# Evaluation of Large Language Models on hierarchical entity matching for cultural heritage metadata

Bram Bakker[1], and Iris Hendrickx[2]

[1] Institute for Computing and Information Sciences, Radboud Universiteit, Nijmegen, The Netherlands
[2] Centre for Language Studies, Centre for Language and Speech Technology, Radboud University, Nijmegen, The Netherlands

**Abstract**

Data organization is a key challenge for many cultural heritage institutions. The size of current digital collections and the heterogeneity of metadata creates problems for interoperability, reliability and information retrieval. This work-in-progress paper aims to provide the groundwork for research into effectively and transparently performing automatic hierarchical entity matching within library metadata. We explore to what extent LLMs can detect whether an entity relationship for a pair of records (book descriptions) exists, based on their bibliographic metadata. We focus solely on the edge cases: those pairs of records that are potentially difficult to label with the correct entity relationship. We compare several different LLMs; we study the effect of the amount of detail in the metadata description and we also look at zero-shot versus few-shot prompting. Our results show that LLMs are a very promising technique for parts of this task of hierarchical entity matching of metadata book descriptions.

**Keywords:** digital humanities, entity matching, entity resolution, LLMs, natural language processing, library metadata

## 1 Introduction

Rapid digitization has allowed cultural heritage institutions to collect, process and distribute a staggering amount of cultural information. The KB national library of the Netherlands (KB) houses the national bibliography, which contains digital metadata of millions of books and newspapers.[1] However, the massive size of datasets like this can make *information retrieval* challenging. Libraries face the task of providing reliable access to valuable information for a variety of users like librarians, publishers, citizens and researchers.

One obstacle in this regard is a lack of digital bibliographic structure. Experts bemoan the fact that digital catalogs often lack the ”context in which each new translation, edition, and adaptation of a work inevitably takes place” ([12], p.4), which was such an integral aspect of analog library catalogs. When searching through the digital catalog of the KB, librarians would perhaps want to gather a list of all prints of a particular work like ”Max Havelaar”, excluding parodies or scholarly discussions about the work. Researchers may want to select and compare specific revised editions or particular translations. This would require a data model which categorizes individual books into larger umbrella entities. In the late 1990’s, the IFLA (International Federation of Library Association and Institutions) developed the FRBR framework with this goal in mind [4; 8].

Within the FRBR framework, a hierarchical entity-relationship model is defined called ”WEMI” (Work, Expression, Manifestation and Item) [16]. In this model, the highest level entity

Bram Bakker, and Iris Hendrickx. “Evaluation of Large Language Models on hierarchical entity matching for cultural heritage metadata.” In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 1146–1163. https://doi.org/10.63744/UKsLY7DKNPvA.



[1] `https://webggc.oclc.org/cbs/DB=3.9/`

is that of a ‘work’, an abstract cultural creation. An example could be the famous work ”Max Havelaar”, which captures the story, characters and general intellectual content related to the entity. A work is given form in a particular ‘expression’, this could be the Portuguese translation, a version with modernized Dutch language or even a comic book. However, if an adaptation makes significant changes to the original creation, this then often becomes its own new work.

An expression is made tangible through a ‘manifestation’, which can be characterized by lettering, cover-art, format and other physical attributes. These manifestations in turn can consist of various real-world ‘items’, for example three physical copies of a particular print of Max Havelaar with the same physical characteristics. The FRBR framework can be realized through metadata standards such as RDA (Resource Description Access) [6]. Using RDA, libraries can organize their often heterogeneous records into hierarchically connected entities, a desirable trait for linked open data.

However, many public bibliographies and library catalogs lack such a structure. Researchers report several important factors that make practical implementations of the model challenging [16], like the lack of a standardized definition of the entities, and the speed of digitization, which makes the amount of data that needs to be compared for connecting individual records into a hierarchical database of entities daunting [10].

One potential exception is the popular union catalog Worldcat, which does organize records into ”works” based on authoritative entity clusters using the OCLC work-set algorithm.[2] However, in this data model, work and expression entities are combined into a single entity. Additionally, multiple other algorithms are required to handle multilingual data and duplicate detection [9]. The KB library itself has also worked on such a rule-based clustering algorithm with moderate success.[3] As rules cannot capture every type of variation in the metadata, a more flexible solution is needed.

