Improving Concept Specificity in LLM-Based Concept Linking for Oral History Segments Using the WO2 Thesaurus

Feruza Bakhtiyorova

Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
f.bakhtiyorova@student.vu.nl
(2740961)

Supervisor: Victor de Boer
Second reader: Lea Krause

Abstract. Digital oral history collections need metadata that makes long interviews searchable without flattening personal testimony. This thesis studies concept linking in WO2Net, a Dutch Second World War oral history project where interview segments are linked to concepts from the WO2 Thesaurus. Earlier WO2Net validation work showed that concept links can be too broad, missing, or ambiguous, especially when segments refer to specific camps, events, places, or organisations [1].
This thesis investigates whether retrieval from the WO2 Thesaurus improves large language model (LLM)-based concept linking for oral history segments. Three conditions were compared on five selected WO2Net interview segments: baseline prompting, refined prompting, and knowledge-graph retrieval-augmented generation (KG-RAG) prompting. All three conditions used GPT-5.5 Instant. The outputs were evaluated through a blind expert questionnaire completed by five WO2Net experts, who rated proposed concepts for correctness and specificity.
The results show that KG-RAG was most useful for improving specificity. Among applicable specificity ratings, KG-RAG reached 79.17%, compared with 74.36% for refined prompting and 65.62% for baseline prompting. It recovered specific missing concepts such as Reichenbach, Birkenau, Kristallnacht, and Bergen-Belsen. However, this specificity gain did not translate into higher overall correctness, because the expanded candidate list also introduced noisy or less central links. The thesis concludes that retrieval from the WO2 Thesaurus improves specificity in the selected cases, but does not automatically improve overall correctness.

Keywords: oral history · concept linking · knowledge graph · retrieval-augmented generation · large language models · WO2 Thesaurus · expert evaluation

1 Introduction

Digital oral history archives preserve personal accounts that are difficult to represent through short metadata labels. A single interview segment can contain

explicit historical references, such as camps, organisations, and events, while also communicating broader themes such as persecution, displacement, trauma, family separation, or survival. Metadata is therefore important for search and exploration: it gives users structured access points for finding relevant material within large interview collections [2,3]. In this thesis, concept linking means connecting an interview segment to one or more controlled vocabulary concepts. The challenge is that these links need to be accurate enough to support retrieval and specific enough to be useful.

This thesis studies concept linking in the context of WO2Net, a Dutch Second World War oral history project connected to the Oorlogsbronnen data infrastructure. Oorlogsbronnen provides access to Dutch Second World War source material, while the WO2 oral history matching pipeline processes oral history interview transcripts and enriches them with concepts from the WO2 Thesaurus [4,5]. The pipeline works with .vtt files, a subtitle/transcript format used to store time-aligned interview text. In the existing pipeline, these transcripts are segmented, relevant segments are selected, and the selected segments are linked to Second World War concepts using embeddings, exact matches, and LLM validation [5]. Earlier WO2Net validation work showed that this workflow can produce useful segment and concept suggestions, but also that concept links can be too broad, ambiguous, or missing [1]. In particular, prompt refinement can only select from the concepts available to the model. If a specific concept is missing from the candidate list, stricter prompting alone cannot recover it.

The central idea of this thesis is to use the WO2 Thesaurus not only as the controlled vocabulary from which final concept labels are selected, but also as a retrieval source before concept selection. In the previous WO2Net/Oorlogsbronnen matching pipeline, the WO2 Thesaurus is used as the source vocabulary for linking interview segments to concepts. This thesis uses the same thesaurus not only as the final vocabulary for concept labels, but also earlier in the concept-linking process to retrieve additional candidate concepts that may be missing from the original candidate list. This thesis proposes a knowledge-graph retrieval-augmented generation (KG-RAG) approach for the WO2Net concept-linking task. In this approach, candidate concepts are first retrieved from the WO2 Thesaurus, and the large language model (LLM) then selects which of these candidates are supported by the segment text. This changes the task from selecting among a fixed set of original matched concepts to selecting from a candidate list expanded through thesaurus retrieval. Related work on ontology-aware prompting and knowledge-source integration suggests that external structured resources can support LLMs in specialized domains, but also shows that such integration needs empirical evaluation rather than assuming uniform improvement [9,10,11].

The research question is:

To what extent does retrieval from the WO2 Thesaurus improve the correctness and specificity of LLM-based concept linking for oral history segments?

The contribution of this thesis is threefold. First, it implements a small KG-RAG pipeline for WO2Net concept linking using real WO2Net segments and the WO2 Thesaurus. Second, it compares KG-RAG with two prompting-only

conditions: the original baseline prompting and the refined prompt from earlier WO2Net work. Third, it evaluates the outputs through a focused expert-informed evaluation, using correctness and specificity as the main criteria. The thesis does not claim to solve oral history interpretation as a whole. It tests a narrower question: whether retrieval from a domain thesaurus improves the concepts selected for specific interview segments.

2 Related Work

2.1 Computational access to oral history collections

Oral history collections preserve personal memories, lived experience, and interpretations of the past, but they are difficult to access computationally. Interviews are long, often non-linear, and shaped by memory, emotion, silence, and conversational context. Metadata therefore plays a practical role: it helps researchers and the public identify where relevant historical material appears in large collections.

Cocciolo [2] studies the use of generative AI for oral history metadata in an LGBTQ+ archival context. The study shows that AI can support metadata creation, but also that AI-generated descriptions should be reviewed before publication. This is relevant here because this thesis also treats AI as a metadata support tool rather than as an unsupervised archival authority.

Chen et al. [3] apply NLP and data analytics to the Densho Digital Collection, a large oral history collection about Japanese American incarceration. Their work shows that computational methods can reveal spatial, thematic, and emotional patterns across oral history archives. However, it operates at collection level, while this thesis evaluates segment-level concept linking.

Together, these studies show that AI and NLP can support access to oral history collections, but that review and interpretation remain necessary. This thesis therefore evaluates concept linking as a support method whose outputs are checked by domain experts.

2.2 LLMs, ontologies, and knowledge graphs for information extraction

LLMs can perform information extraction tasks without extensive task-specific training, but they may struggle when a task depends on specialized domain knowledge. In oral history collections, this is especially relevant because interviews often contain references to places, events, organisations, social groups, and historical experiences that require contextual interpretation rather than simple keyword matching [2,3]. Similar issues appear in domain-specific information extraction, where relevant concepts may include specialized terms or common noun phrases that do not fit standard named entity categories [12,10].

Ontology-aware prompting offers one way to connect LLMs to domain structures. Zeginis et al. [11] evaluate ontology-aware zero-shot LLM prompting for

information extraction from Greek public administration documents. Although their study is not about oral history, it is methodologically relevant because it shows how an ontology can define the target structure for extracting metadata from unstructured text.

Knowledge graph and ontology integration has also been studied in specialized domains such as biomedicine. Chang and Sung [9] review the integration of SNOMED CT into LLM pipelines and show that external knowledge sources can improve performance, but not consistently across all studies and evaluation settings. This mixed evidence is important for this thesis because it motivates separate evaluation of correctness and specificity instead of assuming that a knowledge source improves every metric.

Xiao [10] frames ontology-guided, LLM-assisted information extraction as a response to long-tail domain knowledge and limited annotation data. This thesis follows the same broad logic at a smaller bachelor-project scale: it uses an external thesaurus to provide candidate concepts at inference time instead of fine-tuning an LLM.

2.3 Domain-specific concept recognition beyond standard named entities

Concept linking in WO2Net is related to named entity recognition (NER), but it is not identical to standard NER. Standard NER usually focuses on entities such as persons, organisations, and locations. WO2Net concept linking also includes camps, events, abstract themes, social categories, and thesaurus-specific labels. Some relevant concepts are proper names, such as Auschwitz or Philips. Others are common noun phrases or broader historical categories.

Chierchiello et al. [12] study ontology-guided domain entity recognition in environmental texts and distinguish ordinary named entities from domain-specific entities expressed through common noun phrases. This distinction also applies to WO2Net. Some segment concepts are named entities, but others require domain-specific recognition and controlled-vocabulary alignment.

For this thesis, the implication is that concept linking should not be evaluated only as whether the model recognizes named places. A good method also needs to choose the right level of thesaurus concept. For example, Bergen-Belsen is more specific and historically appropriate than Bergen in a camp-related segment. Similarly, Kristallnacht is more informative than only place or family-related concepts when the segment explicitly discusses that event.

2.4 Evaluation and reproducibility in LLM-based studies

LLM-based empirical studies require careful reporting because outputs can vary with model choice, prompt wording, candidate lists, and post-processing. Baltes et al. [13] provide guidelines for empirical studies involving LLMs and emphasize reporting the LLM role, model versions, prompts, human validation, baselines, and limitations. These recommendations align with the design choices in this

thesis: the model is fixed across conditions, prompts and outputs are stored, and human expert validation is used.

Krause [14] argues that conversational AI systems need to handle context, uncertainty, and structured representations carefully. This is relevant to oral history concept linking because a segment excerpt may not contain all context needed for broader interpretation. The current thesis therefore treats segment-level concept linking as one layer of metadata creation, while recognizing that interview-level context may be needed for thematic interpretation.

