cigliano, A., & Fallucchi, F. (2025). The convergence of open data, linked data, ontologies, and large language models: Enabling next-generation knowledge systems. In Communications in Computer and Information Science (pp. 197–213). https://doi.org/10.1007/978-3-031-81974-2_17

Dobreski, B., Park, J., Leathers, A., & Qin, J. (2020). Remodeling archival metadata descriptions for linked archives. International Conference on Dublin Core and Metadata Applications, 1–11. https://dcpapers.dublincore.org/pubs/article/download/4223/2417

Doumanas, D., Bouchouras, G., Soularidis, A., Kotis, K., & Vouros, G. (2024). From human- to LLM-centered collaborative ontology engineering. Applied Ontology, 19(4), 334–367. https://doi.org/10.1177/15705838241305067

Encoded Archival Description (EAD). (n.d.). Society of American Archivists. https://www2.archivists.org/groups/technical-subcommittee-on-encoded-archival-standards-ts-eas/encoded-archival-description-ead

GO FAIR initiative. (2022, January 21). FAIR principles – GO FAIR. https://www.go-fair.org/fair-principles/

Gracy, K. F. (2014). Archival description and linked data: A preliminary study of opportunities and implementation challenges. Archival Science, 15(3), 239–294. https://doi.org/10.1007/s10502-014-9216-2

Gracy, K. F. (2017). Enriching and enhancing moving images with linked data. Journal of Documentation, 74(2), 354–371. https://doi.org/10.1108/jd-07-2017-0106

Hawkins, A. (2021). Archives, linked data and the digital humanities: Increasing access to digitised and born-digital archives via the semantic web. Archival Science, 22(3), 319–344. https://doi.org/10.1007/s10502-021-09381-0

Hawkins, A., & University of Liverpool. (2021). Advocating for linked archives: The benefits to users of archival linked data. University of Liverpool.

Hoseini, S., Burgdorf, A., Paulus, A., Meisen, T., Quix, C., & Pomp, A. (2025). Challenges and opportunities of LLM-augmented semantic model creation for dataspaces. In Lecture Notes in Computer Science (pp. 183–200). https://doi.org/10.1007/978-3-031-78955-7_17

ICA. (2025, February 12). ISAD(G): General International Standard Archival Description – Second edition. https://www.ica.org/resource/isadg-general-international-standard-archival-description-second-edition/

International Council on Archives Records in Contexts Ontology (ICA RiC-O) version 1.0.2. (n.d.). https://www.ica.org/standards/RiC/RiC-O_1-0-2.html

Koch, I., Lopes, C. T., & Ribeiro, C. (2023). Moving from ISAD(G) to a CIDOC CRM-based linked data model in the Portuguese archives. Journal on Computing and Cultural Heritage, 16(4), 1–21. https://doi.org/10.1145/3605910

Linked Open Vocabularies (LOV). (n.d.). https://lov.linkeddata.es/dataset/lov/

LinkedData – W3C Wiki. (n.d.). https://www.w3.org/wiki/LinkedData#:~:text=The%20term%20Linked%20Data%20refers,can%20look%20up%20those%20names.

Maratsi, M. I., Ahmed, U., Alexopoulos, C., Charalabidis, Y., & Polini, A. (2024). Towards cross-domain linking of data: A semantic mapping of cultural heritage ontologies. https://doi.org/10.1145/3657054.3657077

Maratsi, M. I., Gialoussi, N., Alexopoulos, C., & Charalabidis, Y. (2025). A proposed methodology for sub-ontology development in comprehensive scientific investigation methods

and tooling. In Communications in Computer and Information Science (pp. 28–43). https://doi.org/10.1007/978-3-031-81974-2_3

Maratsi, M. I., Haskiya, D., Berggren, M., Charalabidis, Y., & Alexopoulos, C. (2025). Towards open archival linked data (ALD): The case of Swedish National Archives. In Lecture Notes in Computer Science (pp. 3–23). https://doi.org/10.1007/978-3-031-94578-6_1

Mazzini, S., & Ricci, F. (2011). EAC-CPF ontology and linked archival data. (pp. 72–81). http://ceur-ws.org/Vol-801/paper6.pdf

Ministère de la Culture de France, Service interministériel des Archives de France. (n.d.). FranceArchives. https://francearchives.gouv.fr/en

Nguyen, L., Barcelos, E., French, R., Wu, Y. (2026). KROMA: Ontology Matching with Knowledge Retrieval and Large Language Models. In: Garijo, D., et al. The Semantic Web – ISWC 2025. ISWC 2025. Lecture Notes in Computer Science, vol 16140. Springer, Cham. https://doi.org/10.1007/978-3-032-09527-5_34