In recent years, it has been shown that Large Language models have high zero-shot performance on entity matching benchmarks [17]. For this reason, we aim to evaluate the zero-shot performance of LLMs on the task of automatic hierarchical entity matching within library metadata.

We focus our study on the edge cases in a sample from the Dutch national library catalog data: the non-trivial and challenging links between records that are difficult to generate automatically. Because the data is mostly in Dutch, we further want to investigate the performance of a Dutch model, Fietje, compared with the larger hosted commercial GPT and the open-source Llama models. We also investigate the use of demonstrative example matches in the prompt (few-shot prompting), and the effect of including or leaving out chunks of metadata. In the end, we discuss the performance, weaknesses and strengths of using LLMs on entity matching in this particular domain.

## 2 Related Work

The task of entity matching has been well established, and goes back more than fifty years [7]. It addresses the problem of heterogeneous data records referring to the same real world entity [14]. In our case we additionally define two *types* of entities, ‘works’ and ‘expressions’. These describe two levels of alignment, where ‘expression’ is more strict than ‘work’.

Early automatic entity matching systems were rule-based [3]. These rules consist mostly of matching specific attribute/metadata fields based on domain knowledge. The introduction of transformer models has severely changed the landscape of entity matching [2]. By feeding entire records in natural language as inputs, BERT-based models have achieved massive gains in performance on multiple benchmarks [13]. The downside of these methods is their lack of explainability [1].

[2] `https://www.oclc.org/research/areas/data-science/workrecs.html`
[3] `https://lab.kb.nl/tool/rda-entity-finder`

Rule-based models are inherently transparent and explainable. When a mistake is made, a simple investigation of the rules can reveal the reason why. This is far more difficult for a model like BERT with millions of parameters.

Most recently, Large Language Models or LLMs have achieved significant zero-shot performance in this task [11; 20]. This was first tested by Narayan et al., in 2022 [15]. In various experiments, LLMs like GPT, Mistral and Llama outperform SOTA pre-trained deep learning models. Pre-trained BERT-based models tend to struggle most with data outside of the domain of their training set, but even on data within their training set domain LLMs often tend to perform on-par with the pre-trained models [17].

## 3 Methodology

We investigate whether LLMs can predict entity relationships between pairs of books, using only their bibliographic metadata descriptions. We evaluate multiple LLMs, examining how the level of detail in the metadata affects performance. Additionally, we compare zero-shot and few-shot prompting strategies. In this section, a description is provided of the datasets, entity descriptions, general experimental setup, evaluation metrics and the models used.

In the experiments we select and prompt LLMs with candidate pairs of records as visualized in figure 1. These candidate pairs are each associated with a positive or negative value which denotes whether they match or not based on their WEMI umbrella entity. These triples are thus formulated in the following way: <record A> <record B> <pos/neg>.

### 3.1 Data

The data used to construct these triples is derived from a subset of the national bibliography, specifically focusing on metadata for children's books and novels. This metadata, already provided by the KB in RDA format, includes manually verified WEMI relationships that can serve as our evaluation set.

The selection of metadata in the prompt matters. For example, Peeters et al. show that removing price metadata in strings for the Walmart-Amazon benchmark dataset improves performance of the matching model [18]. Our metadata is in the format Pica+, which defines a large variety of fields and subfields to describe each book, the national bibliography uses 600 fields and subfields in total. These range from barely informative (physical characteristics like page number), to highly informative (title and author) and variably informative (the free annotation field). We discussed with the catalog experts of the KB which of the fields would be relevant. We created two versions of our record representations, a 'Long' version with 22 fields, and a 'Short' version with only the highly informative metadata (16 fields for expressions and 11 for the works). These fields used are listed in appendix B.

