

Chatting with Papers: A Hybrid Approach Using LLMs and Knowledge Graphs

Vyacheslav Tykhonov^a, Han Yang^b, Philipp Mayr^b, Jetze Touber^a and Andrea Scharnhorst^a

^a*Data Archiving and Networked Services, Royal Netherlands Academy of Arts and Sciences (DANS-KNAW), 2593 HW The Hague, Anna van Saksenlaan 51, The Netherlands*

^b*GESIS – Leibniz-Institute for the Social Sciences, Cologne, Germany,*

Abstract

This demo paper reports on a new workflow *GhostWriter* that combines the use of Large Language Models and Knowledge Graphs (semantic artifacts) to support navigation through collections. Situated in the research area of Retrieval Augmented Generation, this specific workflow represents the creation of local and adaptable chatbots. Based on the tool-suite *EverythingData* at the backend, *GhostWriter* provides an interface that enables querying and “chatting” with a collection. Applied iteratively, the workflow supports the information needs of researchers when interacting with a collection of papers, whether it be to gain an overview, to learn more about a specific concept and its context, and helps the researcher ultimately to refine their research question in a controlled way. We demonstrate the workflow for a collection of articles from the *method data analysis* journal published by GESIS – Leibniz-Institute for the Social Sciences. We also point to further application areas.

1. Introduction

Who would not love to have the possibility to navigate through a collection of documents by asking all kinds of questions to it? Traditionally, knowledge is communicated by language in a more or less codified way. On the less codified, natural language side, we find the development and application of Generative AI, which are primarily based on Large Language Models (LLMs) and mimic chatting with an “expert” [1]. At the other end of the spectrum of how knowledge is communicated, we find Knowledge Organization Systems (KOS) [2] (and more specifically, semantic artifacts [3]). To be queried, the latter require some prior understanding of the structure in which the knowledge is codified. Think here about how one would have used a systematic catalog in a library in which books are indexed and ordered according to classification systems.

✉ vyacheslav.tykhonov@dans.knaw.nl (V. Tykhonov); Han.Yang@gesis.org (H. Yang); philipp.mayr@gesis.org (P. Mayr); jetze.touber@dans.knaw.nl (J. Touber); andrea.scharnhorst@dans.knaw.nl (A. Scharnhorst)

🌐 <https://pure.knaw.nl/portal/en/persons/vyacheslav-tykhonov> (V. Tykhonov);
<https://www.gesis.org/institut/ueber-uns/mitarbeitendenverzeichnis/person/Han.Yang> (H. Yang);
<https://philippmayr.github.io/> (P. Mayr); <https://pure.knaw.nl/portal/en/persons/jj-touber> (J. Touber);
<https://pure.knaw.nl/portal/en/peoples/andrea-scharnhorst> (A. Scharnhorst)

>ID 0000-0001-9447-9830 (V. Tykhonov); 0009-0000-0913-824X (H. Yang); 0000-0002-6656-1658 (P. Mayr);
0000-0003-4532-1273 (J. Touber); 0000-0001-8879-8798 (A. Scharnhorst)

 © 2025 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

Metaphorically, one could describe those two ends of the spectrum of how to engage with knowledge as “asking an expert” versus “asking a librarian”.

The motivation for this explorative research is multifold. First, it starts with being aware of the ambiguity of natural language. There is power in this ambiguity: It serves serendipity and enables the emergence of associations across contexts. But this ambiguity might also lead to possible misunderstandings. So, to compensate for the ambiguity of natural language, one might look in the direction of more defined, codified terminology as expressed in KOS. But terms from KOS also carry a burden. By their nature, they are defined in specific contexts and usually do not easily travel across knowledge domains. The experimental workflow and demo we present here aspire to combine the best of both ways to communicate knowledge when it comes to information retrieval. This demo paper is scientifically located in the area of Retrieval Augmented Generation (see 2). An additional motivation for developing this workflow lies in the fact that many generative AI solutions are provided by large tech companies. They are trained on specific materials, the origin of which is not always known. Additionally, these new generations of AI are based on stochastic methods, and so the results cannot be reproduced in a deterministic way. The so-called ‘hallucinations’, assertions that misrepresent facts and context, are one known effect of these stochastic models [4]. Despite these features, the power of Large Language Models (LLMs) is increasingly unfolded in all kind of research. The goal of the development of a fully functional prototype as reported in this demo paper was also to explore how data streams can be effectively connected to LLMs on a smaller scale, ensuring that relevant data segments are delivered to the model at the right time. Knowledge graphs appear to be a promising approach for sharing structured information across different models, as they include provenance and data origin, which are essential for trust and context, and combined with LLM’s can enhance information retrieval. With this demo paper, we discuss how LLM’s and industry provided solutions can become embedded in workflows which can be controlled locally, so that AI is applied in a way following Open Science principles, and being reproducible and ethical.

