

Is It a Plant, Activity, or Medicine?

Constructing Thick Knowledge Graphs

Marieke van Erp¹, Paul Groth²

¹KNAW Humanities Cluster, Oudezijds Achterburgwal 185, 1012 DK Amsterdam, The Netherlands

²University of Amsterdam, Science Park 904, 1090 GH Amsterdam, the Netherlands

Abstract

A central benefit of knowledge graphs is that they coalesce information about entities. However, in doing so, they often flatten entities failing to represent important nuance about them, for example, the different contexts which entities are situated in or how they evolve over time. Such information is important for many applications particularly in the humanities. In this demo, we present a knowledge graph construction pipeline for creating thick knowledge graphs that explicitly represent important subtleties about entities such as near-identity, entity change, and perspectives. We use the construction pipeline to create a historical commodities knowledge graph and provide a browsing interface that highlights these aspects. The knowledge graph can be browsed at: <https://trifecta.dhlab.nl/idchkg-demo/>. You can find the demo video at: <https://www.youtube.com/watch?v=8eadoUb5lt8>.

Keywords

identity, change, contextual information extraction, computational history

1. Introduction

Knowledge graphs are a useful tool for representing information about entities in a structured manner supporting a variety of analysis tasks [1]. However, many knowledge graphs (KGs) are not designed to represent the different contexts of entities and how these might change over time [2]. In this demo paper, we present a knowledge graph construction pipeline for creating what we term *thick knowledge graphs* that explicitly represent these important subtleties, specifically, near-identity, entity change, and perspectives on claims about entities. Here, we are adopting the concept of ‘thick descriptions’ introduced by the philosopher Gilbert Ryle [3] and expanded by Clifford Geertz [4]. Ryle developed this concept to describe actions in the world as not only by their physical characteristics (e.g. a wink is someone quickly closing and opening one eye) but also their context as interpreted by the actors (e.g. a wink indicates a shared joke). Geertz expanded this to cultural anthropology where he advocated for ‘a web of meaning’. We find that this concept of thick descriptions is exceptionally well-suited to referring to adding layers of context to KGs. Here, we illustrate this concept by constructing a thick KG from historical sources as it showcases the need for such additional context.

The work that most closely resembles our aims are around KGs for cultural heritage and history (cf. [5, 6, 7, 8]), however, they are not primarily focused on both the complexity of entities as well as the change they undergo. Furthermore, their construction is not specifically aimed at creating thick KGs. Hence, we present the following contributions in this paper:

- A pipeline for the construction of KGs from text where the resulting KG explicitly represents near-identity among entities, claim perspectives and historical evolution;
- A static site generator that allows for the browsing of the resulting KG;
- A historical commodities KG.

The remainder of this paper is organised as follows. In Section 2, we present our data model, extraction pipeline, browser, and implementation details. In Section 3, we present the created thick historical knowledge graph. Section 4 demonstrates an exploration scenario of the knowledge graph, followed by conclusions and future work in Section 5.

ISWC 2026 - Posters and Demos 25 - 29 October 2026, Bari, Italy

✉ marieke.van.erp@dh.huc.knaw.nl (M. van Erp); p.t.groth@uva.nl (P. Groth)

ORCID [0000-0001-9195-8203](https://orcid.org/0000-0001-9195-8203) (M. van Erp); [0000-0003-0183-6910](https://orcid.org/0000-0003-0183-6910) (P. Groth)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. The idckg pipeline

The idckg¹ pipeline is focused on constructing thick knowledge graphs primarily from document corpora. A focal point of the pipeline is the data model that is targeted during construction as this operationalises the thick knowledge graph notion. We begin there and then discuss the pipeline’s extraction and browsing components. Lastly, we discuss key implementation details.

2.1. Data Model

The data model centres around *entity senses* these are different contexts in which the meaning of an entity is highlighted, for example **coffee** (*beverage*) vs **coffee** (*plant*). Relations are made between entity senses and are referred to as *claims*. The use of claim emphasizes the contingency of such relations. Each claim has a *time* (to anchor the claim temporally), a *perspective* (whether the claim is positive, neutral, or negative), and a *provenance* (the mention in the source document).