### 3.2 Large Language Models

We used the GPT API and proprietary models to test out our various prompts and conditions. The following models were tested:

1. **GPT 4.1 (2025-04-14)**: This LLM was at the time of testing the flagship GPT model, released in April 2025.[4]

2. **GPT 3.5 turbo**: This LLM is a smaller model than 4.1, but still performs well at various benchmarks.[5]

[4] `https://platform.openai.com/docs/models/gpt-4.1`
[5] `https://platform.openai.com/docs/models/gpt-3.5-turbo`

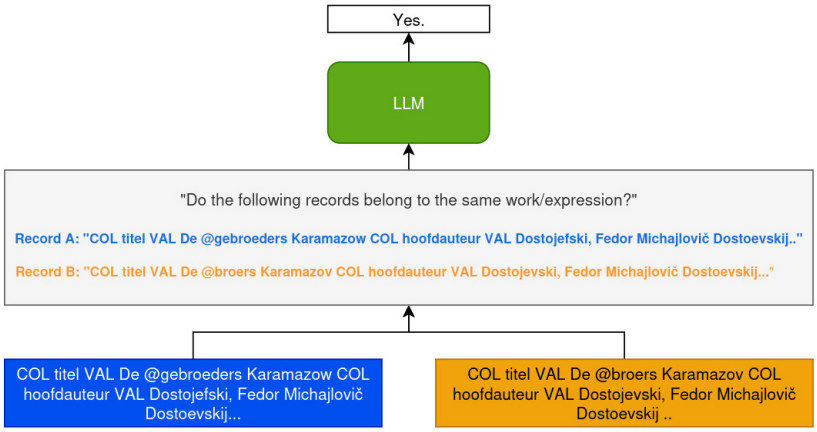


**Figure 1:** Visual example of LLM prompt procedure

3. **Llama 3.3 70B Instruct**: This open-source model from Meta offers a balance between size and performance.[6]

4. **Fietje 2**: This small-scale LLM with just 2 billion parameters is developed by Bram van Roy and is an adaptation of Phi to improve performance on the Dutch language specifically.[7]

### 3.3 Test set creation

The test-set for work pairs consists solely of edge-cases, as the majority of record pairs are trivial to match as positive when they share the exact same title and author, or as a negative when they do not. To find the edge-cases we manually investigated the WEMI connections from the RDA data provided by the KB.

Examples of edge-cases are two records from the same work without an exact match for author and title (example 1). We also selected for records from the same author which belong to different works but with high string similarity between titles (example 2). Furthermore, we selected translations, books from the same work but have different title strings (example 3). There were no strict guidelines when searching for these edge-cases. We also keep another 'random' test-set of candidate work-pairs which were not handpicked to compare performance on the more simple cases.

(1) COL taal VAL ned COL jaar VAL 1930 COL titel VAL @Gadzji Moerat COL hoofdauteur vermelding VAL door L.N. Tolstoy COL 2e auteur vermelding VAL vertaling [uit het Russisch] van dr Anna Kosloff COL hoofdauteur VAL Tolstoy, L.N. Tolstoj 1828-1910 COL 2e auteur VAL Kosloff, Anna Kosloff COL plaats VAL Amsterdam COL uitgever VAL De Spieghel COL vertaling van VAL Chadzi-Murat. - 1912

COL taal VAL ned COL jaar VAL 2015 COL titel VAL @Hadzji Moerat COL hoofdauteur vermelding VAL Leo Tolstoj COL 2e auteur vermelding VAL vertaald [uit het Russisch] door Froukje Slofstra COL hoofdauteur VAL Tolstoj, L.N. Tolstoj 1828-1910 COL 2e auteur VAL Slofstra, Froukje Slofstra 1977- COL editie VAL Eerste druk COL plaats VAL Amsterdam COL uitgever VAL Uitgeverij Van Oorschot COL vertaling van VAL Hadzji Moerat

(2) COL taal VAL ned COL titel VAL De @atoomtrillingen COL hoofdauteur VAL Toonder, Marten Toonder 1912-2005

COL taal VAL ned COL titel VAL De @toornviolen COL hoofdauteur VAL Toonder, Marten Toonder 1912-2005 COL Eerder verschenen in VAL Een heer moet alles alleen doen. - Amsterdam : De Bezige Bij, 1969. - (Literaire reuzenpocket ; 310) COL auteur/primair VAL Toonder

[6] `https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct`
[7] `https://huggingface.co/BramVanroy/fietje-2-instruct`