Our paper starts with a short overview of some related works in section 2. We introduce the use case in section 3. The description of the *GhostWriter* workflow and some demonstrations of its abilities form the core of this paper in section 4. The paper closes with conclusions and an outlook into future research topics in section 5.

2. Related work

Information Retrieval (IR) is the task of finding specific information from a large collection of documents [5] given a query. This is not only an important step in the scientific endeavor to answer a specific research question, but also, more generally, a crucial way to reduce information overload and cognitive overload [6]. With the help of LLMs, augmented search generation (RAG) [7] solves this task in a modern approach, in which documents, and particular paragraphs from them, are stored in the vector database, together with their semantic embeddings. Once a query is proposed, it will be converted into the embedding space and matched with the embedding vectors of the paragraphs stored in the vector database by calculating the cosine similarity. The matched paragraphs are consequently summarized by the LLM, and the query is then answered

with natural language.

With the RAG method, a query could be answered appropriately with less likelihood of triggering the hallucination of the LLMs, as long as the paragraphs can be matched precisely. But there are situations in which the standard RAG performs less well. For instance, when questions require a comprehensive understanding of the documents on a more global scale, and especially in cases when the concepts which are vital to answer a query are located far apart in the documents. As an improvement, the GraphRAG [8] targets the same kind of query and constructs the paragraphs in a graph structure, where the entities and relationships are extracted and are used as the vertices and edges of a graph, together with a text description generated by the LLM. This graph is then partitioned hierarchically into subgraphs, called the graph communities. To answer a query, the graph communities are summarized by the LLM. The LLM toolbox contains scores which indicate whether this summarization answers the query. To generate the final answer, the summaries (of the graph communities) are ranked to the LLM-based score and concatenated until the token limit is reached. The GraphRAG method has been proven to perform well on tasks that require the information to be located in different parts of documents. However, the graph might be sparse when the entities and relationships are not explicitly explained or paraphrased, and hence the connections will break. In contrast, the workflow of this demo paper while based on the RAG paradigm, uses ontology and knowledge graphs to eventually define graph communities. This way, the workflow emphasizes the generality between entities and relationships in order to improve the matching accuracy in a wider range within the documents.

For the purpose of science studies, the new generation of LLMs sparks new ideas to further improve information retrieval and large corpus analysis, like Named Entity Recognition [9]. Information retrieval is full of approaches to organize personal information spaces [10]. If it comes to exploring the past and present of science dynamics, a toolbox is available for bibliometric analysis of formal scholarly communication [11, 12]. Platforms have also been tested to collectively build, annotate, and analyze collections of scientific papers, targeting their metadata and full text [13]. Some recent representations in industry include the Ai2 PaperFinder from Allen Ai¹, which assists users with an interactive workflow and clarifies the user's goal step-by-step, as well as the ScienceDirect AI tool from Elsevier², which emphasizes the transparency of the retrieved papers. In short, in Information Retrieval in general and in quantitative studies of the science, we find more and more explorations with tooling based in generative AI approaches. In contrast to these efforts, our workflow is designed to be applied to an arbitrary collection of documents. It provides the user with a web interface but the majority of the tooling appears at the backend where a couple of new techniques (all connected via API services) is bundled together to enhance the provided information by data and metadata and to give feedback to the user in the form of natural language-based statements.