3 Method

This chapter describes the setup of the thesis before the results are reported. First, it defines the concept-linking task and the main terms used throughout the thesis. It then describes the data sources and case-selection procedure, followed by the three concept-linking conditions: baseline prompting, refined prompting, and KG-RAG prompting. Finally, it explains the expert evaluation design, rating scales, metrics, and analysis levels.

The evaluation in this thesis is based on five selected oral history segments rated by five WO2Net experts. The larger set of enriched segments and validation groups was used to identify and select informative cases, not as a separate full performance evaluation. The expert ratings are analysed in two complementary ways: first as an aggregate comparison between the three conditions, and second as an in-depth case-level analysis to understand why the methods behave differently.

3.1 Task definition and terminology

The task evaluated in this thesis is concept linking. Given one oral history segment, the system must select one or more relevant concepts from the WO2 Thesaurus. A good concept link should be correct, meaning that the concept is supported by the segment, and specific, meaning that the concept is neither too broad nor unnecessarily narrow for the segment.

A segment is a short excerpt from a longer interview transcript. In this thesis, the five expert-facing segment excerpts range from 121 to 177 words. A case is one selected segment together with its role in the evaluation, such as missing camp concepts, missing event concept, or place/camp disambiguation. A condition refers to one of the three concept-linking alternatives compared in this thesis: baseline prompting, refined prompting, and KG-RAG prompting.

The expert evaluation uses concept rows and expert ratings. A concept row is one row in the blind expert questionnaire, containing one proposed concept for one selected case. An expert rating is one judgement by one expert for one concept row. The main metrics are correctness and specificity. For specificity, the thesis also reports an applicable-specificity rate, which excludes ratings marked not applicable, unsure, missing, or blank.

3.2 Workflow, data sources, and case selection

Figure 1 gives an overview of the thesis workflow. The workflow starts from existing WO2Net/Oorlogsbronnen oral history data and focuses on the concept-linking stage. The same five selected segments are processed under three conditions: baseline prompting, refined prompting, and KG-RAG prompting. Baseline and refined prompting use the original candidate concepts produced by the previous WO2 oral history matching pipeline. KG-RAG adds an intermediate retrieval step from the WO2 Thesaurus before LLM-based concept selection. The outputs of the three conditions are then merged into a blind expert questionnaire and evaluated for correctness and specificity.

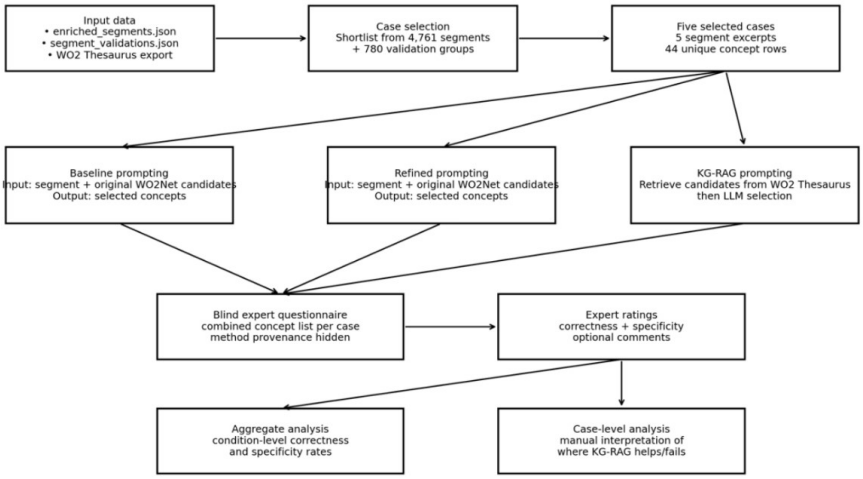


Fig. 1. Overview of the thesis workflow. The three conditions differ mainly in the source of candidate concepts: baseline and refined prompting use the original WO2Net candidate concepts, while KG-RAG retrieves additional candidates from the WO2 Thesaurus before concept selection.

The data used in this thesis comes from the WO2Net/Oorlogsbronnen oral history workflow. The earlier WO2 oral history matching pipeline processes oral history interview transcripts in .vtt format, creates transcript segments, selects relevant segments, and enriches them with concepts from a Second World War thesaurus [5]. The raw subtitle files are available in the public segmenten_ondertiteling repository [6]. This thesis uses local processed JSON and CSV files derived from these sources, so the experiments can be reproduced without querying the public data sources during each run.

The controlled vocabulary used for concept linking is the WO2 Thesaurus. The WO2 Thesaurus is available as linked data through the Oorlogsbronnen WO2 Thesaurus dataset and can also be queried through the WO2 Oorlogsbronnen SPARQL endpoint [7,8]. In this thesis, the thesaurus was used through

a local parsed export. This parsed version contained 17,820 concepts, including concept URIs, concept IDs, labels, concept types, broader concepts, narrower concepts, related concepts, and scheme membership. Concept URIs are treated as stable identifiers because different labels or languages can refer to the same thesaurus concept.

The local thesis pipeline used two main input files for case selection. The first file, `enriched_segments.json`, contains transcript segments produced by the earlier WO2 oral history matching pipeline together with automatically generated enrichment information, including candidate concept links. The second file, `segment_validations.json`, contains previous crowdsource validation information for segments and proposed concept links. These files were not used as a final evaluation dataset in this thesis. Instead, they were used to identify informative cases for a focused expert evaluation.

The goal of case selection was to choose a small set of segments that could test different concept-linking problems. The selected cases were not intended to form a statistically representative benchmark of the full WO2Net collection. Instead, they were chosen because previous validation information suggested different types of concept-linking behaviour: missing specific concepts, ambiguous concepts, concept conflicts, and one mostly accepted segment used as a control case.

For case selection, the local pipeline processed 4,761 enriched segments and 780 validation groups. A validation group refers to the previous validation information associated with a segment and its proposed concept links. This produced a shortlist of 487 candidate segments: 414 rejected segments, 38 conflict segments, and 35 accepted segments. In this context, rejected, conflict, and accepted refer to previous validation status. Rejected segments were segments for which previous validation indicated that the proposed output was problematic. Conflict segments had mixed validation signals. Accepted segments were segments for which previous validation indicated that the output was mostly suitable.

From this shortlist, five segments were selected for the expert evaluation. In this thesis, a case means one selected segment together with its role in the evaluation. The five cases cover different concept-linking situations: organisation/work concepts, missing camp concepts, a missing event concept, an accepted control case, and place/camp disambiguation. The five excerpts come from four interview contexts: two Lotty Huffener segments and one segment each from Annie Sulzbach, Rob Wurms, and Geertruida Zeehandelaar. The expert-facing excerpts are between 121 and 177 words.

Table 1 lists the selected cases and the reason each case was included.

Table 1: Selected evaluation cases and reasons for inclusion.

Case	Segment ID	Interviewee/ Context	Role	Reason for inclusion
Case 01	02%20stiso_ 20091008_ Huffener_Lotty_ 1.nl_8	Lotty Huffener	Organisation/work conflict case	Tests whether the method handles Philips and work/labour-related concepts.
Case 02	01_JKKV_2003_ LOTTY_ HUFFENER-Veffer-h264. nl_1	Lotty Huffener	Missing camp concepts case	Validators mentioned missing Reichenbach and Birkenau.
Case 03	04_JKKV_2003_ ANNIE_ SULZBACH-h264. nl_3	Annie Sulzbach	Missing event concept case	Validators mentioned that Kristallnacht was missing.
Case 04	08%20stiso_ 20091104_Wurms_ Rob_1.nl_10	Rob Wurms	Accepted control case	Tests whether KG-RAG avoids worsening a mostly accepted case.
Case 05	22%20stiso_ 20110501_ Zeehandelaar_ Geertruida_1.nl_ 4	Geertruida Zeehandelaar	Specific place/camp disambiguation case	Tests Bergen-Belsen versus Bergen and other camp/place concepts.

The code and selected case data for this step are available in the thesis repository under 03_thesis_kg_rag/scripts/03_select_evaluation_segments.py and 03_thesis_kg_rag/data/evaluation_segments.json.

3.3 Concept-linking conditions

All three conditions were applied to the same five selected cases using GPT-5.5 Instant. Keeping the model constant makes the comparison focus on the method setup rather than on model choice. The three conditions differ in two

main ways: the source of the candidate concepts and the prompt logic used to select concepts.

Baseline prompting and refined prompting are prompting-only conditions. They receive the segment excerpt and the original candidate concepts from the previous WO2 oral history matching pipeline. KG-RAG differs because it first retrieves additional candidate concepts from the WO2 Thesaurus and then asks the LLM to select the supported concepts from this retrieved list.

Table 2 summarizes the three conditions.

Table 2: Overview of the three concept-linking conditions.

Condition	**Input to LLM**	**Candidate source**	**Main mechanism**	**Output**
Baseline prompting	Segment excerpt and original candidate concepts	Previous WO2 oral history matching pipeline	Original archived prompt logic	Selected concepts from the original candidate list
Refined prompting	Segment excerpt and original candidate concepts	Previous WO2 oral history matching pipeline	Stricter prompt logic from earlier prompt-refinement work	Selected concepts from the original candidate list
KG-RAG prompting	Segment excerpt and retrieved thesaurus candidates	WO2 Thesaurus retrieval	Candidate retrieval followed by LLM-based concept selection	Selected concepts from the retrieved candidate list

The exact prompt templates and prompt-file locations are reported in Appendix A.