(3) COL taal VAL dee COL jaar VAL 1952 COL titel VAL @Tom Puss COL hoofdauteur vermelding VAL af Marten Toonder COL hoofdauteur VAL Toonder, Marten Toonder 1912-2005 COL plaats VAL København COL uitgever VAL A/S Ota COL omschrijving VAL Plaatjesalbum als reclame voor Ota gryn

COL taal VAL ned COL jaar VAL 1952 COL titel VAL @Tom Poes COL hoofdauteur vermelding VAL door Marten Toonder COL hoofdauteur VAL Toonder, Marten Toonder 1912-2005 COL plaats VAL Rotterdam COL uitgever VAL Quaker Oats graanproducten COL omschrijving VAL Datering: volgens Strip Katalogus Matla COL omschrijving VAL Met ingeplakte plaatjes

As a third test set, we select candidate expression-pairs from records belonging to the same work. This allows us to evaluate the models' ability to differentiate between revised editions and distinct manifestations of the same expression. Differentiating between pairs of expressions presents a greater challenge than distinguishing between work-pairs, due to limitations in the available metadata, which can lack sufficient detail to determine the correct label, like in example 4. Both of the records are Dutch translations of Oliver Twist with illustrations, but in the first record the translator is not included in the description, so we do not know if both of the manifestations make use of the same translation. This means the textual content could differ based on information not present in the metadata record. Further examples of the dataset can be found in appendix D.

(4) COL taal VAL ned COL jaar VAL 196X COL annotatie VAL Beeldverhaal COL titel VAL @Oliver Twist COL hoofdauteur vermelding VAL Charles Dickens COL 2e auteur vermelding VAL [vert. uit het Engels] COL hoofdauteur VAL Dickens, Charles John Huffam Dickens 1812-1870 COL plaats VAL [S.l.] COL uitgever VAL Rotogravure Pers COL omschrijving VAL Datering: niet in Brinkman, spelling Marchant, schatting

COL taal VAL ned COL jaar VAL 1941 COL titel VAL @Olivier Twist COL hoofdauteur vermelding VAL door Charles Dickens COL 2e auteur vermelding VAL nieuwe vert. [uit het Engelsch] COL 2e auteur vermelding VAL met 28 houtgr. naar teek. van J. Mahony COL hoofdauteur VAL Dickens, Charles John Huffam Dickens 1812-1870 COL 2e auteur VAL Mahony, J. Mahony COL editie VAL 12e dr COL plaats VAL Amsterdam [etc.] COL uitgever VAL Gebr. Graauw Uitg.-Mij. COL vertaling van VAL Oliver Twist. - London : Bentley, 1838 COL Oorspr. titel VAL Olivier Twist, of het leven van een weesjongen. - Amsterdam : Gebr. Diederichs, 1840

The work test-sets are imbalanced to reflect a real world scenario as there are fewer positive pairs for the same work than negative matches in a dataset. The number of pairs for each test-set is listed in table 1.

| dataset | Positives | Negatives | total pairs |
|---|---|---|---|
| expression-test-sample | 164 | 186 | 350 |
| work-test-sample | 86 | 264 | 350 |
| work-random-sample | 50 | 450 | 500 |

**Table 1:** Ratio of positive and negative pairs for each test-set

### 3.4 Prompting

Previous work on applying LLMs for entity matching has shown that few-shot prompting is more effective than zero-shot prompting [15]. We select 4 examples for both works-pairs and expression-pairs. In both cases these 4 examples consist of 2 positive and 2 negative pairs.

The prompts are constructed by first instructing the model with its role: "You are a data assistant". Then a definition of either 'work' or 'expression' is provided. Unfortunately the FRBR framework does not provide definitions to support the needs of many different types of users [5]. We thus have to choose specific definitions that suite our case, and add into our prompt a definition for textual expressions from O'neill et al. [16]. For 'work' we use the IFLA definition by specifying that the term 'work' comprises distinct intellectual content. The exact prompts can be found in appendix A.

### 3.5 Experimental Setup

We perform experiments on three different test-sets, containing two different input types, long and short prompts, using two different strategies, few-shot and zero-shot. In this way, we test the models' sensitivity for metadata and demonstrative examples. We select precision, recall and F1 evaluation measures of positive pair classification. A link to the GitHub with the full code and data can be found here [8].