¹<https://paperfinder.allen.ai/chat>

²<https://elsevier.shorthandstories.com/sciedirect-ai/>

3. Use case - journal *mda*

As a use case, we present a collection of social science articles from the journal *mda - methods, data and analysis*. The journal itself was founded in 2007 and is published by GESIS - Leibniz Institute for the Social Sciences. The journal started originally with a mixture of German and English language articles “on all questions important to quantitative methods in the social sciences, with a special emphasis on survey methodology”³ to now be fully published in English as a Diamond Open Access journal. From this journal, we created a test collection of 100 random articles in PDF format and as demonstrated in the next section 4, we applied the *GhostWriter* workflow on this collection⁴.

4. GhostWriter demo

4.1. System Design

The workflow has in essence two parts: a data extraction and enrichment pipeline, called *Everything Data*, leveraging a knowledge graph and a human-computer-interface (hci) called *GhostWriter* [14]. It makes use of a recent metadata format standard (called Croissant for Machine Learning) for datasets used for machine learning, and techniques to contextualise and semantically enrich information [15]. If we talk about a workflow or pipeline here, one has to be aware that this is a rather simplifying label. For many of the current generative AI experiments, we find that their machinery consists of a ‘bag of various tools’ combined based on the tacit knowledge of engineers and researchers.⁵

For this demo paper, we decided to present elements of the workflow schematically in Figure 1. In the upper part, we find a description of *EverythingData*⁶ part of the workflow, while the lower part represents the *GhostWriter* interface. *EverythingData* is a generic way to extract and enrich information from documents of any kind. After documents are split and paragraphs are extracted (1, in Figure 1), terms from these paragraphs are extracted (2) and those terms are further enriched (3). For each of these steps, embeddings are created, using a vector database in which those operations take place. As a result, we find terms which are already selected, weighted, extended according to the way the embeddings are executed, with the result that terms become contextualized further. In the lower part of Figure 1, we schematically summarize the *GhostWriter* interface. It starts with a query (already taken from our use case) for which information is retrieved (4). Eventually, the human-computer interface presents a natural language answer and pointers to possible source documents (5). This last step can be repeated, thus enabling an iterative way of further specifying the original query (chatting with the local collection). The data enrichment in *EverythingData* is enabled by the LLM-based term extraction and the knowledge graph & ontology-based term enrichment, while the HCI interface is achieved

³<https://mda.gesis.org/>

⁴<https://gesis.now.museum/?page=about> Beta version, full text close reading approach implemented, workflow and interface still in development.

⁵Personal communication with Arno Simons at the Workshop “Large Language Models for the History, Philosophy and Sociology of Science” Berlin, April 2-4 <https://www.tu.berlin/en/hps-mod-sci/workshop-llms-for-hpss>.