Furthermore, each *entity sense* is connected to an *entity*, which can have a *type*. Entities aggregate entity senses together. *Near-identity links* express that multiple senses contribute to a multifaceted entity representation. Figure 1a shows a graphical representation of the data model and Figure 1b shows an example.

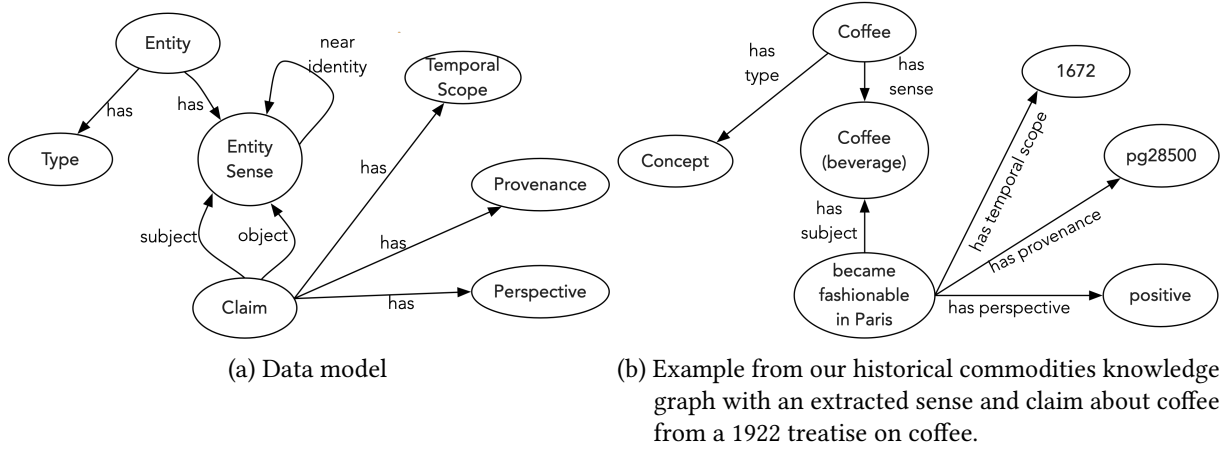


Figure 1: Overview of the identity and an example of data representation in the idckg model from a passage of *All About Coffee* by William H. Ukers (1922)

2.2. Extraction

The extraction pipeline performs six steps as shown in Figure 2. In Step 1, the ingested input is chunked into text chunks based on paragraph breaks. The size of a chunk is parameterized but has a default of 400 tokens. In Step 2, an LLM is employed to extract entity mentions, senses, and claims using zero-shot prompting. Prompts are available in the Github repository. The pipeline extracts entity mentions of type concept, person, organization, place, event, work or other. Claims are relationships between entities with additional qualifiers (e.g. temporal scope, perspective). Both senses and relationship predicates are represented as strings and unbound. This choice allows both to emerge from the corpora itself.

In Step 3, two LLM judges evaluate each claim for groundedness and qualifier fidelity. The groundedness check, inspired by [9], rates to what extent the claim is directly stated in the text - this check is built in to combat hallucinations. The qualifier fidelity check, related to modifier detection in aspect extraction (cf. [10]), assesses whether the claim elements such as the attribution and stance are sufficiently supported by the text. Both judges assign scores between 0 and 1. In case a claim receives a low score, Step 4 sends the claim back to the Information Extraction module in Step 3 for repairs.