3.3.1 Baseline prompting The baseline condition represents the original prompt-based concept-linking approach from earlier WO2Net work. For each selected case, the prompt receives two inputs: the segment excerpt and the original candidate concepts already produced by the previous WO2 oral history matching pipeline. The model is then asked to select the candidate concepts that are supported by the segment.

The important limitation of this condition is that it can only choose from the original candidate list. If a relevant concept is missing from that list, the baseline prompt cannot recover it. This makes the baseline condition useful for

testing the quality of the earlier prompt logic, but not for testing whether new thesaurus concepts can be added.

For the baseline condition, prompts were generated from the selected segment excerpts and the original WO2Net candidate concepts. The prompt asked the model to choose only from this given candidate list and to return the selected concepts in a structured output format. The prompt and output locations are listed in Appendix A and Appendix D.

3.3.2 Refined prompting The refined condition uses the improved top-down prompt from earlier WO2Net prompt-refinement work. Like the baseline condition, it receives the segment excerpt and the original candidate concepts from the previous WO2 oral history matching pipeline. It therefore uses the same candidate source as the baseline condition.

The difference between baseline prompting and refined prompting is the prompt logic. The refined prompt applies stricter selection instructions: the model is asked to select only concepts that are supported by the segment and to avoid weak or unsupported matches. This condition tests whether better prompt instructions improve concept selection when the candidate list stays the same.

For the refined condition, prompts were generated from the same segment excerpts and original WO2Net candidate concepts as the baseline condition. The difference is that the refined prompt uses stricter top-down selection rules. It asks the model to prefer concepts that are clearly supported by the segment, avoid weak or incidental matches, and be more conservative with generic concepts. The prompt and output locations are listed in Appendix A and Appendix D.

3.3.3 KG-RAG prompting The KG-RAG condition changes the candidate source. Instead of using only the original candidate concepts from the previous WO2 oral history matching pipeline, it retrieves a new candidate list from the local parsed WO2 Thesaurus before LLM-based selection. In this thesis, KG-RAG means knowledge-graph retrieval-augmented generation: the thesaurus is used as a structured retrieval source, and the LLM is then used to decide which retrieved concepts are actually supported by the segment.

The KG-RAG retrieval step has four layers. First, concepts already matched by the previous WO2Net pipeline are kept as seed candidates. This prevents useful original concepts from being lost when they are not literally mentioned in the segment. Second, the retriever performs explicit lexical matching between the segment text and searchable WO2 Thesaurus labels. This layer retrieves concepts whose labels are directly mentioned in the segment, such as named places, camps, organisations, or events. Weak one-word labels and common stopwords are filtered to reduce noisy matches.

Third, the retriever applies contextual event retrieval. This layer searches specifically for event concepts that may be supported by combinations of textual clues, such as dates, months, locations, and event-related words. For example, a segment mentioning a date, a place, and terms related to bombing or fire may support an event candidate even when the exact event label is not quoted

literally. Fourth, the retriever performs a conservative event graph expansion. For the strongest contextual event candidates, it may add a small number of related or broader event concepts from the thesaurus graph. This expansion is restricted to event concepts and is intentionally limited, so that the prompt is not filled with loosely related concepts.

The retrieved candidates from these layers are merged by concept ID, ranked by retrieval score and source priority, and limited to a maximum of 40 candidates per segment. Each candidate is stored with its label, URI, concept type, scheme, retrieval score, candidate source, and retrieval reason. This makes the candidate list inspectable before the LLM selection step.

In the second stage, the segment excerpt and retrieved candidate concepts are inserted into the KG-RAG prompt. For each candidate, the prompt includes the concept label, URI, and concept type. Retrieval scores, retrieval sources, and retrieval reasons are not shown to the LLM, because these fields are useful for inspection but could bias the model's selection. The model is instructed to select only concepts that are supported by the segment text, reject noisy or weakly related candidates, and prefer specific concepts when the segment supports them. In this setup, retrieval expands the space of possible concepts, while the LLM selection step filters the retrieved candidates.

This setup tests whether retrieval from the WO2 Thesaurus helps the model recover specific concepts that are missing from the original WO2Net candidate list. It also tests the main risk of KG-RAG: expanding the candidate list may improve specificity, but it may also introduce weak or only incidentally related concepts if retrieval and selection are not strict enough.

To avoid evaluation leakage, previous validation comments, consensus status, removed concepts, and missing-concept comments were not included in the retrieval input or in the KG-RAG prompts. This means the model did not receive the same validation information that was later used to interpret why the cases were selected.

3.4 Evaluation design

The evaluation in this thesis is based on a blind expert questionnaire completed by five WO2Net experts familiar with oral history annotation and Second World War data. The questionnaire evaluated the concept links produced by the three conditions for the five selected cases. The larger set of 4,761 enriched segments and 780 validation groups was used only for shortlisting and case selection, as described in Section 3.2. It was not used as a separate full performance evaluation.

The expert ratings are analysed at two levels. First, the ratings are aggregated by condition to compare baseline prompting, refined prompting, and KG-RAG. This aggregate comparison answers the main research question by showing how the three conditions differ in correctness and specificity. Second, the same expert ratings are examined at case level to understand why the aggregate results look the way they do. The case-level analysis is therefore an in-depth interpretation of the five selected cases, not a separate evaluation dataset.

The expert-facing questionnaire combined all concepts selected by the three conditions into one blind list per case. Experts did not see whether a concept came from baseline prompting, refined prompting, KG-RAG, or more than one condition. Method provenance was stored only in an internal table and mapped back after the questionnaire responses were collected. This design was used to reduce bias and to make experts judge the concepts themselves rather than the methods that produced them.

3.4.1 Expert questionnaire and rating scales The expert questionnaire was prepared as an Excel workbook with three sheets: Instructions, Case Overview, and Expert Evaluation. The Instructions sheet explained the task and rating categories. The Case Overview sheet listed the five cases. The Expert Evaluation sheet contained 44 blind concept rows. Each concept row showed the case ID, concept ID, segment excerpt, concept label, and concept URI, followed by empty fields for ratings and optional comments.

The experts were asked to judge each proposed concept on two dimensions: correctness and specificity. Correctness asked whether the concept was supported by the segment. The allowed correctness values were: incorrect, partially correct, correct, and unsure. Specificity asked whether the concept was at the right level of detail for the segment. The allowed specificity values were: too broad, appropriate, too narrow, not applicable, and unsure. Experts could also add optional comments, including missing-concept suggestions and general comments.

The questionnaire did not ask experts for their name or email address. Consent forms were collected separately from the questionnaire responses and were not included in the processed analysis files. In the processed dataset, experts were anonymized as expert_01 to expert_05. The structure of the questionnaire, including the field names, rating values, and screenshots of the expert-facing workbook, is shown in Appendix B.

3.4.2 Metrics and analysis The evaluation uses correctness and specificity as the two main criteria. Correctness measures whether a proposed concept is supported by the segment. Specificity measures whether the proposed concept is at the right level of detail.

Correctness is reported in two ways. The strict correct rate is the proportion of expert ratings marked correct. The partial-credit correctness score assigns correct = 1, partially correct = 0.5, and incorrect = 0. Ratings marked unsure or missing are excluded from the partial-credit calculation.

Specificity is also reported in two ways. The appropriate-specificity rate across all ratings is the proportion of all specificity ratings marked appropriate. The applicable-specificity rate is calculated after excluding ratings marked not applicable, unsure, missing, or blank. This metric is useful because some concepts cannot meaningfully be judged for specificity if the expert first decides that specificity is not applicable or is unsure.

Because the three conditions produced different numbers of selected concepts, raw counts alone are not sufficient for comparison. A condition that selects more

concepts has more opportunities to receive both correct and incorrect ratings. For this reason, the thesis reports rates rather than only absolute counts.

One additional complication is that the expert questionnaire used one combined blind concept list per case. If the same concept was selected by more than one condition, a single expert rating for that concept contributes to each relevant condition after method provenance is mapped back. The aggregate comparison by condition is therefore descriptive rather than a statistical test with independent observations.

After the expert responses were collected, the ratings were processed into summary tables by condition, case, concept, and comment theme. Shared concept rows were expanded back to each condition that had selected the concept. The main processed output files used in the Results section are listed in Appendix D.

4 Results

This chapter reports the results of the three concept-linking conditions. Section 4.1 first shows what each condition selected for the five cases. Section 4.2 reports the aggregate expert-rating results by condition. Section 4.3 then examines the five cases in more detail to explain the main patterns in the aggregate results. Section 4.4 summarizes the qualitative expert comments.

4.1 Outputs of the three concept-linking conditions

Before discussing expert ratings, it is useful to compare the raw selected concepts produced by the three conditions. Table 3 shows the concepts selected by baseline prompting, refined prompting, and KG-RAG for each case. These are the outputs that were later merged into the blind expert questionnaire.

Table 3: Selected concepts by case and condition.