⁶<https://github.com/Dans-labs/everythingdata>

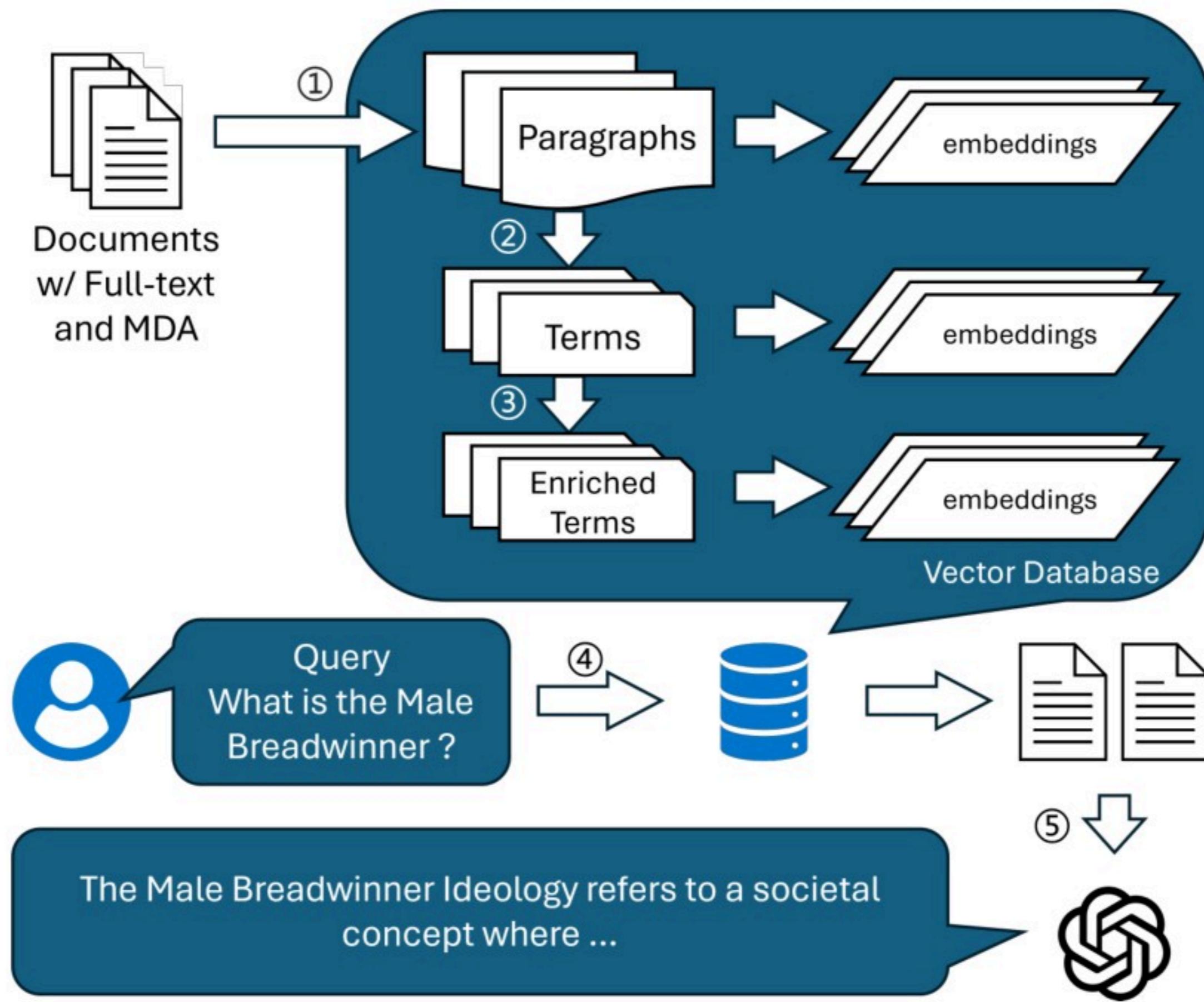


Figure 1: Schematic Workflow of *EverythingData* and *Ghostwriter*

by the LLMs, which enables natural language. One interesting feature of this approach is that multilinguality is naturally part of it, due to the coupling to knowledge graphs which connect various language spaces (such as wikidata).

4.2. A detailed chat with *method-data-analysis* papers

To showcase our workflow, we created a collection with 100 articles from the *mda* journal⁷. We then proposed questions to the interface and inspected the results. For this demo paper, we choose to display the question: “explain male breadwinner model to me”. The first response is shown in a snapshot from the website (see Figure 2). One can see that the question is explained in detail, and related articles are listed as well, together with a score of confidence (calculated inside the workflow). This indicates that this workflow can answer questions in an academic language and support further investigation into the collection.

It is interesting to note that the listed articles do not necessarily need to contain the exact query term to be found relevant. The user is invited to further close-reading to explore the relevance of the listed hits.

But what is even more interesting is the fact that the question can be refined, as shown in Figure 3. In the provided source documents, there is no direct information about how data was collected in connection with the male breadwinner model. Still, if we refine the question in this direction, we get answers. For instance, one of the listed articles mentions that a study used data from Germany and employed a mixed-methods research strategy, including a survey experiment

⁷Ghostwriter on MDA papers, betaversion, see <https://gesis.now.museum>, select collect collection mda

Enter your question:

Ask
 Good
 Bad
Share
Copy
Google

Answer

The Male Breadwinner Ideology refers to a societal concept where men are expected to be the primary earners and financial providers for their families. This ideology emphasizes the importance of men's work and earning potential in maintaining social status, respectability, and family well-being. In the context of Germany, researchers have studied this concept, examining how it influences individuals' attitudes towards establishing market relationships and engaging in entrepreneurial activities. The model suggests that men who strongly adhere to traditional breadwinner norms are more likely to establish market relationships and pursue business opportunities.