¹The acronym ‘idckg’ stands for IDentity and CHange Knowledge Graph

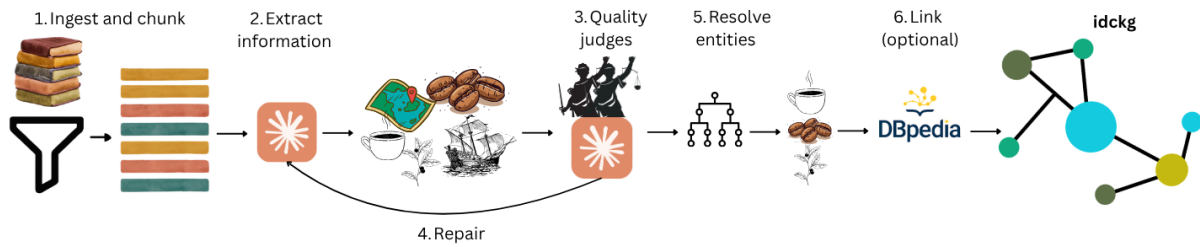


Figure 2: Overview of core steps in the idckg construction pipeline

In Step 5, entity senses are resolved from the entity mentions and then linked to an overall entity. This is done based on string equivalence between the entity and entity sense tuple. An optional step is to link the entities to DBpedia² using DBpedia Spotlight [11]. Lastly, the knowledge graph is made available as a finalised database.

This pipeline provides a proof of principle that extraction of thick knowledge graphs is possible. There are a number of areas where it can be improved such as on the evaluation of the information extraction and judging steps as well as the integration of existing structured data during construction.

2.3. Browsing

The pipeline provides the ability to create a browsing interface as a static website. The web interface shows the different entity senses and their associated perspectives. Frequently occurring entity senses are presented in more detail, low frequency entities and entity senses can be accessed via facet search and text search. Several views that support the type of explorations described in Section 4 are included such as a timeline, stream diagrams, disagreements, and sources and their reliability. For each claim, a link to the source paragraph is readily available to inspect the validity of the extracted claim. Figure 3 shows some of the views present in the demo.