Osman, I., Pileggi, S. F., & Yahia, S. B. (2024). Uncertainty in automated ontology matching: Lessons from an empirical evaluation. Applied Sciences, 14(11), 4679. https://doi.org/10.3390/app14114679

Overview for “linked-open-data.” (n.d.). Europeana PRO. https://pro.europeana.eu/itemtype/linked-open-data

Pan, X., Jacco, V. O., Victor, D. B., & Huang, Z. (2024). A RAG approach for generating competency questions in ontology engineering. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2409.08820

Riksarkivet. (n.d.). Start – söktjänstens startsida – Riksarkivet – Sök i arkiven. https://sok.riksarkivet.se/

Rinaldi, A.M., Russo, C. & Tommasino, C. A semantic approach for cultural heritage ontology matching and integration based on textual and multimedia information. Soft Computing 29, 1019–1034 (2025). https://doi.org/10.1007/s00500-025-10517-y

Soni, P. (2025, October 7). Confusion matrix, precision, and recall. Train in Data's Blog. https://www.blog.trainindata.com/confusion-matrix-precision-and-recall/

List of Tables

Table 1. An excerpt of the mapping output

Custom Archives Schema Element	Description (Summary)	RiC-O Equivalent Class / Property	Notes	Revised RiC-O Mapping (based on RA) (GPT-4o)	Evaluation (Human)
Level	The hierarchical level of the material	rico:RecordResource + rico:hasLevel	RiC-O uses rico:hasOrHadLevelOfDescription as a datatype property	rico:RecordResource + rico:hasLevel	TP
Legal Status	Public Record or not (under law)	rico:hasLegalStatus	A datatype or object property; Legal status can also be an instance of rico:LegalStatus	rico:hasLegalStatus	TP
Language	Language(s) of the record	rico:hasLanguage	Object property linking to rico:Language	rico:hasLanguage	TP
Reference	Unique identifier to link records	rico:hasIdentifier	Identifier is often modeled as a literal, possibly as an instance of rico:Identifier	rico:hasIdentifier	TP
Former Reference (Dept)	Previous ID by originating body	rico:hasOrHadIdentifier	Used for historical identifiers	rico:hasOrHadIdentifier	TP
Map Designation	Map series info for maps	rico:hasIdentifier / rico:hasTitle	If used for structured series title or	rico:hasIdentifier / rico:hasTitle	FN

			map ID		
Place of Deposit	Repository where records are held	rico:heldBy	Links to rico:Agent (institution)	rico:Place	FP
Access Conditions	Conditions affecting access	rico:hasAccessRestriction	Could use rico:hasAccessCondition or link to rico:Rule	rico:hasAccessRestriction	FP (Hallucination)
Restrictions on Use	Use or reproduction restrictions	rico:hasUsageRestriction	May be a note or Rule	rico:hasUsageRestriction	FP (Hallucination)
Custodial History	Description of prior custody	rico:hasCustodialHistory	Object or literal property	rico:hasCustodialHistory	FP (Hallucination)
Map Designation	Map series info for maps	rico:hasIdentifier / rico:hasTitle	If used for structured series title or map ID	rico:hasIdentifier / rico:hasTitle	FN
Former Reference (Pro)	Former TNA identifier	rico:hasOrHadIdentifier	For previous/provenance-based identifiers	rico:hasOrHadIdentifier	TP
Publication Note	Reference to published finding aids	rico:hasBibliographicReference	Modeled with rico:Bibliography	rico:hasBibliographicReference	FP
Administrative, Bibliographical Background	History of the record creator	rico:hasHistory	Linked to rico:Agent	rico:hasHistory	FP
Gender Indicator	Gender of the individual	rico:hasGender	Can use literal or vocabulary	rico:hasGender	FP (Hallucination)
Index Terms: Subjects	Subject terms	rico:hasSubject	Could be Topic or Thing	rico:hasSubject	TP

Table 2. The confusion matrix for all use cases

Confusion Matrix for all Use Cases	
TP	**FN**
GPT-4o (zero-shot): 34	**GPT-4o (zero-shot):** 6
GPT-4o (RA 1): 35	**GPT-4o (RA 1):** 7
GPT-4o (RA 2): 26	**GPT-4o (RA 2):** 1
GPT-4.5 (zero-shot): 14	**GPT-4.5 (zero-shot):** 35
GPT-4.5 (RA as validation): 13	**GPT-4.5 (RA as validation):** 27
FP	**TN**
GPT-4o (zero-shot): 26	**GPT-4o (zero-shot):** 0
GPT-4o (RA 1): 23	**GPT-4o (RA 1):** 0
GPT-4o (RA 2): 13	**GPT-4o (RA 2):** 0
GPT-4.5 (zero-shot): 6	**GPT-4.5 (zero-shot):** 13
GPT-4.5 (RA as validation): 19	**GPT-4.5 (RA as validation):** 8