Sources

	Chat	Title	Reference	Score
0	Ask	The Past, Present and Future of Factorial Survey Experiments: A Review for the Social Sciences Treischl methods, data, analyses	https://mda.gesis.org/index.php/mda/article/download/2021.07/293	0.427998
1	Ask	Gender and Survey Participation: An Event History Analysis of the Gender Effects of Survey Participation in a Probability-based Multi-wave Panel Study with a Sequential Mixed-mode Design Becker methods, data, analyses	https://mda.gesis.org/index.php/mda/article/download/2021.08/299	0.418127

Figure 2: Question on “male breadwinner model”

[16]. Additionally, it references another article that examined the male breadwinner model [17], but does not provide information on how data was collected for this study. Obviously, the associations to other terms created in the *EverythingData* part of the workflow lead to suggestions to other sources. Knowledge-based enrichment of terms with their multilingual properties can increase the number of sources that can be queried and improve recall without compromising precision. In general, this demonstration shows that a way to chat with papers can be created that goes far beyond any query term or related query term matching. To evaluate such workflows in a proper scientific manner is not an easy task and requires many well-defined user tests and evaluations, but at least we can demonstrate a working technology which can be used for the multilingual sources.

5. Conclusion and Outlook

This demo paper bases on experiences gained in a series of project-based explorations concerning information processes in the social sciences and humanities, and more specifically in the area cultural heritage and their information infrastructural needs, such as the European project MuseIT⁸ [18] and the Dutch research infrastructure project ODISSEI⁹.

As we demonstrated in Section 4, the workflow belongs to the RAG approach and relies on

⁸<https://www.muse-it.eu>

⁹<https://odissei-data.nl>

The screenshot shows a user interface for a question-answering system. At the top left is a "Ask Questions" button with a thinking face icon. Below it is a text input field containing the question "explain how data was collected on male breadwinner model". A red "Ask" button is positioned next to the input field. To the right of the input field are several interaction buttons: "Good" (with a thumbs up icon), "Bad" (with a thumbs down icon), "Share", "Copy", and "Google".

Answer

According to the provided text, there is no direct information about how data was collected on the male breadwinner model. However, it mentions that a study used data from Germany and employed a mixed-methods research strategy, including a survey experiment (Hainmueller et al., 2015). Additionally, it references another article that examined the male breadwinner model (Haase et al., 2016), but does not provide information on how data was collected for this study.

Sources

	Chat	Title	Reference	Score
0	Ask	The Past, Present and Future of Factorial Survey Experiments: A Review for the Social Sciences Treischl methods, data, analyses	https://mda.gesis.org/index.php/mda/article/download/2021.07/293	0.516705
1	Ask	Gender of Interviewer Effects in a Multi-topic Centralized CATI Panel Survey Lipps methods, data, analyses	https://mda.gesis.org/index.php/mda/article/download/2016.009/148	0.468180

Figure 3: Finetuning the question on the “male breadwinner model”

entity extraction, where terms are annotated with semantic meaning by mapping them into knowledge graphs built from controlled vocabularies. These ways to connect vector spaces (LLM's) to a knowledge graph (structured knowledge, KOS) open possibilities to enhance the semantic meaning for information retrieval tools.

The challenge with this approach - which in essence is combining semantically related pieces of text - is achieving a coherent interpretation of the context. It often fails when contradictions or differing points of view are present in the source collection. Additionally, semantic similarity does not guarantee logical consistency, and models may incorrectly link unrelated content due to superficial lexical overlap. There's also a risk of the wrong understanding of the context, where nuanced or domain-specific meanings are lost.

As a possible future solution, the current ranking algorithm should be revised to first cluster and rank document fragments at the document level. This means grouping semantically similar fragments by their source document before evaluating their relevance. By doing this, we avoid mixing contradictory content from different documents and preserve contextual coherence. Once ranked, selected clusters can be passed to the LLM, along with explicit metadata about provenance such as paper titles, authors, publication dates, and source repositories. This enhances the reliability of the generated responses by making the origins of information traceable and allowing the model to distinguish between perspectives from different sources. Linking entities to even more knowledge graph representations (such as wikidata) brings the possibility of multilingual search as one immediate benefit of such enrichment.