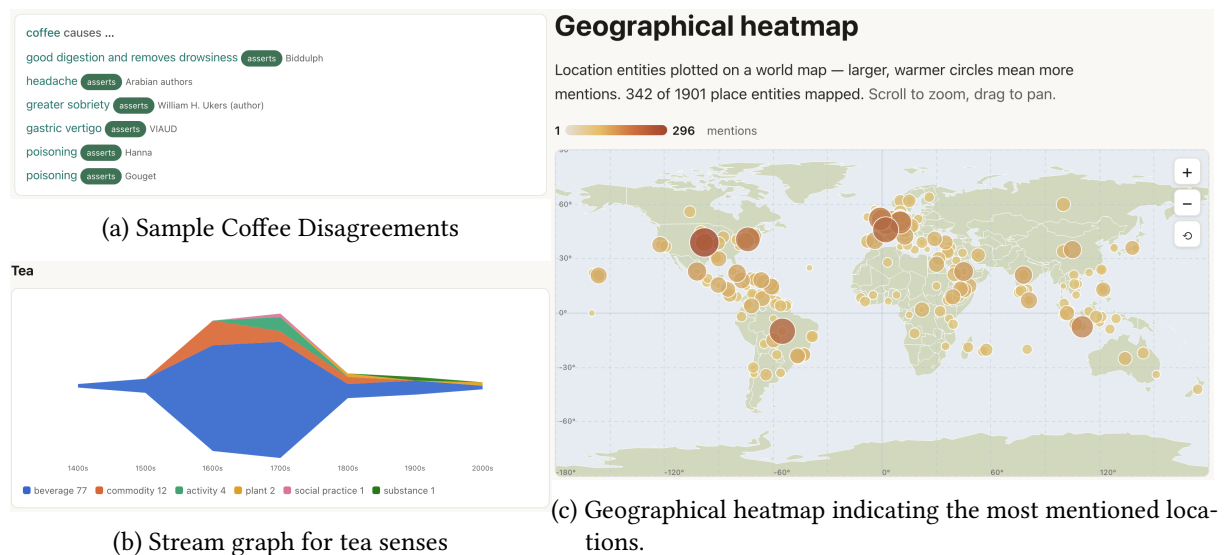


Figure 3: Example views of the browsing interface

²<https://www.dbpedia.org/>

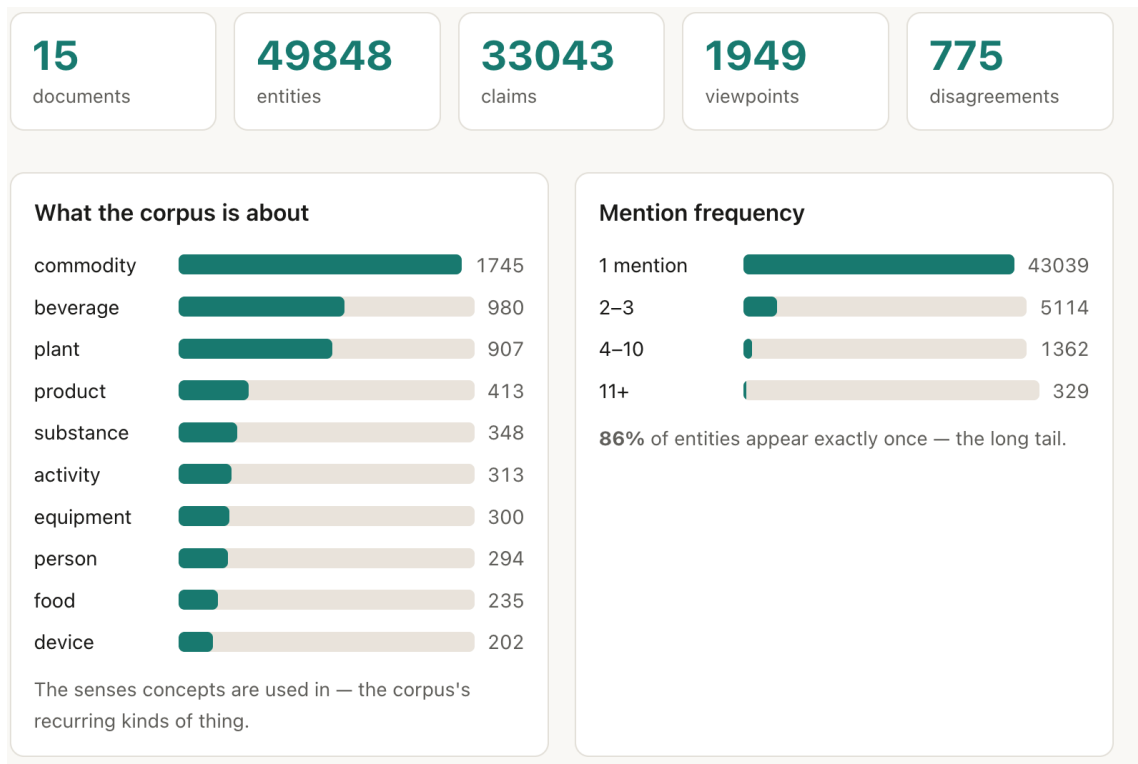


Figure 4: KG statistics overview

2.4. Implementation

The pipeline and web interface generator are implemented in Python and designed to be modular and to run locally. It offers access to local LLMs through the use of Ollama and can use models hosted behind APIs (e.g. Claude platform). Extraction is conducted concurrently where possible. A key design decision was the ability to have the entire KG stored as a single portable file. Specifically, the KG is stored as an SQLite database. The implementation also offers the ability to export to RDF/Turtle and a JSON format. The web interface is generated as a series of static pages from the SQLite database to optimise speed during browsing and removes the need to run an online database e.g. SPARQL endpoint. The codebase for the project can be found at: <https://github.com/trifecta-project/idchkg>.

3. A Historical Commodities KG

We showcase the use of the idchkg pipeline to generate a historical commodities KG. The selection of this domain was driven by our collaboration with historians specialized in commodities history as part three projects: SevenFrontiers,³ Globalise,⁴ and TRIFECTA.⁵ In these collaborations, we have elicited competency questions that require thick knowledge graphs. The CQs follow roughly four categories: 1) type and context classification of entity mentions (e.g. which mentions of tea refer to a beverage and which to a medicine), 2) the source and perspective of the information (e.g. is this a policy view or an opinion), 3) temporal information (e.g. did trade volumes of cocoa increase or decrease with time), and 4) which agents and locales were involved (e.g. was the adoption of coffee widespread in Italy). The resulting KG aims to help address these CQs.