Table 3. The evaluation metrics for all cases

Trial / Metrics	Accuracy	Precision	Recall	F1-Score	Hallucination Cases
GPT-4o (zero-shot)	51.5%	56.66	85%	67.9%	31.8%
GPT-4o (RA 1)	53.84%	60.34%	83.33%	70%	26%
GPT-4o (RA 2)	65%	66.66%	96.29%	78.77%	30%
GPT-4.5 (zero-shot)	39.7%	70%	28.5%	40%	16%
GPT-4.5 (RA as validation)	31.34%	40%	32.5%	35.86%	0

Table 4. An excerpt of the LOV recommendations output

RA Element	Description	RiC-O Mapping	LOV Recommendations (GPT-4o CH)	LOV Recommendations (GPT-4o cross-domain)	Hallucinations (Human Evaluation)	Acceptable Recommendations (Human Evaluation)	Non-valid Recommendations (Human Evaluation)
Arkiv	Archive (fond)	schema:ArchiveComponent, rico:RecordSet	crm:E78_Curated_Holding, edm:Aggregation	crm:E78_Curated_Holding, edm:Aggregation, prov:Entity, schema:CreativeWork, void:Dataset	0	crm:E78_Collection, edm:Aggregation, schema:CreativeWork, void:Dataset, prov:Entity, edm:Collection, frbr:Manifestation, bibo:Collection	0
Typ	Archive type	rico:RecordSetType	crm:E78_Curated_Holding, dc:type, edm:Aggregation, rdau:P60047, skos:Concept	crm:E78_Curated_Holding, dc:type, edm:Aggregation, prov:Entity, rdau:P60047, schema:CreativeWork, schema:additionalType, skos:Concept, void:Dataset	0	rdau:P60047 (has type of agent), schema:additionalType	crm:E78_Curated_Holding, skos:Concept, prov:Entity, void:Dataset, schema:CreativeWork, edm:Aggregation

Typ	Types of geographical units and divisions	rico:PlaceType	dc:type, rdau:P60047, skos:Concept	dc:type, rdau:P60047, schema:additionalType, skos:Concept	0	schema:additionalType, dc:type, rdau:P60047	skos:Concept
Arkivbildare/ upphov	Archivist	schema:creator, rico:Agent	foaf:Person, schema:Person	foaf:Person, schema:Person, vcard:Individual	0	foaf:Person, vcard:Individual, schema:Person	0
Arkivinstitution	Archival institution	rico:CorporateBody	edm:Agent, foaf:Organization, schema:Organization	edm:Agent, foaf:Organization, gr:BusinessEntity, org:Organization, schema:Organization	0	edm:Agent, foaf:Organization, gr:BusinessEntity, org:Organization, schema:Organization	0
Dokument	Document	rico:Record	bibo:Document, ddesc:Document, foaf:Document	bibo:Document, ddesc:Document, foaf:Document, schema:CreativeWork	ddesc:Document	bibo:Document, foaf:Document	schema:CreativeWork
Fotografi	Photograph	rico:Record	crm:E38_Image, edm:ProvidedCHO	crm:E38_Image, edm:ProvidedCHO, media:Image, schema:ImageObject, vcard:Photo	media:Image	crm:E38_Image, edm:ProvidedCHO, schema:ImageObject, vcard:Photo	0

Titel	Archive title	rico:Name	crm:E78_Curated_Holding, dcterms:title, edm:Aggregation, skos:prefLabel	crm:E78_Curated_Holding, dcterms:title, edm:Aggregation, prov:Entity, schema:CreativeWork, schema:name, skos:prefLabel, void:Dataset	0	dcterms:title, schema:name, skos:prefLabel	prov:Entity, void:Dataset, schema:CreativeWork, edm:Aggregation, crm:E78_Collection
Senast ändrad	Date and system time of latest modification	rico:isModificationDateOf	crm:P4_has_time-span, dcterms:date	crm:P4_has_time-span, dcterms:date, schema:dateCreated, time:TemporalEntity	0	0	crm:P4_has_time-span, dcterms:date, schema:dateCreated, time:TemporalEntity
Upphovsrätt	Copyright	rico:ConditionsOfUse	dcterms:isVersionOf, edm:WebResource, rico:Instantiation	dcterms:isVersionOf, edm:WebResource, rico:Instantiation, schema:isBasedOn	0	0	dcterms:isVersionOf, edm:WebResource, rico:Instantiation, schema:isBasedOn
Topografihän visningar	Topography references	rico:IsAssociatedWithPlace			0	geonames:locatedIn	gn:Feature, crm:E53_Place, dc:spatial, gn:Place, schema:Place, wd:Q515 (City)