As demonstrated in the case of chatting with (some) *mda* papers, any query input starts a kind of distant reading machinery which returns with an answer (using natural language) together with pointers to the sources from the original collection supporting this provisional answer. This way, it mimics a close-reading level and still enables the user to continue her/his close-reading journey.

The *GhostWriter* workflow works for any “collection”. You can create a “semantic index” and a local chatbot for collections from webpages/websites, RSS feeds, but also from trusted data repositories such as Dataverse instances. (see DataChat¹⁰). In essence, you enable local chatting with papers – this way pointing a human to specific texts which one can then engage with by way of close-reading. You confine the ‘search space’ by placing your question (and the collection) into a particular area in the networked space of scientific knowledge. We also demonstrated that you gain information beyond what is explicitly expressed in the text and annotated in the metadata (by creating associations on the level of natural language and on the level of Knowledge Organisation Systems). It is like you start chatting with the experts or invisible colleges behind the papers. But, there are also many preconditions to make such a workflow work. It starts with API’s from which collections can be harvested, and API’s which provide KOS services (for the risks of KOS services see e.g. [19]). Well-curated metadata and digitized full text are another precondition. Implementation of the Croissant Standard [15] as a new standard for datasets for machine-learning is another precondition. As part of our future work, we have developed the Model Context Protocol (MCP)¹¹ and initiated the practice of depositing metadata for papers separately in the Dataverse data repository. This setup allows MCP to be used as an automated factual query source. The approach appears more reliable than other approaches, as it retrieves up-to-date metadata directly from data sources and enables verification of the information provided. It’s also currently being deployed as “Ask Dataverse” services and tested by Harvard IQSS on <https://ask.dataverse.org>. So in short, while this technology offers great opportunities, it is far from being a ’ready-from-the-shelf-implementable’ software. It is still very much in development.

Having said this, one large application field is collections of documents from the Cultural Heritage - for instance, the many collections that Galleries, Libraries, Archives and Musea hold [20] contain. Developing skills to develop your workflows and to tailor them to your own needs is a very important way to empower Global Open Research Commons [21], a way to execute Open Science, even more needed today.

Acknowledgments

Parts of this paper have been funded by projects such as ODISSEI (see <https://odissei-data.nl>, SSHOC.nl (NWO fund), FAIRImpact (EC Grant DOI 10.3030/101057344), MuseIT (EC Grant DOI 10.3030/101061441) and NFDI4DataScience. We thank Yves Rozenholc (University Paris Descartes) and the now.museum for their collaboration and support. Han Yang received funding from the Deutsche Forschungsgemeinschaft (DFG) under grant number: MA 3964/15-3 (Socio-Hub project). Han Yang and Philipp Mayr received additional funding from the European Union

¹⁰<https://github.com/gdcc/datachat>

¹¹<https://mcp.dataverse.org>

under the Horizon Europe grant OMINO – Overcoming Multilevel INformation Overload¹² under grant number 101086321 [6].