Corpus: To showcase the history of various commodities, we have created a corpus of 10 books from

³<https://cordis.europa.eu/project/id/101141427>

⁴<https://globalise.huygens.knaw.nl/>

⁵<https://trifecta.dhlab.nl>

Project Gutenberg⁶ concerning a specific commodity or set of commodities as well as 5 Wikipedia articles detailing the history of Coffee, Tea, Sugar, Spices, and the Colombian Exchange (the trade between the New World (Americas) and the Old World (Europe and Africa)). The books we processed were published between 1730 and 1922 and comprise a total number of tokens of 94,964. The Wikipedia articles add an additional 685 tokens.

Construction: The documents were processed using the extraction pipeline described in Section 2.2. We used Claude Haiku 4.5 [12] for a higher quality extraction and speed rather than a smaller, local model. The knowledge graph took 2.5 hours to construct using 24 concurrent workers. Construction is dominated by the LLM judges about 81% of the 79,000 LLM calls going to them with the rest going to extraction and repairs. In total, 102 million tokens were used and the cost of extraction was \$172.

The Knowledge Graph: Figure 4 presents the core KG statistics showing how it captures the kind of thick information discussed previously, for example, a multitude of viewpoints, senses and disagreements. The knowledge graph as a SQLite database is available at doi.org/10.5281/zenodo.21515713. The knowledge graph can be browsed at <https://trifecta.dhlab.nl/idchkg-demo/>.

Addressing the competency questions: Through the browsing interface to the KG, we aimed to start to address the competency questions of historians. The interface shows for each entity sense a list of other senses in the corpus as well as a list of defining properties that contribute to that sense. Furthermore, claims from different sources are presented along with links to the text source and information on the extraction model’s confidence and a timeline. For the most frequently mentioned entities, a separate disagreements page shows the different claims made. On occasion these contradict each other, but often these provide updated information, such as coffee exports totalling 22,000,000 pounds in 1810 and 88,700 pounds in 1908 demonstrating ups and downs in the trade.

4. Demonstration Scenario

During the demo, visitors will be able to explore the historical knowledge graph. Users land on the starting page where they can browse the most entity senses, filter by facets or search, view disagreements and access the sources. Upon selecting any of the entity senses or claims, further information is displayed. For entities this consists of a connections graph, other senses, a DBpedia link (where available), defining properties, claims, and a timeline. Each of these elements is interactive allowing them to navigate to different parts of the KG.

To further dive into mentions of different senses over time, stream graphs show the commodities with the most mentions in our corpus (coffee, tea, chocolate, and sugar) and how they evolve over time. Users can explore the authority of the extracted information, through an overview of the different sources and their reliability according to the LLM judges from the pipeline. Lastly, a map view offers the option to zoom in on particular geographical locations that could be extracted.

Beyond browsing the KG, visitors will also be able to see the pipeline in operation and generate a small thick KG from different sources.

5. Conclusions and Future Work

We have detailed a pipeline for constructing thick knowledge graphs, a site generator to create a browser for this knowledge graph, and we have shown how it can be operationalised by generating a knowledge graph from 18th century to contemporary sources. We have sketched a demonstration scenario on how users can browse through the knowledge graph using the interface, while inspecting commodities in different contexts, as well as the sources from which the claims were extracted.

Future work will focus on evaluation and quality control of the extraction pipeline and scaling the knowledge graph. In this paper, we focused on demonstrating the feasibility of the approach, but domain-specific information extraction models [13, 14] and clustering can potentially provide higher

⁶<https://www.gutenberg.org/>

quality knowledge graph statements. Entity linking could also be improved by specifically considering entity senses [15] and integration with commonsense knowledge graphs such as ConceptNet [16], Candle [17], and Mango [18].

Additionally, to further evaluate the data model and interface, a systematic user evaluation needs to be conducted to provide further insights into the useability of the current browser as well as avenues for other analyses and visualisations.

More detailed metadata regarding the sources and their editorial orientation is another area of future work and would provide further options for in-depth comparisons between sources. This requires more fine-grained corpus curation.