We also use two baselines for both expression and work matching. One baseline ('rule') relies on rule-based title and author-field matching, details can be found in appendix C. The other baseline ('encoder') matches embeddings generated from a model called e5-small [19] using the Skrub package.[9] The e5 model is trained using contrastive multilingual text pairs. The embeddings are matched based on a cosine similarity threshold of 0.5 for works, and 0.7 for expressions. Both of these methods are significantly faster and computationally more efficient compared to the LLMs, while being easy to implement.

## 4 Results

| Input type | Model | Precision | | Recall | | F1 score | |
|---|---|---|---|---|---|---|---|
| | | Zero-shot | Few-shot | Zero-shot | Few-shot | Zero-shot | Few-shot |
| Short | Fietje | 0.266 | 0.260 | 1.000 | 1.000 | 0.421 | 0.413 |
| | GPT 3.5 turbo | 0.535 | 0.633 | 0.988 | 0.802 | 0.694 | 0.708 |
| | GPT 4.1 | 0.891 | 0.929 | 0.954 | 0.907 | 0.921 | 0.918 |
| | Llama 3.3 70b | 0.667 | 0.766 | 0.941 | 0.847 | 0.781 | 0.805 |
| | baseline rule | 0.212 | - | 0.361 | - | 0.267 | - |
| | baseline encoder | 0.308 | - | 0.907 | - | 0.460 | - |
| Long | Fietje | 0.202 | 0.249 | 1.000 | 0.980 | 0.337 | 0.397 |
| | GPT 3.5 | 0.494 | 0.608 | 0.941 | 0.765 | 0.648 | 0.677 |
| | GPT 4.1 | 0.920 | 0.939 | 0.930 | 0.895 | 0.925 | 0.917 |
| | Llama 3.3 70b | 0.851 | 0.755 | 0.663 | 0.861 | 0.745 | 0.804 |
| | baseline rule | 0.212 | - | 0.361 | - | 0.267 | - |
| | baseline encoder | 0.302 | - | 0.698 | - | 0.421 | - |

**Table 2:** Evaluation Measures for 'work' prompts

In table 2 we can see that GPT 4.1 is the best performing model in the task of matching records for works, with a highest F1 score of 0.925, zero-shot with long metadata. Fietje performs worst and classifies nearly all pairs as positive. In general, we note that adding demonstrative examples in the prompt and the inclusion of more metadata seems to increase precision at the cost of recall. One interesting aspect is that only GPT 4.1 seems to be able to fully take advantage of the extra metadata, both increasing precision and recall. The baseline encoder performs better than the rule-based baseline, but both do not reach the performance of bigger LLMs.

[8] https://github.com/BramBakker/Test_LLM_WEMI_kb

[9] https://skrub-data.org/stable/

| Input type | Model | Precision | | Recall | | F1 score | |
|---|---|---|---|---|---|---|---|
| | | Zero-shot | Few-shot | Zero-shot | Few-shot | Zero-shot | Few-shot |
| Short | GPT 4.1 | 1.000 | 1.000 | 0.940 | 0.900 | 0.969 | 0.947 |
| | baseline rule | 1.000 | - | 0.880 | - | 0.936 | - |
| | baseline encoder | 0.800 | - | 0.960 | - | 0.873 | - |
| Long | GPT 4.1 | 1.000 | 1.000 | 0.920 | 0.920 | 0.958 | 0.958 |
| | baseline rule | 1.000 | - | 0.880 | - | 0.936 | - |
| | baseline encoder | 0.764 | - | 0.840 | - | 0.800 | - |

**Table 3:** Evaluation Measures for ‘work’ prompts: non difficult cases

Table 3 shows that not handpicking edge-cases results in comparable performance of the LLMs compared to the baselines. This calls for careful analysis as the token-related and environmental costs associated with using GPT models can be substantial.