Table 5. The evaluation results

Total Cases	69	
Hallucinations	≈ 0	
Acceptable Recommendations	32 cases (46.37%)	**Both acceptable and non-acceptable recommendations:** 23 times (71.8%)
		Only acceptable recommendations (exclusively): 10 cases (31.25%)
		Only non-acceptable recommendations (exclusively): 8 times (25%)
No Recommendations	37 cases (53.62%)	**True Negative (TN) cases:** 16 (43.24%)
		False Negative (FN) cases: 21 (56.75%)

List of Figures

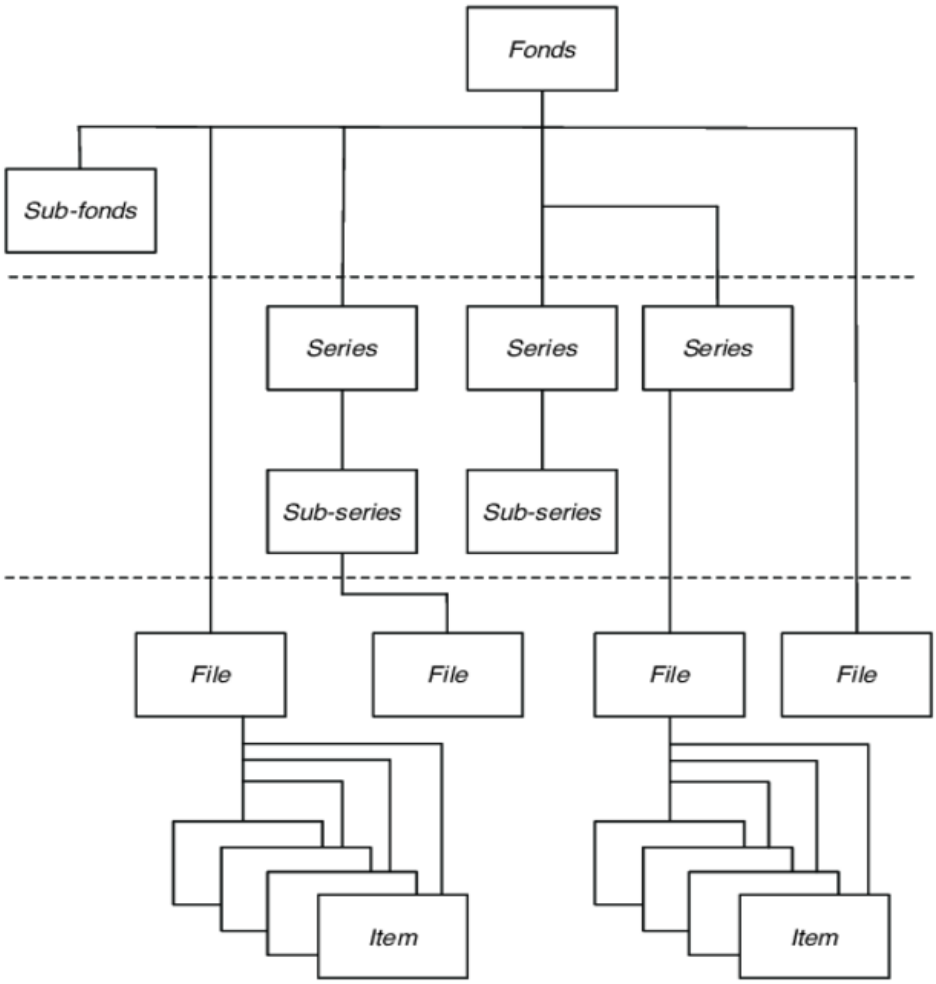


Figure 2. The mapping process to align RA and RiC-O (Maratsi, Haskiya et. al, 2025)