We hope that this demo encourages further development of methods and approaches for thick knowledge graphs such that tea as a medicine is not confused with tea as a social activity.

Acknowledgments

Funded by the European Union under grant agreement 101088548 - TRIFECTA. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. Things that helped get this demo and paper out: copious amounts of oolong tea and flat whites made from single origin Brazilian coffee beans.

Declaration on Generative AI

During the preparation of this work, the authors used Claude Code to support development and create documentation. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] C. Peng, F. Xia, M. Naseriparsa, F. Osborne, Knowledge graphs: Opportunities and challenges, *Artificial intelligence review* 56 (2023) 13071–13102.
- [2] M. Van Erp, Unflattening knowledge graphs, in: *Proceedings of the 12th Knowledge Capture Conference 2023*, 2023, pp. 223–224.
- [3] G. Ryle, The thinking of thoughts: What is 'le penseur' doing?, *Studies in Anthropology* 11 ([1968] 1996).
- [4] C. Geertz, The interpretation of cultures, Basic books, [1973] 2016.
- [5] A. Wong, M. Fox, M. Katsumi, Semantically interoperable census data: unlocking the semantics of census data using ontologies and linked data, *International Journal of Population Data Science* 9 (2024) 2378.
- [6] C. Shimizu, P. Hitzler, S. Gonzalez-Estrecha, J. Goeke-Smith, D. Rehberger, C. Foley, A. Sheill, The wikibase approach to the enslaved. org hub knowledge graph, in: *International Semantic Web Conference*, Springer, 2023, pp. 419–434.
- [7] P. Lisena, D. Schwabe, M. van Erp, R. Troncy, W. Tullett, I. Leemans, L. Marx, S. C. Ehrich, Capturing the semantics of smell: the odeuropa data model for olfactory heritage information, in: *European semantic web conference*, Springer, 2022, pp. 387–405.
- [8] E. Hyvönen, Digital humanities on the semantic web: Sampo model and portal series, *Semantic Web* 14 (2023) 729–744.
- [9] M. Song, S. H. Sim, R. Bhardwaj, H. L. Chieu, N. Majumder, S. Poria, Measuring and enhancing trustworthiness of llms in rag through grounded attributions and learning to refuse, in: *International Conference on Learning Representations*, volume 2025, 2025, pp. 30783–30823.
- [10] O. Gunes, T. Furche, G. Orsi, Structured aspect extraction, in: Y. Matsumoto, R. Prasad (Eds.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*:

Technical Papers, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 2321–2332. URL: <https://aclanthology.org/C16-1219/>.

- [11] J. Daiber, M. Jakob, C. Hokamp, P. N. Mendes, Improving efficiency and accuracy in multilingual entity extraction, in: Proceedings of the 9th International Conference on Semantic Systems (I-Semantics), 2013.
- [12] Anthropic, Claude haiku 4.5 system card, <https://www.anthropic.com/claude-sonnet-4-5-system-card>, 2025.
- [13] S. Verkijk, P. Vossen, P. Sommerauer, Language models lack temporal generalization and bigger is not better, in: Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 20629–20637.
- [14] T. Paccosi, M. Koolen, Smelling the past: Investigating historical models for olfactory event extraction, in: The Fourth Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2026), 2026.
- [15] M. van Erp, P. Groth, Towards entity spaces, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 2129–2137. URL: <https://aclanthology.org/2020.lrec-1.261/>.
- [16] R. Speer, J. Chin, C. Havasi, Conceptnet 5.5: An open multilingual graph of general knowledge, in: Proceedings of the AAAI conference on artificial intelligence, volume 31, 2017.
- [17] T.-P. Nguyen, S. Razniewski, A. Varde, G. Weikum, Extracting cultural commonsense knowledge at scale, in: Proceedings of the ACM web conference 2023, 2023, pp. 1907–1917.
- [18] T.-P. Nguyen, S. Razniewski, G. Weikum, Cultural commonsense knowledge for intercultural dialogues, in: Proceedings of the 33rd ACM international conference on information and knowledge management, 2024, pp. 1774–1784.