| Input type | Model | Precision | | Recall | | F1 score | |
|---|---|---|---|---|---|---|---|
| | | Zero-shot | Few-shot | Zero-shot | Few-shot | Zero-shot | Few-shot |
| Short | Fietje | 0.468 | 0.494 | 1.000 | 0.781 | 0.638 | 0.605 |
| | GPT 3.5 turbo | 0.726 | 0.551 | 0.274 | 0.598 | 0.398 | 0.573 |
| | GPT 4.1 | 0.786 | 0.807 | 0.402 | 0.409 | 0.532 | 0.543 |
| | Llama 3.3 70b | 1.000 | 0.000 | 0.024 | 0.000 | 0.048 | 0.000 |
| | baseline rule | 0.620 | - | 0.677 | - | 0.647 | - |
| | baseline encoder | 0.601 | - | 0.799 | - | 0.686 | - |
| Long | Fietje | 0.469 | 0.523 | 0.976 | 0.700 | 0.634 | 0.600 |
| | GPT 3.5 | 0.718 | 0.571 | 0.171 | 0.439 | 0.276 | 0.497 |
| | GPT 4.1 | 0.864 | 0.869 | 0.232 | 0.323 | 0.365 | 0.471 |
| | Llama 3.3 70b | 1.000 | 1.000 | 0.006 | 0.006 | 0.012 | 0.012 |
| | baseline rule | 0.620 | - | 0.677 | - | 0.647 | - |
| | baseline encoder | 0.573 | - | 0.549 | - | 0.561 | - |

**Table 4:** Evaluation Measures for ‘expression’ prompts

Table 4 shows that performance on the expression matching of LLMs is generally worse than the rule-based baseline function. Llama only classifies pairs as negatives. Fietje classifies nearly every pair as a positive. Only the GPT models mix their predictions, but they are still not accurate. Even though the definition of the concept of ‘expression’ used in the prompt was precise, every model struggled with it. We will investigate this more deeply in the discussion section.

### 4.1 Error Analysis

GPT 4.1 has better F1 scores than other models in every experiment. A large part of this improvement can be attributed to excluding false matches with high string similarity. Especially book series with very similar titles by the same writers get often misclassified by other models as positives such as example 5.

(5) COL taal VAL ned COL taal VAL eng COL jaar VAL 2013 COL titel VAL @Nijntjes eerste woordenboek COL hoofdauteur vermelding VAL Dick Bruna COL 2e auteur vermelding VAL [vert. uit het Engels] COL hoofdauteur VAL Bruna, Hendrik Magdalenus Bruna 1927-2017 COL plaats VAL Amsterdam COL uitgever VAL Mercis Publishing COL omschrijving VAL Teksten in het Engels en Nederlands

COL taal VAL ned COL taal VAL eng COL jaar VAL 2013 COL titel VAL @Nijntjes eerste telboek COL hoofdauteur vermelding VAL Dick Bruna COL 2e auteur vermelding VAL [vert. uit het Engels] COL hoofdauteur VAL Bruna, Hendrik Magdalenus Bruna 1927-2017 COL plaats VAL Amsterdam COL uitgever VAL Mercis Publishing COL omschrijving VAL Teksten in het Nederlands en Engels

In order to understand why the models struggle with matching expressions, we investigated the mistakes they make. Even the large and advanced GPT 4.1 often does not match positive pairs when there are added illustrators or slight spelling variations in the title, as is the case in example 6. These are not integral textual changes of the content of the book, yet the models keep misclassifying these positive pairs.

(6) COL taal VAL ned COL jaar VAL 1958 COL titel VAL @Ierse nachten COL ondertitel VAL roman COL hoofdauteur vermelding VAL S. Vestdijk COL hoofdauteur VAL Vestdijk, Simon Vestdijk 1898-1971 COL editie VAL Vijfde druk COL plaats VAL 's-Gravenhage COL uitgever VAL N.V. Uitgeverij Nijgh & Van Ditmar

COL taal VAL ned COL jaar VAL 1946 COL titel VAL @Iersche nachten COL ondertitel VAL roman COL hoofdauteur vermelding VAL S. Vestdijk COL hoofdauteur VAL Vestdijk, Simon Vestdijk 1898-1971 COL plaats VAL Rotterdam COL uitgever VAL Nijgh & Van Ditmar

## 5 Discussion & Conclusion

This study explores the capacity of large language models (LLMs) to infer entity relationships between pairs of books, relying solely on their bibliographic metadata descriptions. Our results have shown that especially GPT 4.1 achieves remarkable performance with zero-shot prompting on this task. Furthermore, GPT 4.1 is able to incorporate knowledge from the large input metadata descriptions effectively.