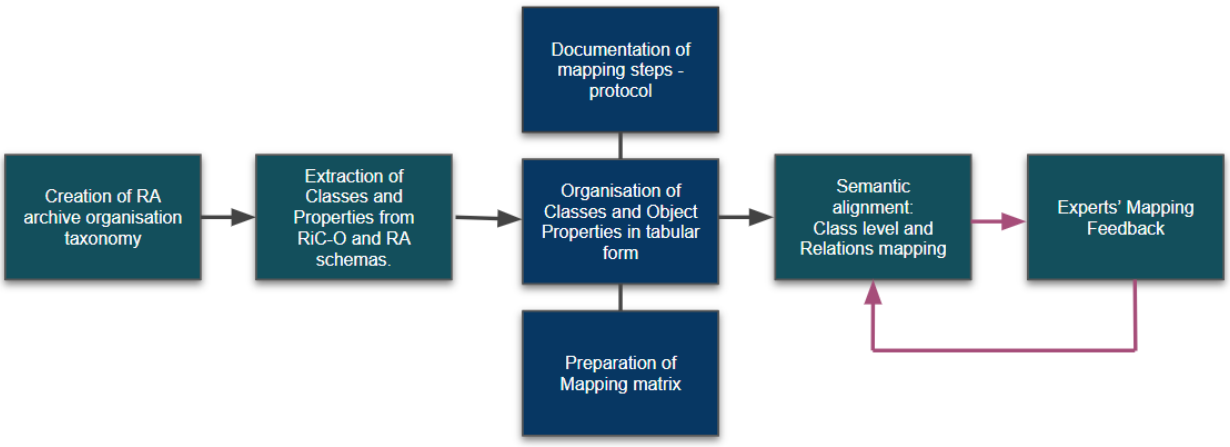


Figure 3. The methodological set-up for the two-fold LLM experiment

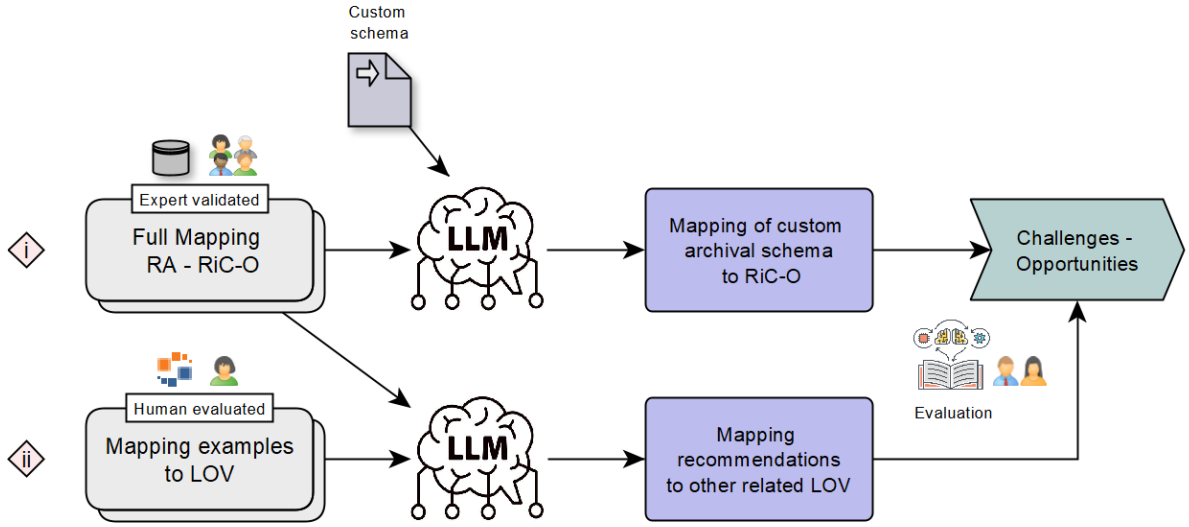
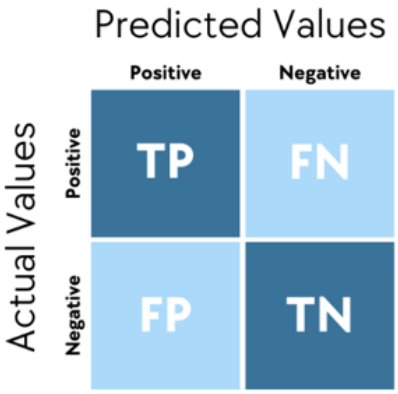


Figure 4. Confusion matrix and the formulas (Soni, 2025)



$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Figure 5. The domains (and LOV vocabulary per domain) included in the recommendations