Case	**Baseline selected concepts**	**Refined selected concepts**	**KG-RAG selected concepts**
Case 01	Gevangenen; Dwangarbeid; Philips	Philips; Dwangarbeid; Gevangenen	Philips
Case 02	Vught; Auschwitz; Kamp Vught; Transporten; Gevangenen	Kamp Vught; Auschwitz	Kamp Vught; Auschwitz; Reichenbach; Birkenau; Amsterdam

Table 3: Selected concepts by case and condition.

Case	**Baseline selected concepts**	**Refined selected concepts**	**KG-RAG selected concepts**
Case 03	Familieleven; Huisvesting; Woningen; Verhuizingen; Amsterdam	Amsterdam; Familieleven; Huisvesting; Woningen; Verhuizingen	Kristallnacht; Frankfurt am Mainz; Duitsland; Amsterdam; Familienleben; Wohnraumbeschaffung; Verhuizingen; Residence
Case 04	Auschwitz; Sobibor; Barakken; Oorlogsgetroffenen; Monnickendam	Auschwitz; Sobibor; Barakken; Monnickendam	Gas chambers; Baracken; Ondergedoken; Auschwitz; Auschwitz-Birkenau; Sobibor; Krakau; Oswiecim; Netherlands; Utrecht
Case 05	Sobibor; Westerbork; Bergen; Auschwitz; Pearl Harbor	Sobibor; Westerbork; Bergen; Auschwitz; Pearl Harbor	Red Cross; Contact Holland; Auschwitz; Bergen-Belsen; Sobibor; Pearl Harbor; Westerbork; Overleden

The output comparison shows the main difference between the prompting-only conditions and KG-RAG. Baseline and refined prompting can only select from the original candidate concepts provided by the previous WO2 oral history matching pipeline. KG-RAG can select concepts that were retrieved from the WO2 Thesaurus and were not available to the prompting-only conditions.

This difference is visible in several cases. In Case 02, KG-RAG selected Reichenbach and Birkenau, while the prompting-only conditions did not. In Case 03, KG-RAG selected Kristallnacht, while baseline and refined prompting selected family, housing, moving, and Amsterdam-related concepts. In Case 05, KG-RAG selected Bergen-Belsen and did not select Bergen, while both prompting-only conditions selected Bergen. These outputs suggest where KG-RAG may improve specificity. At the same time, KG-RAG often produced longer concept lists, especially in Cases 03, 04, and 05. Whether these additional concepts are useful or noisy is evaluated through the expert ratings reported below.

4.2 Aggregate expert-rating results by condition

Table 4 summarizes the expert ratings by condition. The table is based on `expert_rating_summary_by_method.csv`, generated by `14_process_expert_evaluation.py`. Although the output file uses the word 'method", the thesis refers to the three alternatives as conditions: baseline prompting, refined prompting, and KG-RAG. The column 'unique concept rows" refers to the number of blind questionnaire rows associated with each condition after method provenance was mapped back from the internal table.

Table 4: Expert evaluation summary by condition. Correctness is reported as strict correctness and partial-credit correctness. Specificity is reported across all ratings and among applicable ratings only.

Condition	**Expert ratings**	**Unique concept rows**	**Strict correct rate**	**Partial-credit correctness**	**Appropriate specificity, all ratings**	**Appropriate specificity, applicable only**
Baseline	115	23	66.96%	75.23%	54.78%	65.62%
Refined	95	19	68.42%	77.47%	61.05%	74.36%
KG-RAG	160	32	65.62%	74.68%	59.38%	79.17%

The aggregate results show that KG-RAG did not improve overall correctness in these five cases. Refined prompting achieved the highest strict correct rate at 68.42%, followed by baseline prompting at 66.96% and KG-RAG at 65.62%. The same pattern appears for partial-credit correctness: refined prompting scored 77.47%, baseline prompting scored 75.23%, and KG-RAG scored 74.68%.

The strongest result for KG-RAG is specificity. Among applicable specificity ratings, KG-RAG reached 79.17%, compared with 74.36% for refined prompting and 65.62% for baseline prompting. This means that when experts considered a specificity judgement applicable, they more often judged KG-RAG concepts to be at the right level of detail. However, this specificity advantage did not translate into a higher correctness score.

These percentages should be interpreted descriptively. The evaluation contains five selected cases, and the method-expanded ratings are not statistically independent because the same concept row can contribute to more than one condition when multiple conditions selected the same concept.

4.3 Case-level in-depth results

The aggregate results show a trade-off, but they do not explain which concepts caused it. For that reason, the same expert-rating data was also examined at

case level. Table 5 summarizes the main KG-RAG effect in each case. The full numeric case-level summary is reported in Appendix C.

Table 5: Qualitative case-level summary of the main KG-RAG effects.

Case	**Main KG-RAG effect**	**Main limitation**
Case 01: Philips/work	Selected Philips conservatively and received relatively strong correctness.	Did not fully capture the work/labour context discussed by experts.
Case 02: missing camps	Recovered Reichenbach and Birkenau; both received positive expert ratings.	This was the clearest success case.
Case 03: missing event	Recovered Kristallnacht, the known missing event concept.	Also selected several additional concepts judged less relevant or too broad.
Case 04: accepted control	Added several specific concepts for an already accepted segment.	Over-expanded the output compared with refined prompting.
Case 05: Bergen-Belsen	Selected Bergen-Belsen rather than Bergen and recovered Red Cross.	Also introduced debatable concepts such as Contact Holland and Overleden.

Case 02 shows the clearest benefit of candidate expansion. KG-RAG selected Birkenau and Reichenbach, which were not selected by the prompting-only conditions because they were not available in the original candidate list. Expert ratings supported both concepts with the majority judgments of correct and appropriate specificity. This case demonstrates the central advantage of KG-RAG: it can recover specific missing concepts when thesaurus retrieval adds them to the candidate list.

Case 03 shows a more mixed result. KG-RAG retrieved and selected Kristallnacht, which was the known missing event concept. This is an important qualitative success. However, the KG-RAG output also selected additional location, family, housing, and residence-related concepts. Experts judged several of these as less relevant or too broad. As a result, KG-RAG had a lower strict correctness rate than the two prompting-only methods for this case, even though it recovered the key missing event concept.

Case 05 shows the disambiguation value of KG-RAG. The KG-RAG output selected Bergen-Belsen and rejected Bergen. This directly addresses the validation issue for which the case was selected. At the same time, KG-RAG also selected additional concepts such as Contact Holland and Overleden, which experts considered more debatable. The case therefore supports the same general

conclusion as Case 03: KG-RAG improves access to specific concepts but needs filtering to avoid less central links.

Case 01 and Case 04 show that KG-RAG is not uniformly better. In Case 01, KG-RAG selected only Philips, which made the output conservative and relatively correct, but it did not fully capture the broader work/labour context. In Case 04, KG-RAG selected many concepts for an already accepted segment. Some were useful and specific, but the increased number of concept links reduced the strict correctness rate compared with refined prompting.

4.4 Expert comments

The expert comments provide an important qualitative layer beyond the numerical ratings. Several comments point to concept-specific issues that are not fully visible in the aggregate metrics. For example, experts suggested that Philips-Kommando may be more specific than Philips in the organisation/work case. Experts also noted that Vught can refer to the city rather than the concentration camp, which illustrates the importance of disambiguating place names from camp-related concepts.

One follow-up comment raised a broader limitation of the evaluation setup. The expert argued that some segments are difficult to judge without access to the wider interview context, and that the main subject of a segment may sometimes lie at a broader thematic level than the proposed keywords suggest. Examples mentioned in this feedback include religious differences, persecution, emigration, post-war trauma, and anti-Jewish measures in the Netherlands. This feedback suggests that segment-level concept linking can recover explicit entities and events, but may miss broader interpretive themes that require more contextual understanding.

The comment analysis therefore supports two conclusions. First, the KG-RAG approach appears useful for recovering explicit missing concepts from the thesaurus. Second, concept linking at the segment level does not fully address the thematic interpretation of oral history testimony. A wider interview context, or a separate thematic annotation step, would be needed for that task.

5 Discussion

5.1 Answering the research question

The research question asked to what extent retrieval from the WO2 Thesaurus improves the correctness and specificity of LLM-based concept linking for oral history segments. The results show a mixed answer. In the selected cases, KG-RAG improved specificity, but it did not improve overall correctness.

The strongest KG-RAG result was the applicable-specificity score. When experts considered a specificity judgement applicable, KG-RAG concepts were more often judged to be at the right level of detail than concepts from baseline or refined prompting. However, KG-RAG did not achieve the highest correctness

score. Refined prompting had the highest strict correctness rate and partial-credit correctness score. This means that retrieval from the WO2 Thesaurus should not be interpreted as a general improvement in concept-linking quality. It is better understood as a candidate-expansion strategy that can improve specificity when relevant retrieved concepts are available.

5.2 Interpreting the KG-RAG trade-off

KG-RAG is most useful when the original candidate list does not containsectionInterpreting the KG-RAG trade-off

KG-RAG is most useful a specific concept that is directly supported by the segment. This was clearest in Case 02, where KG-RAG selected Reichenbach and Birkenau, and in Case 05, where it selected Bergen-Belsen instead of the broader or ambiguous Bergen. Case 03 showed a similar benefit for event concepts because KG-RAG selected Kristallnacht.