For most of the models, including examples in the prompt often boosted precision at the cost of recall. From the perspective of a potential human-in-the-loop cataloger in the library checking these results, the less accurate model with higher recall is more desirable. It could be the case that our selection of edge-cases in the prompt is not the most appropriate for such application.

We had expected that a smaller Dutch model would give an advantage, but our results showed that Fietje underperformed. The smaller baseline encoder even outperformed it. This could be because the e5 model used for the baseline was trained using contrastive text pairs, fitting better with our objective.

For the work matching task, we observe that LLMs can get great performance, but for expression matching this was not the case. We found that even simple rule-based matching systems perform on similar levels as LLMs on the task of matching expressions.

The disappointing expression matching results can be explained by the fact that the LLMs lack the ability to discern expressions based on the rich and variable metadata fields, even when given a clear definition, as shown in the error analysis. Another reason could be that it is sometimes difficult to discern a ground truth from the metadata only, which we can see in example 4. The ability to identify integral changes to textual content might require access to that content. In the future, it would be interesting to automatically sample the content of digitized books for significant textual changes.

A potential limitation of these results is the fact that our test-set contains many books by famous authors. LLMs have a large pool of knowledge about these authors and their literature from outside of the dataset. We could investigate the performance of the model without this knowledge, and in the future include a separate test-set for lesser known authors.

As this paper reports on a preliminary study, many possible paths to explore further remain such as the above mentioned larger diversity in the test sample, but also adaption of the prompts, a more refined selection of metadata fields to be included in the record representation, and a stronger rule-based baseline algorithm.

Another open question is the performance of fine-tuned models. In the current study we only used off the shelf models, but perhaps a fine-tuned language model could perform well, with the added benefit that it can be made explainable using XAI techniques developed specifically for entity matching [1]. Therefore, to follow up this work-in-progress paper, we aim to compare the performance of a LLM fine-tuned on domain data from the KB to that of the general LLMs. To

enhance transparency and support catalogers in their decision-making processes, we also intend to include an explainable component in this implementation.

## 6 Acknowledgments

This work is part of the HAICu project with file number NWA.1518.22.105 which is financed by the Dutch Research Council (NWO). We would like to thank Djoke Dam and Michel de Gruijter, Eric Vos, Sita Bhagwandin and Meta van der Waal-Gentenaar from the KB National Library for their valuable input and sharing of the data.

## References

[1] Barlaug, Nils. “LEMON: explainable entity matching”. In: *IEEE Transactions on Knowledge and Data Engineering 35*, no. 8 (2022), pp. 8171–8184.

[2] Barlaug, Nils and Gulla, Jon Atle. “Neural networks for entity matching: A survey”. In: *ACM Transactions on Knowledge Discovery from Data (TKDD) 15*, no. 3 (2021), pp. 1–37.

[3] Binette, Olivier and Steorts, Rebecca C. “(Almost) all of entity resolution”. In: *Science Advances 8*, no. 12 (2022), eabi8021.

[4] Carlyle, Allyson. “Understanding FRBR as a conceptual model”. In: *Library resources & technical services 50*, no. 4 (2006), pp. 264–273.

[5] Coyle, Karen. “FRBR, twenty years on”. In: *Cataloging & Classification Quarterly 53*, no. 3-4 (2015), pp. 265–285.

[6] Coyle, Karen and Hillmann, Diane. “Resource description and access (RDA)”. In: *D-Lib magazine 13*, no. 1/2 (2007), pp. 1082–9873.

[7] Fellegi, Ivan P and Sunter, Alan B. “A theory for record linkage”. In: *Journal of the American statistical association 64*, no. 328 (1969), pp. 1183–1210.

[8] Functional Requirements for Bibliographic Records, IFLA Study Group on the, Library Associations, International Federation of, and Cataloguing. Standing Committee, Institutions. Section on. *Functional requirements for bibliographic records*. Vol. 19. De Gruyter Saur, 1998.

[9] Gatenby, Janifer, Greene, Richard O, Oskins, W Michael, and Thornburg, Gail. “GLIMIR: Manifestation and content clustering within WorldCat”. In: *Code4Lib journal* , no. 17 (2012).