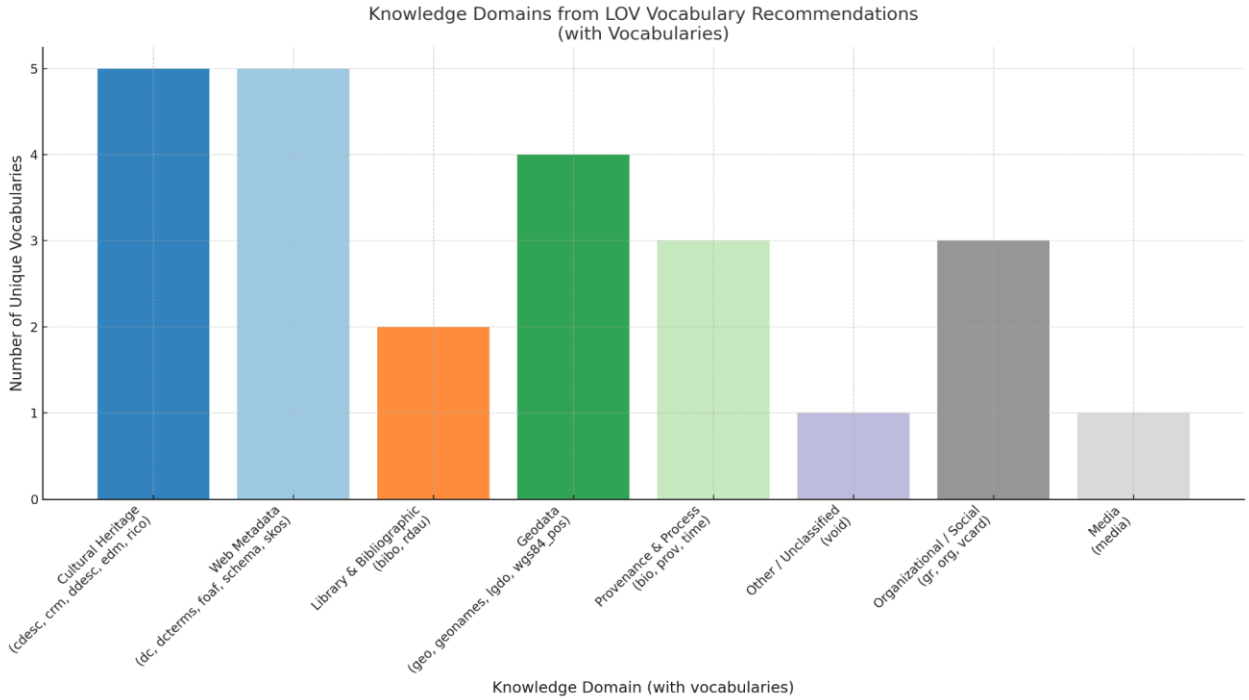


Figure 6. Number of occurrences for each LOV

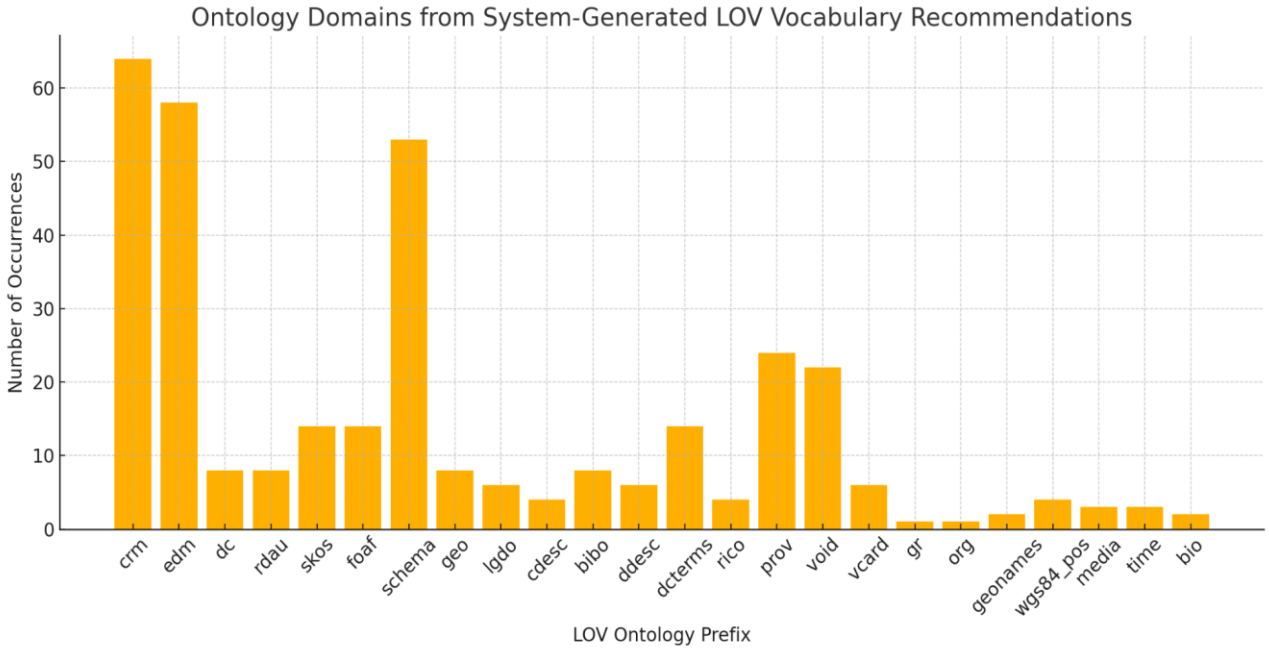