At the same time, the same expansion mechanism also explains why KG-RAG did not improve correctness overall. By retrieving more candidate concepts, KG-RAG gives the model more opportunities to select useful specific concepts, but also more opportunities to select weak, broad, or only incidentally related concepts. Case 03 illustrates this clearly: KG-RAG recovered Kristallnacht, but also selected additional family, housing, residence, and location concepts that reduced its correctness score. The results therefore suggest that retrieval alone is not enough. A KG-RAG pipeline also needs strict filtering and selection so that additional candidates improve specificity without adding too much noise.

5.3 Segment-level concept linking and wider interview context

The evaluation also shows a boundary of segment-level concept linking. The current method receives a segment excerpt and a list of candidate concepts. This setup is suitable for identifying explicit places, camps, organisations, and events that are mentioned or clearly implied in the excerpt. It is less suitable for identifying broader themes that require knowledge of the full interview or the narrator's life story.

This limitation matters for oral history. A segment that mentions a place, camp, or event may also communicate broader themes such as persecution, trauma, migration, family separation, religious identity, or survival strategies. Expert feedback showed that some relevant meanings depended on wider context and were not fully captured by the proposed concept links. This connects to work on context in conversational AI, where structured representations can support interpretation but cannot replace situational and contextual understanding [14].

For WO2Net, this suggests that concept linking and thematic interpretation should be treated as related but different tasks. Concept linking can support search and browsing through controlled thesaurus concepts, while thematic interpretation may require surrounding segments, full-interview summaries, or a separate annotation step.

5.4 Limitations and future work

This thesis has several limitations. First, the evaluation is based on five selected cases. These cases were chosen because they represent known concept-linking problems, not because they statistically represent the full WO2Net collection. Second, the evaluation involved five WO2Net experts. This provides useful domain-informed judgement, but it is still a small expert sample. The findings should therefore be read as a focused expert-informed case study rather than as broad statistical evidence.

Third, the KG-RAG retrieval step was fixed after limited refinement. The final retrieval version was sufficient for testing the main research question, but it still produced noisy candidates in some cases. Better filtering of technical concepts, broad lexical matches, multilingual labels, and weak contextual matches could improve future versions. Fourth, the same LLM was used across all three conditions. This controlled for model variation, but it means the thesis does not show whether the same pattern would appear with other models. Finally, the evaluation focused on correctness and specificity, not on usefulness for end users.

Future work should therefore scale the evaluation to more segments and experts, improve retrieval and filtering, normalize concept labels before expert review, and test the method with other LLMs or open models. A future WO2Net pipeline could also separate segment-level concept linking from interview-level thematic annotation, so that explicit thesaurus concepts and broader historical meanings are handled as related but distinct tasks.

6 Conclusion

This thesis evaluated whether retrieval from the WO2 Thesaurus improves LLM-based concept linking for WO2Net oral history segments. The study compared three conditions on the same five selected cases: baseline prompting, refined prompting, and KG-RAG prompting. All three conditions used the same LLM, so the comparison focused on the method setup rather than model choice. The outputs were evaluated through a blind expert questionnaire completed by five WO2Net experts.

The results show that retrieval from the WO2 Thesaurus improved specificity in the selected cases, but did not improve overall correctness. KG-RAG achieved the highest applicable-specificity rate, meaning that experts more often judged its concepts to be at the right level of detail when a specificity judgement was applicable. However, KG-RAG did not achieve the highest correctness rate. Refined prompting performed slightly better on both strict correctness and partial-credit correctness.

The case-level analysis explains this trade-off. KG-RAG was useful when the original candidate list lacked specific concepts that were supported by the segment. It recovered concepts such as Reichenbach, Birkenau, Kristallnacht, and Bergen-Belsen, which the prompting-only conditions did not select. At the same time, the expanded candidate list also introduced more opportunities for weak,

broad, or only incidentally related concepts. This reduced the overall correctness of the KG-RAG output in some cases.

The answer to the research question is therefore limited but clear: retrieval from the WO2 Thesaurus can improve the specificity of LLM-based concept linking when relevant missing concepts are retrieved and made available to the model. However, retrieval alone does not guarantee a more correct concept set. For WO2Net, KG-RAG is best understood as a candidate-expansion and specificity-support method that still requires careful retrieval design, strict LLM selection, URI-based normalization, and expert validation.

Reflection / Use of Generative AI

Generative AI was used as a support tool during this thesis for wording, restructuring, debugging, and revision. It was not used as a source of expert judgment or as a replacement for the evaluation. The research design, implementation decisions, data processing, interpretation of results, and final thesis text were reviewed and revised by the author. All reported results come from the documented scripts and processed data files in the thesis repository.

This project highlighted both the potential and the limits of using large language models in cultural heritage workflows. A key insight was that model outputs may appear useful at a general level while still lacking the specificity needed for archival search and interpretation. Working with WO2 Thesaurus retrieval showed that external knowledge sources can help recover more specific concepts, but also that retrieval does not automatically make the final concept set more correct. The project therefore showed the value of a middle ground: lightweight retrieval and prompting can support metadata creation, but expert evaluation remains necessary for historically sensitive oral history material.

References

1. Bakhtiyorova, F., Boon, D.: Optimizing AI-driven segmentation of oral history: A feedback-loop approach using crowdsourced validation in WO2Net. Course paper, Vrije Universiteit Amsterdam, WO2Net (2026). https://www.wo2net.nl/app/uploads/2026/02/Group1_Final_Paper.pdf
2. Cocciolo, A.: Oral history metadata and AI: A study from an LGBTQ+ archival context. Preservation, Digital Technology & Culture 54(1), 27–33 (2025). https://doi.org/10.1515/pdtc-2024-0054
3. Chen, H., Kim, J.A., Chen, J., Sakata, A.: Demystifying oral history with natural language processing and data analytics: A case study of the Densho digital collection. The Electronic Library (2024). https://doi.org/10.1108/EL-12-2023-0303
4. Oorlogsbronnen: Bronnen. https://www.oorlogsbronnen.nl/bronnen
5. Oorlogsbronnen: WO2 oral history matching pipeline. GitHub repository (2025). https://github.com/Oorlogsbronnen/wo2-oral-history-matching-pipeline
6. Oorlogsbronnen: Segmenten ondertiteling. GitHub repository (2025). https://github.com/Oorlogsbronnen/segmenten_ondertiteling

7. Oorlogsbronnen: WO2 Thesaurus dataset. Linked data dataset, Spinque Data Catalog (2025). https://data.spinque.com/ld/data/oorlogsbronnen/wo2_thesaurus/
8. LDMax: WO2 Oorlogsbronnen SPARQL endpoint. https://query.ldmax.nl/wo2-oorlogsbronnen/nfeAGR
9. Chang, E., Sung, S.: Use of SNOMED CT in large language models: Scoping review. JMIR Medical Informatics 12, Article e62924 (2024). https://doi.org/10.2196/62924
10. Xiao, J.: An ontology-guided, language model-assisted approach for theme-specific information extraction. Doctoral dissertation, University of Illinois Urbana-Champaign, IDEALS (2025). https://hdl.handle.net/2142/129450
11. Zeginis, D., Kalampokis, E., Tarabanis, K.: Applying an ontology-aware zero-shot LLM prompting approach for information extraction in Greek: The case of DI-AVGEIA.gov.gr. In: Proceedings of the 28th Pan-Hellenic Conference on Progress in Computing and Informatics. ACM (2024). https://doi.org/10.1145/3716554.3716603
12. Chierchiello, E., Chiril, P., Pagano, A.: Ontology-guided domain entity recognition in environmental texts: Evaluating syntax-driven and LLM approaches using BabelNet and GEMET. In: Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), pp. 233–244. CEUR Workshop Proceedings (2025). https://aclanthology.org/2025.clicit-1.25/
13. Baltes, S., Angermeir, F., Arora, C., Muñoz Barón, M., Chen, C., Böhme, L., Calefato, F., Ernst, N., Falessi, D., Fitzgerald, B., Fucci, D., He, J., Treude, C., Kalinowski, M., Lambiase, S., Russo, D., Lungu, M., Martinez Montes, C., Prechelt, L., Ralph, P., van Tonder, R., Wagner, S.: Guidelines for empirical studies in software engineering involving large language models. arXiv (2026). https://arxiv.org/abs/2508.15503v7
14. Krause, L.: Contextualising conversational AI. Doctoral dissertation, Vrije Universiteit Amsterdam (2025). https://doi.org/10.5463/thesis.1210

A Prompt reporting and prompt-file locations

The three concept-linking conditions used different prompt setups. The exact generated prompt files are stored in the thesis repository under `03_thesis_kg_rag/results/llm_prompts/`. This appendix reports the prompt locations and the structure of each prompt. The full generated prompt text can be inspected in the repository.

Table 6: Prompt locations and prompt structure.

Condition	**Prompt folder**	**Prompt structure**
Baseline prompting	`03_thesis_kg_rag/results/llm_prompts/baseline/`	Segment excerpt; original WO2Net candidate concepts; instruction to select supported concepts; JSON output format.

Table 6: Prompt locations and prompt structure.