References

- [1] W. Fan, Y. Ding, L. Ning, S. Wang, H. Li, D. Yin, T. Chua, Q. Li, A survey on RAG meeting llms: Towards retrieval-augmented large language models, in: R. Baeza-Yates, F. Bonchi (Eds.), Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024, ACM, 2024, pp. 6491–6501. doi:10.1145/3637528.3671470.
- [2] F. Mazzocchi, Knowledge organization system (kos): an introductory critical account, *KO Knowledge Organization* 45 (2018) 54–78. URL: <https://www.isko.org/cyclo/kos>.
- [3] Ó. Corcho, F. J. Ekaputra, I. Heibi, C. Jonquet, A. Micsik, S. Peroni, E. Storti, A maturity model for catalogues of semantic artefacts, *CoRR* (2023). doi:10.48550/ARXIV.2305.06746.
- [4] J. Yao, K. Ning, Z. Liu, M. Ning, L. Yuan, LLM lies: Hallucinations are not bugs, but features as adversarial examples, *CoRR* (2023). doi:10.48550/ARXIV.2310.01469.
- [5] H. Schütze, C. D. Manning, P. Raghavan, *Introduction to information retrieval*, volume 39, Cambridge University Press Cambridge, 2008.
- [6] J. A. Hołyst, P. Mayr, M. Thelwall, I. Frommholz, S. Havlin, A. Sela, Y. N. Kenett, D. Helic, A. Rehar, S. R. Maček, et al., Protect our environment from information overload, *Nature Human Behaviour* 8 (2024) 402–403. doi:10.1038/s41562-024-01833-8.
- [7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, December 6-12, 2020, virtual, 2020. URL: <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>.
- [8] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, J. Larson, From local to global: A graph rag approach to query-focused summarization, *arXiv preprint arXiv:2404.16130* (2024).
- [9] X. Tong, N. Smirnova, S. Upadhyaya, R. Yu, J. H. Culbert, C. Sun, W. Otto, P. Mayr, Utilizing large language models for named entity recognition in traditional chinese medicine against COVID-19 literature: Comparative study, *CoRR* (2024). doi:10.48550/ARXIV.2408.13501.
- [10] N. Goharian, N. Tonellootto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part I*, volume 14608 of *Lecture Notes in Computer Science*, Springer, 2024. doi:10.1007/978-3-031-56027-9.
- [11] N. J. Van Eck, L. Waltman, Visualizing bibliometric networks, in: *Measuring scholarly impact: Methods and practice*, Springer, 2014, pp. 285–320.

¹²<https://ominoproject.eu/>

- [12] P. Mayr, A. Scharnhorst, *Scientometrics and Information Retrieval - weak-links revitalized*, *Scientometrics* 102 (2015) 2193–2199. doi:10.1007/s11192-014-1484-3.
- [13] D. Chavalarias, Q. Lobbé, A. Delanoë, *Draw me science: Multi-level and multi-scale reconstruction of knowledge dynamics with phylomemries*, *Scientometrics* (2021) 1–31.
- [14] V. Tykhonov, F. van Rijsselberg, E. Endarto, *The next generation of data management with artificial intelligence* (2024). doi:10.5281/zenodo.14507120.
- [15] O. Benjelloun, E. Simperl, P. Marcenac, P. Ruyssen, C. Conforti, M. Kuchnik, J. van der Velde, L. Oala, S. Vogler, M. Akthar, N. Jain, S. Tykhonov, *Croissant format specification* (2024). URL: <https://mlcommons.github.io/croissant/docs/croissant-spec.html>.
- [16] J. Hainmueller, D. Hangartner, T. Yamamoto, *Validating vignette and conjoint survey experiments against real-world behavior*, *Proceedings of the National Academy of Sciences* 112 (2015) 2395–2400.
- [17] M. Haase, I. Becker, A. Nill, C. J. Shultz, J. W. Gentry, *Male breadwinner ideology and the inclination to establish market relationships: Model development using data from germany and a mixed-methods research strategy*, *Journal of Macromarketing* 36 (2016) 149–167.
- [18] M. Johansson, V. Tykhonov, S. Alexandersson, K. Ferguson, J. Hanlon, A. Scharnhorst, N. Osborne, *A knowledge base for arts and inclusion—the dataverse data archival platform as a knowledge base management system enabling multimodal accessibility*, arXiv preprint arXiv:2504.05976 (2025).
- [19] M. L. Zeng, P. Mayr, *Knowledge organization systems (kos) in the semantic web. a multi-dimensional review*, in: R. P. Smiraglia, A. Scharnhorst (Eds.), *Linking Knowledge: Linked open data for knowledge organization and visualization*, Ergon-Verlag, Ergon, 2021, pp. 34–63. doi:10.5771/9783956506611-34.
- [20] Sally Chambers and Frédéric Lemmers, *Collections as Data*, META (ANTWERPEN) (2021) 36–37. URL: <https://www.vvbad.be/meta/meta-nummer-20213/collections-data>.
- [21] S. Jones, M. Leggott, J. Lopez Albacete, C. Pascu, K. Payne, M. Schouppe, A. Treloar, G. O. R. C. IG, *Gorc ig: Typology and definitions*, 2023. doi:10.15497/RDA00087.