Figure 7. RA schema elements and their relation to the recommended LOV

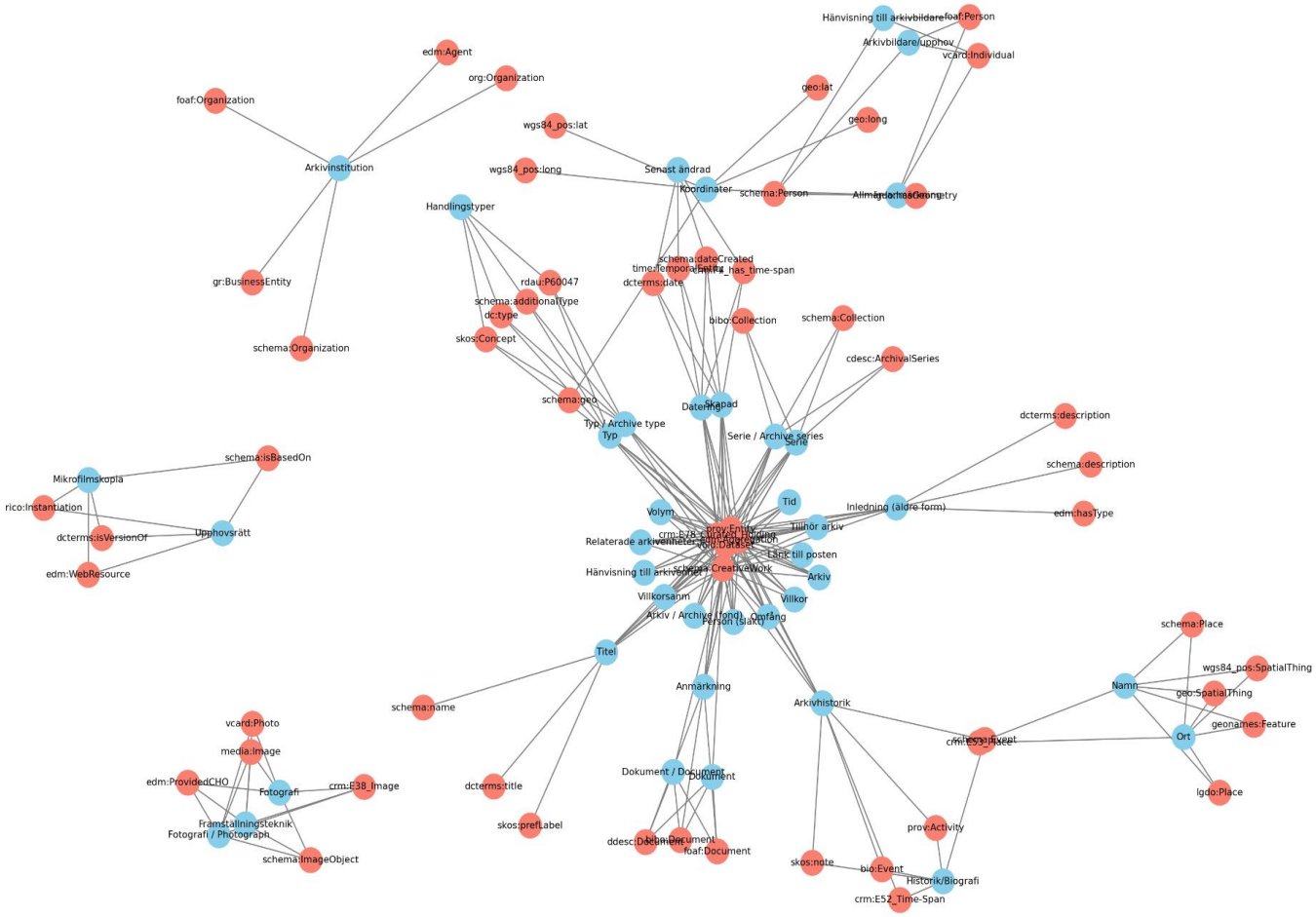
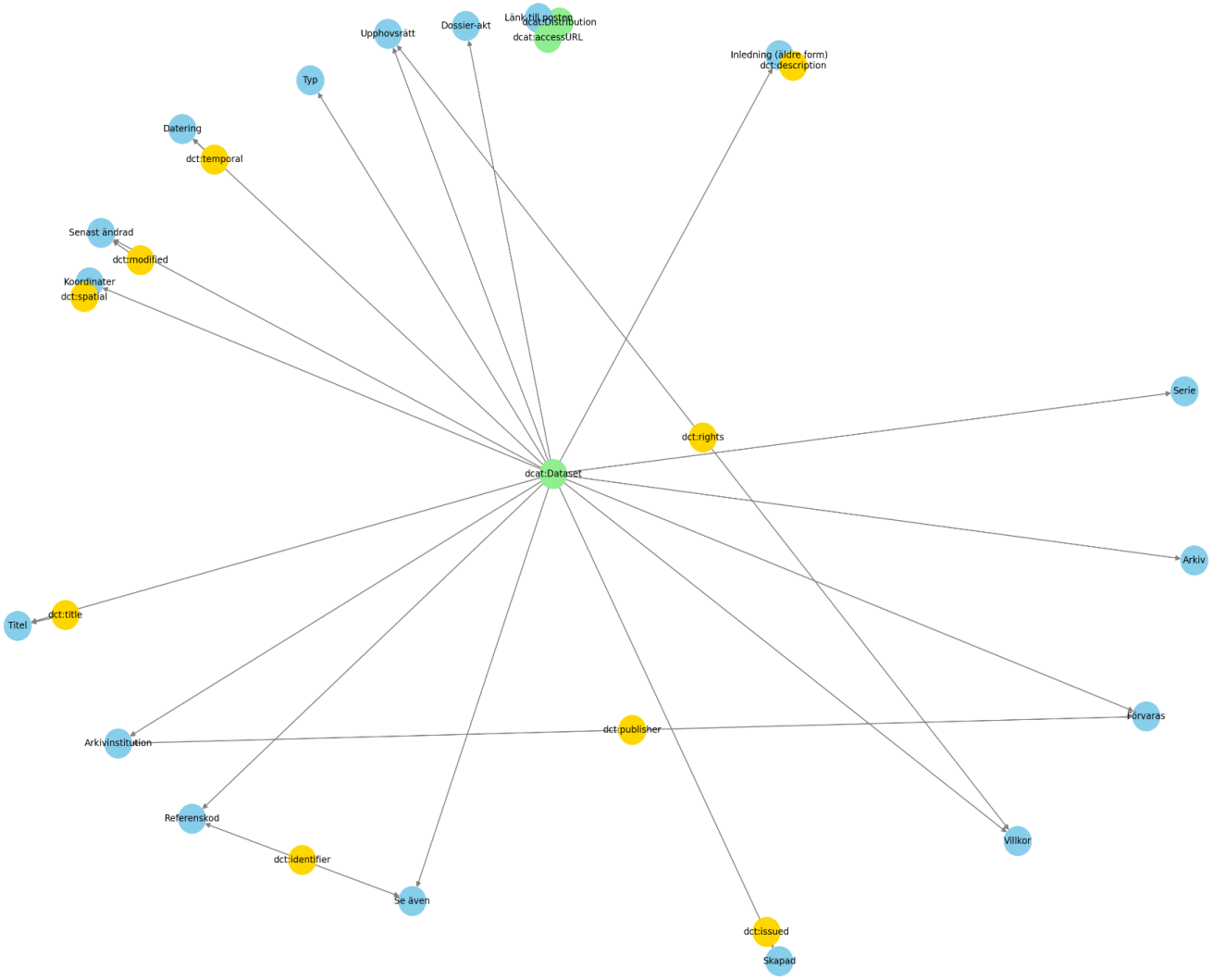