Condition	**Prompt folder**	**Prompt structure**
Refined prompting	`03_thesis_kg_rag/results/llm_prompts/refined/`	Segment excerpt; original WO2Net candidate concepts; stricter top-down selection rules; JSON output format.
KG-RAG prompting	`03_thesis_kg_rag/results/llm_prompts/kg_rag/`	Segment excerpt; retrieved WO2 Thesaurus candidate concepts with labels and URIs; instruction to select supported concepts, reject noisy candidates, prefer specific supported concepts, and return JSON output.

Table 7: Main scripts used to generate prompts and outputs.

Step	**Script**	**Purpose**
Case selection	`03_thesis_kg_rag/scripts/03_select_evaluation_segments.py`	Selects the five evaluation cases from the larger set of enriched segments and validation groups.
KG-RAG retrieval logic	`03_thesis_kg_rag/scripts/retrieve_candidates.py`	Defines the retrieval layers used for KG-RAG candidate retrieval: original seeds, explicit lexical retrieval, contextual event retrieval, and conservative event graph expansion.
KG-RAG candidate retrieval	`03_thesis_kg_rag/scripts/04_run_candidate_retrieval.py`	Retrieves candidate concepts from the local parsed WO2 Thesaurus for each selected segment.

Table 7: Main scripts used to generate prompts and outputs.

Step	**Script**	**Purpose**
KG-RAG prompt generation	`03_thesis_kg_rag/scripts/05_build_kg_rag_prompts.py`	Builds KG-RAG prompts from the selected segments and retrieved thesaurus candidates.
Refined prompt generation	`03_thesis_kg_rag/scripts/08_build_refined_prompts.py`	Builds refined prompts using the original candidate concepts and the refined prompt logic.
Baseline prompt generation	`03_thesis_kg_rag/scripts/10_build_baseline_prompts.py`	Builds baseline prompts using the original candidate concepts and archived baseline prompt logic.
Expert evaluation processing	`03_thesis_kg_rag/scripts/14_process_expert_evaluation.py`	Processes expert questionnaire responses and generates summary tables for the Results section.

B Expert evaluation questionnaire structure

The expert evaluation was prepared as an Excel workbook. The workbook contained three sheets: Instructions, Case Overview, and Expert Evaluation. Experts only saw the blind questionnaire and did not see which condition produced each concept. Method provenance was stored separately in `internal_expert_evaluation_table.csv` and mapped back after the responses were collected.

Table 8: Expert evaluation workbook structure.

Sheet	**Purpose**
Instructions	Explains the evaluation task and the meaning of the rating categories.
Case Overview	Lists the five selected cases and gives experts an overview of the evaluated segments.
Expert Evaluation	Contains the blind concept rows to be rated by the experts.

Table 9: Fields in the Expert Evaluation sheet.

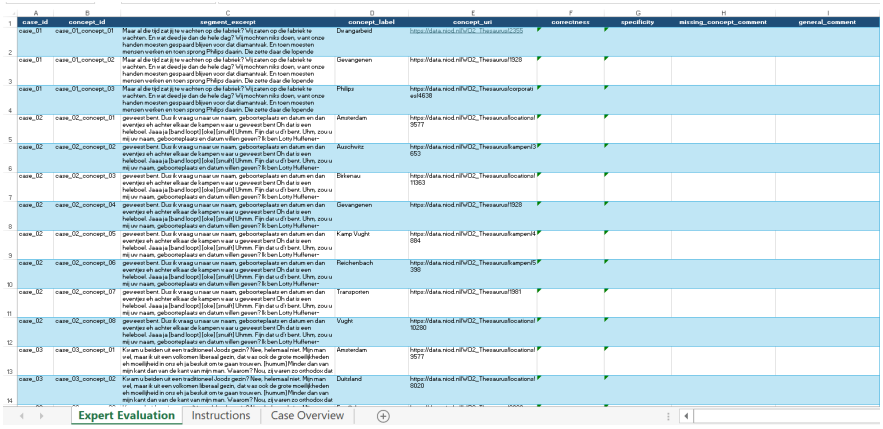
Field	**Meaning**
case_id	Identifier for the selected case, e.g. case_01.
concept_id	Identifier for the concept row within the case, e.g. case_01_concept_03.
segment_excerpt	Segment text shown to the expert.
concept_label	Human-readable label of the proposed WO2 Thesaurus concept.
concept_uri	Stable URI of the proposed concept where available.
correctness	Expert rating of whether the concept is supported by the segment.
specificity	Expert rating of whether the concept is at the right level of detail.
missing_concept_comment	Optional field for missing or alternative concepts.
general_comment	Optional free-text comment.

Table 10: Rating values used in the questionnaire.

Rating dimension	**Allowed values**
Correctness	incorrect; partially correct; correct; unsure
Specificity	too broad; appropriate; too narrow; not applicable; unsure

The Expert Evaluation sheet contained 44 blind concept rows. Each row represented one proposed concept for one selected case. Because the questionnaire was blind, experts did not see which condition produced each concept. After the responses were collected, each concept row was linked back to one or more conditions using the internal method-provenance table.

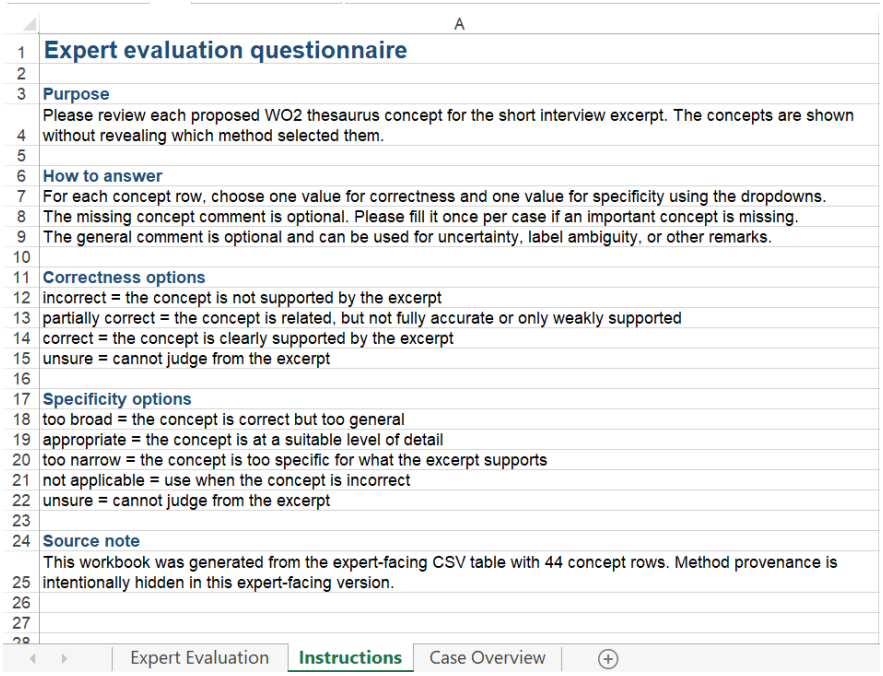
Figures 2, 3 and 4 show the structure of the expert evaluation workbook. The workbook contained three sheets: Instructions, Case Overview, and Expert Evaluation. The screenshots are included to make clear what experts saw during the blind evaluation. Method provenance was not visible in the expert-facing workbook.



case_id	concept_id	segment_excerpt	concept_label	concept_uri	correctness	specificity	missing_concept_comment	general_comment
case_01	case_01_concept_01	Maar al die tijd zat jij te wachten op die fabriek? Vij zaten op die fabriek te wachten. En vrar deedje dan de hele dag? Vijmochten niks doen, vrant onze handen moesten gespaard blijven voor dat diamantvak. En toen moesten mensen vellen en toen sprong Philips daarin. Die zette daar die topende	Dwangarbeid	https://data.niod.nl/WO2_Thesaurus/id/355				
case_01	case_01_concept_02	Maar al die tijd zat jij te wachten op die fabriek? Vij zaten op die fabriek te wachten. En vrar deedje dan de hele dag? Vijmochten niks doen, vrant onze handen moesten gespaard blijven voor dat diamantvak. En toen moesten mensen vellen en toen sprong Philips daarin. Die zette daar die topende	Gevangenen	https://data.niod.nl/WO2_Thesaurus/id/1928				
case_01	case_01_concept_03	Maar al die tijd zat jij te wachten op die fabriek? Vij zaten op die fabriek te wachten. En vrar deedje dan de hele dag? Vijmochten niks doen, vrant onze handen moesten gespaard blijven voor dat diamantvak. En toen moesten mensen vellen en toen sprong Philips daarin. Die zette daar die topende	Philips	https://data.niod.nl/WO2_Thesaurus/id/corporation/id/4638				
case_02	case_02_concept_01	gewreest bent. Dus ik vraag u naar uw naam, geboorteplaatz en datum en dan eventjes eh achter elkaar de kampen u aar u gewreest bent Oh dat is een helleboel. Jaaa ja [band loopt] [oke] [smuit] Uhmm. Fijn dat u dit bent. Uhm, zou u mij uw naam, geboorteplaatz en datum willen geven? Ik ben Lotty Huffener-	Amsterdam	https://data.niod.nl/WO2_Thesaurus/id/location/id/9577				
case_02	case_02_concept_02	gewreest bent. Dus ik vraag u naar uw naam, geboorteplaatz en datum en dan eventjes eh achter elkaar de kampen u aar u gewreest bent Oh dat is een helleboel. Jaaa ja [band loopt] [oke] [smuit] Uhmm. Fijn dat u dit bent. Uhm, zou u mij uw naam, geboorteplaatz en datum willen geven? Ik ben Lotty Huffener-	Auschvitz	https://data.niod.nl/WO2_Thesaurus/id/kampen/id/653				
case_02	case_02_concept_03	gewreest bent. Dus ik vraag u naar uw naam, geboorteplaatz en datum en dan eventjes eh achter elkaar de kampen u aar u gewreest bent Oh dat is een helleboel. Jaaa ja [band loopt] [oke] [smuit] Uhmm. Fijn dat u dit bent. Uhm, zou u mij uw naam, geboorteplaatz en datum willen geven? Ik ben Lotty Huffener-	Birkenau	https://data.niod.nl/WO2_Thesaurus/id/location/id/1393				
case_02	case_02_concept_04	gewreest bent. Dus ik vraag u naar uw naam, geboorteplaatz en datum en dan eventjes eh achter elkaar de kampen u aar u gewreest bent Oh dat is een helleboel. Jaaa ja [band loopt] [oke] [smuit] Uhmm. Fijn dat u dit bent. Uhm, zou u mij uw naam, geboorteplaatz en datum willen geven? Ik ben Lotty Huffener-	Gevangenen	https://data.niod.nl/WO2_Thesaurus/id/1928				
case_02	case_02_concept_05	gewreest bent. Dus ik vraag u naar uw naam, geboorteplaatz en datum en dan eventjes eh achter elkaar de kampen u aar u gewreest bent Oh dat is een helleboel. Jaaa ja [band loopt] [oke] [smuit] Uhmm. Fijn dat u dit bent. Uhm, zou u mij uw naam, geboorteplaatz en datum willen geven? Ik ben Lotty Huffener-	Kamp Vught	https://data.niod.nl/WO2_Thesaurus/id/kampen/id/684				
case_02	case_02_concept_06	gewreest bent. Dus ik vraag u naar uw naam, geboorteplaatz en datum en dan eventjes eh achter elkaar de kampen u aar u gewreest bent Oh dat is een helleboel. Jaaa ja [band loopt] [oke] [smuit] Uhmm. Fijn dat u dit bent. Uhm, zou u mij uw naam, geboorteplaatz en datum willen geven? Ik ben Lotty Huffener-	Reichenbach	https://data.niod.nl/WO2_Thesaurus/id/kampen/id/398				
case_02	case_02_concept_07	gewreest bent. Dus ik vraag u naar uw naam, geboorteplaatz en datum en dan eventjes eh achter elkaar de kampen u aar u gewreest bent Oh dat is een helleboel. Jaaa ja [band loopt] [oke] [smuit] Uhmm. Fijn dat u dit bent. Uhm, zou u mij uw naam, geboorteplaatz en datum willen geven? Ik ben Lotty Huffener-	Transporten	https://data.niod.nl/WO2_Thesaurus/id/1981				
case_02	case_02_concept_08	gewreest bent. Dus ik vraag u naar uw naam, geboorteplaatz en datum en dan eventjes eh achter elkaar de kampen u aar u gewreest bent Oh dat is een helleboel. Jaaa ja [band loopt] [oke] [smuit] Uhmm. Fijn dat u dit bent. Uhm, zou u mij uw naam, geboorteplaatz en datum willen geven? Ik ben Lotty Huffener-	Vught	https://data.niod.nl/WO2_Thesaurus/id/location/id/10280				
case_03	case_03_concept_01	Kvan u beiden uit een traditioneel Joods gezin? Nee, helemaal niet. Mijn man wel, maar ik uit een volkomen liberaal gezin, dat wa ook de grote moeilijkheden eh moeilijkheid in ons eh ja besluit om te gaan trouwen. [human] Minder dan van mijn kant dan van de kant van mijn man. Waarom? Nou, zij waren zo orthodox dat	Amsterdam	https://data.niod.nl/WO2_Thesaurus/id/location/id/9577				
case_03	case_03_concept_02	Kvan u beiden uit een traditioneel Joods gezin? Nee, helemaal niet. Mijn man wel, maar ik uit een volkomen liberaal gezin, dat wa ook de grote moeilijkheden eh moeilijkheid in ons eh ja besluit om te gaan trouwen. [human] Minder dan van mijn kant dan van de kant van mijn man. Waarom? Nou, zij waren zo orthodox dat	Duitsland	https://data.niod.nl/WO2_Thesaurus/id/location/id/8020				

Expert Evaluation | Instructions | Case Overview

Fig. 2. Expert Evaluation sheet of the blind expert questionnaire. Each row represents one proposed WO2 Thesaurus concept for one selected case. Experts saw the segment excerpt, concept label, and concept URI, and filled in correctness, specificity, missing-concept comments, and general comments. Method provenance was hidden from this sheet.



	A
1	**Expert evaluation questionnaire**
2	
3	**Purpose**
4	Please review each proposed WO2 thesaurus concept for the short interview excerpt. The concepts are shown without revealing which method selected them.
5	
6	**How to answer**
7	For each concept row, choose one value for correctness and one value for specificity using the dropdowns.
8	The missing concept comment is optional. Please fill it once per case if an important concept is missing.
9	The general comment is optional and can be used for uncertainty, label ambiguity, or other remarks.
10	
11	**Correctness options**
12	incorrect = the concept is not supported by the excerpt
13	partially correct = the concept is related, but not fully accurate or only weakly supported
14	correct = the concept is clearly supported by the excerpt
15	unsure = cannot judge from the excerpt
16	
17	**Specificity options**
18	too broad = the concept is correct but too general
19	appropriate = the concept is at a suitable level of detail
20	too narrow = the concept is too specific for what the excerpt supports
21	not applicable = use when the concept is incorrect
22	unsure = cannot judge from the excerpt
23	
24	**Source note**
25	This workbook was generated from the expert-facing CSV table with 44 concept rows. Method provenance is intentionally hidden in this expert-facing version.
26	
27	

Fig. 3. Instructions sheet of the expert evaluation questionnaire. This sheet explains the purpose of the evaluation, how experts should complete the form, and the allowed rating values for correctness and specificity.