Figure 8. DCAT metadata descriptor correspondence to RA elements



Acknowledgements and Declarations

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955569.

The authors declare no conflicts of interest.

Appendix List

Appendix 1 - The Mapping Matrix (Classes and Properties):

https://zenodo.org/records/17601841

Appendix 2 - RiC-O Object Properties Needed for Context Description in RA:

https://zenodo.org/records/17601964

Appendix 3 - Extended Mapping Examples RA – LOV: https://zenodo.org/records/17603808

Appendix 4 - Prompting Templates Used: https://zenodo.org/records/17603894

Appendix 5 - A Custom Archival Institution Data Element Set:

https://zenodo.org/records/17603985

Appendix 6 - Full Table for Experiment 1a: https://zenodo.org/records/17604196

Appendix 7 - Full Table for Experiment 1b: https://zenodo.org/records/17604200

Appendix 8 - Full Table for Experiment 2: https://zenodo.org/records/17604207

Energy Consumption - CO2 Calculation

Experiment 1

- **Number of Tokens:** ~24,000 tokens
- **Energy Consumption:** 0.12–0.60 Wh
- **Carbon Emissions:** 0.05–0.24 grams CO_2e

Experiment 2

- **Number of Tokens:** ~3,420
- **Energy Consumption:** 0.017 – 0.086 Wh
- **Carbon Emissions:** 0.007 – 0.034 g CO_2e

Total

- **Number of Tokens:** ~27,420
- **Energy Consumption:** 0.14 – 0.69 Wh
- **Carbon Emissions:** 0.055 – 0.27 g CO_2e