	A	B	C
1	case_id	number_of_concept	segment_excerpt
2	case_01	3	Maar al die tijd zat jij te wachten op die fabriek? Wij zaten op die fabriek te wachten. En wat deed je dan de hele dag? Wij mochten niks doen, want onze handen moesten gespaard blijven voor dat diamantvak. En toen moesten mensen werken en toen sprong Philips daarin. Die zette daar die lopende banden neer en later mochten wij ook bij Philips werken. Dat was natuurlijk aan een kant wel prettig, want je had wat te doen de hele dag. Ja. [schraapt keel] Philips heeft gezorgd dat we daar aan het werk waren. Maar vóór die tijd deed je gewoon de hele dag niets? Dan deden we niets en dat was heel erg. Ja. Ja. Dat was heel naar. We hingen maar zo'n beetje rond. Ze moesten zorgen dat er met ons niks gebeurde.
3	case_02	8	geweest bent. Dus ik vraag u naar uw naam, geboorteplaats en datum en dan eventjes eh achter elkaar de kampen waar u geweest bent Oh dat is een heleboel. Jaaa ja [band loopt] [oke] [snuift] Uhmm. Fijn dat u d'r bent. Uhm, zou u mij uw naam, geboorteplaats en datum willen geven? Ik ben Lotty Huffener-Veffer. Ik ben geboren 10-7-'21 in Amsterdam. Dat eh Ja, en eh waar bent u uh uh in welke kampen bent u gedeporteerd geweest. Ik ben eerst in Vught geweest, in kamp Vught. Toen zijn we naar Auschwitz gegaan. Daarna naar Reichenbach uhhh zijn we geweest. Birkenau. We hebben daar gewerkt [huhum] en in januari '44 eh '45 zijn we weggegaan daar in Reichenbach. Oke, oke.
4	case_03	11	Kwam u beiden uit een traditioneel Joods gezin? Nee, helemaal niet. Mijn man wel, maar ik uit een volkomen liberaal gezin, dat was ook de grote moeilijkheden eh moeilijheid in ons eh ja besluit om te gaan trouwen. [humum] Minder dan van mijn kant dan van de kant van mijn man. Waarom? Nou, zij waren zo orthodox dat 't een ja bijna onmogelijk leek dat een meisje uit een liberaal huis hun schoondochter zou worden. Waren grote moeilijkheden. [humum] Ja. Heeft dat de band tussen u en uw schoonfamilie eh onder druk gezet? Ja. Tsja [lacht] Zou je wel kunnen zeggen, want de tragiek was dat dan in na de Kristallnacht ook mijn schoonouders wegmoesten en de enige mogelijkheid was dat ze bij ons kwamen. Is dat ook gebeurd? Oh ja. Dat zou je wel kunnen zeggen. Ze waren bij ons, dus we hadden eerst voor hun een woninkje gehuurd en in '40 was die zaak afgelopen, kon mijn man geen twee woningen onderhouden zonder te verdienen en toen zijn ze bij ons in huis gekomen met bovenop nog een schoonzuster, een ongetrouwde schoonzuster. Die was eerst aangevraagd door het meisjesweeshuis, het Joodse [kucht] en die hebben haar ontslagen. En toen kwam ze ook nog bij ons. Dus. Notabene de de uw schoonfamilie die d'r eh behoorlijk moeite mee had, moest u zelf gastvrij herbergen. Oh ja. ja. Maar goed, de ironie van de zaak was die m'n schoonvader was eh in Frankfurt procuratiehouder van een grote bank en aangezien Holland eigenlijk nou tot Duitsland hoorde na de bezetting, kreeg mijn man man mijn schoonvader plotseling salaris nabetaald. En toen na een jaar bij ons geweest te zijn hebben zij nog een kleine woning genomen. En zijn van daaruit gedeporteerd. Alsnog zijn ze gedeporteerd. Mijn schoonzuster is met het eerste transport op 15 juli '42 al gedeporteerd. En m'n schoonouders zijn weggehaald, toen waren wij al weggehaald. hum, hum. Uw uw schoonouders zijn dus als ik het goed heb in '39 naar eh Amsterdam gekomen. Ik geloof, ik kan het niet precies zeggen, ik geloof dat ze al in '38 in direct na de Kristallnacht [humum] kwamen. Al hun kinderen verder mijn man kwam uit een gezin van zes [ja] en daarvan waren eh [kucht] was een zuster de oudste na eerst via Joegoslavië naar Palestina, destijds Palestina [humum]. Eh, z'n enige broer en zijn zuster met haar man plus een jongste zuster zijn naar via Engeland naar Amerika en dus ze waren de enige. ja ja. Van uw eh Praat ik te vlug? Nee, is uitstekend. (...)
	case_04	13	Je bent ook nog niet in Sobibor geweest? Nee. Dat hoeft ook niet van mij hoor. Van mij wel. Maar ik ben wel in Auschwitz geweest en dat is ook nog maar kort geleden. Dat Jiddische koor, dat kreeg het verzoek van een Amerikaanse Jood die hier woont om in Auschwitz dingen te zingen. Hij had ze gehoord en hij zei 'dat is wat ik vroeger hoorde. Ik wil'-hij was ziek- 'ik wil nog naar Auschwitz en ik wil zo graag dat jullie daar zingen. toen heeft dat koor gedacht, daar moet begeleiding bij en die hadden aan mij gedacht en aan Shifra. Die zat toen nog niet in dat koor en ik was nog niet in Auschwitz geweest en Shifra zei, ja dat kan ik wel. Nou, dat is goed, dat zal ik doen als een soort therapeutische begeleiding, mochten er dingen in de groep misgaan. ik had daar grote moeite mee. Ze vertelde wel, hè, dan moest ze overleggen. Ik wou haar niet tegenhouden, maar ik kon het ook niet aan en ik ben op een gegeven moment de kroeg ingedoken, in de Swammerdamstraat. Daar is nu een kroeg, koffie. Daar heb ik zitten schrijven voor mezelf, al mijn woede en frustratie hierover. Over dat ze dat

Expert Evaluation | Instructions | Case Overview

Fig. 4. Case Overview sheet of the expert evaluation questionnaire. This sheet gives experts an overview of the five selected cases, including the case ID, number of proposed concepts, and the segment excerpt used for evaluation.

C Additional result tables

Table 11: Full case-level expert evaluation summary.

Case	Role	Condition	Strict correct rate	Applicable specificity rate
Case 01	Organisation/work conflict	Baseline	73.33%	50.00%

Table 11: Full case-level expert evaluation summary.

Case	**Role**	**Condition**	**Strict correct rate**	**Applicable specificity rate**
Case 01	Organisation/work conflict	Refined	73.33%	50.00%
Case 01	Organisation/work conflict	KG-RAG	80.00%	60.00%
Case 02	Missing camp concepts	Baseline	76.00%	60.87%
Case 02	Missing camp concepts	Refined	100.00%	100.00%
Case 02	Missing camp concepts	KG-RAG	96.00%	100.00%
Case 03	Missing event concepts	Baseline	52.00%	45.45%
Case 03	Missing event concepts	Refined	52.00%	45.45%
Case 03	Missing event concepts	KG-RAG	42.50%	48.28%
Case 04	Accepted control	Baseline	72.00%	78.95%
Case 04	Accepted control	Refined	75.00%	100.00%
Case 04	Accepted control	KG-RAG	66.00%	85.29%
Case 05	Place/camp disambiguation	Baseline	64.00%	94.44%
Case 05	Place/camp disambiguation	Refined	64.00%	94.44%
Case 05	Place/camp disambiguation	KG-RAG	67.50%	90.00%

D Reproducibility notes

The thesis repository is available at:

`https://github.com/ferufera/wo2-segmentation-optimization.git`

The relevant thesis folder is:

`03_thesis_kg_rag/`

The main previous WO2 oral history matching pipeline is available at:
`https://github.com/Oorlogsbronnen/wo2-oral-history-matching-pipeline`
The raw subtitle files are available at:
`https://github.com/Oorlogsbronnen/segmenten_ondertiteling`
The WO2 Thesaurus dataset is available at:
`https://data.spinque.com/ld/data/oorlogsbronnen/wo2_thesaurus/`
The WO2 Oorlogsbronnen SPARQL endpoint is available at:
`https://query.ldmax.nl/wo2-oorlogsbronnen/nfeAGR`
The main thesis scripts are listed below.

- `03_thesis_kg_rag/scripts/03_select_evaluation_segments.py`: selects the five evaluation cases from the larger set of enriched segments and validation groups.
- `03_thesis_kg_rag/scripts/retrieve_andidates.py`: defines the KG-RAG candidate retrieval logic, including original seed concepts, explicit lexical retrieval, contextual event retrieval, and conservative event graph expansion.
- `03_thesis_kg_rag/scripts/04_run_candidate_retrieval.py`: runs KG-RAG candidate retrieval on the five selected evaluation segments and saves the retrieved candidate lists.
- `03_thesis_kg_rag/scripts/05_build_kg_rag_prompts.py`: builds KG-RAG prompts from the selected segment excerpts and retrieved WO2 Thesaurus candidates.
- `03_thesis_kg_rag/scripts/08_build_refined_prompts.py`: builds refined prompts using the original WO2Net candidate concepts and stricter prompt logic.
- `03_thesis_kg_rag/scripts/10_build_baseline_prompts.py`: builds baseline prompts using the original WO2Net candidate concepts and archived baseline prompt logic.
- `03_thesis_kg_rag/scripts/14_process_expert_evaluation.py`: processes expert questionnaire responses and generates the summary tables used in the Results section.

The processed expert evaluation outputs are stored under:
`03_thesis_kg_rag/results/expert_evaluation/responses_processed/`
The core result files are:

- `expert_rating_summary_by_method.csv`
- `expert_rating_summary_by_case.csv`
- `expert_rating_summary_by_concept.csv`
- `expert_comment_themes.csv`
- `combined_expert_ratings.csv`

E Glossary of key terms

Table 12 defines the key terms used in the thesis. These definitions are included because the thesis moves between several levels of analysis: segments, candidate

concepts, selected concepts, expert questionnaire rows, and aggregate condition summaries.

Table 12: Key terms used in the thesis.

Term	**Definition**
Segment	A short excerpt from a longer oral history interview transcript. The five expert-facing excerpts used in this thesis are between 121 and 177 words.
Concept	A controlled-vocabulary term from the WO2 Thesaurus, identified by a stable URI and usually shown with a human-readable label.
Concept linking	The task of matching one segment to one or more relevant WO2 Thesaurus concepts.
Candidate concept	A concept made available to the LLM for a segment before selection.
Original candidate concept	A candidate concept produced by the previous WO2 oral history matching pipeline. Baseline and refined prompting use these concepts.
Retrieved candidate concept	A candidate concept retrieved from the WO2 Thesaurus in the KG-RAG condition.
Selected concept	A candidate concept that the LLM output keeps as relevant for the segment.
Condition	One of the three concept-linking alternatives compared in this thesis: baseline prompting, refined prompting, and KG-RAG prompting.
Case	One selected evaluation segment together with its role in the evaluation, such as missing camp concepts, missing event concept, or place/camp disambiguation.
Validation group	Previous crowdsource validation information associated with a segment and its proposed concept links.
Previous validation status	The status derived from earlier validation signals. In this thesis, accepted, rejected, and conflict indicate whether previous validation mostly accepted the output, rejected it as problematic, or showed mixed judgements.

Table 12: Key terms used in the thesis.

Term	**Definition**
Concept row	One row in the blind expert questionnaire. Each row contains a case ID, segment excerpt, proposed concept label, concept URI, and empty fields for expert ratings and comments.
Expert rating	One judgement by one expert for one concept row. Experts rated correctness and specificity.
Applicable specificity rating	A specificity judgement where the expert could meaningfully judge whether the concept was too broad, appropriate, or too narrow. Ratings marked not applicable, unsure, missing, or blank are excluded from this metric.