[10] Godby, Carol Jean, Wang, Shenghui, and Mixter, Jeffrey K. *Library linked data in the cloud: OCLC’s experiments with new models of resource description*. Springer Nature, 2022. Chap. 1.

[11] Hussein, Mahmoud Mohamed Ashour. *LLM-Enhanced Entity Matching: Comparative Analysis of traditional and modern techniques*. MA thesis. ETH Zurich, 2025.

[12] Patrick Le Boeuf, edited by. *Functional requirements for bibliographic records (FRBR): hype or cure-all?* Routledge, 2005.

[13] Li, Yuliang, Li, Jinfeng, Suhara, Yoshihiko, Doan, AnHai, and Tan, Wang-Chiew. “Deep entity matching with pre-trained language models”. In: *Proceedings of the VLDB Endowment 14*, no. 1 (2020), pp. 50–60.

[14] Li, Yuliang, Li, Jinfeng, Suhara, Yoshihiko, Wang, Jin, Hirota, Wataru, and Tan, Wang-Chiew. “Deep entity matching: Challenges and opportunities”. In: *Journal of Data and Information Quality (JDIQ) 13*, no. 1 (2021), pp. 1–17.

[15] Narayan, Avanika, Chami, Ines, Orr, Laurel, and Ré, Christopher. “Can Foundation Models Wrangle Your Data?” In: *Proceedings of the VLDB Endowment 16*, no. 4 (2022), pp. 738–746. DOI: 10.14778/3574245.3574258.

[16] O’Neill, Edward and Žumer, Maja. “FRBR: application of the model to textual documents”. In: *Library Resources & Technical Services 62*, no. 4 (2018), p. 176.

[17] Peeters, Ralph and Bizer, Christian. “Using chatgpt for entity matching”. In: *European Conference on Advances in Databases and Information Systems*. Springer. 2023, pp. 221–230.

[18] Peeters, Ralph, Steiner, Aaron, and Bizer, Christian. “Entity matching using large language models”. In: *OpenProceedings 2* (2025), pp. 529–541.

[19] Wang, Liang, Yang, Nan, Huang, Xiaolong, Jiao, Binxing, Yang, Linjun, Jiang, Daxin, Majumder, Rangan, and Wei, Furu. “Text embeddings by weakly-supervised contrastive pretraining”. In: *arXiv preprint arXiv:2212.03533* (2022).

[20] Wang, Tianshu, Chen, Xiaoyang, Lin, Hongyu, Chen, Xuanang, Han, Xianpei, Sun, Le, Wang, Hao, and Zeng, Zhenyu. “Match, Compare, or Select? An Investigation of Large Language Models for Entity Matching”. In: *Proceedings of the 31st International Conference on Computational Linguistics*. 2025, pp. 96–109.

## A Prompts used

### Work prompt, original Dutch:

Je bent een data assistent.

Bepaal of twee boekenrecords tot hetzelfde werk behoren.

De term ”werk” omvat de intellectuele inhoud.

Record A:{record_a}

Record B: {record_b}

Behoren deze records tot hetzelfde werk?

Antwoord alleen Ja of Nee:

### English translation:

You are a data assistant.

Determine whether two book records belong to the same work.

The term ”work” refers to the intellectual content.

Record A:{record_a}

Record B: {record_b}

Do these records belong to the same work?

Answer only Yes or No:

### expression prompt, original Dutch:

Je bent een data assistent.

Bepaal of twee boekenrecords tot dezelfde expressie behoren.

De term ”expressie” omvat de specifieke inhoud in de vorm van woorden, zinnen en paragrafen.

Dit betreft alleen aspecten die integraal onderdeel zijn van de (tekstuele) artistieke realisatie.

Vertalingen, herziene edities of bewerkingen van hetzelfde werk worden als andere expressies beschouwd.

Record A:{record_a}
Record B: {record_b}

Behoren deze records tot dezelfde expressie?
Antwoord alleen Ja of Nee:

## English translation:

You are a data assistant.
Determine whether two book records belong to the same expression.
The term ”expression” refers to the specific content in the form of words, sentences, and paragraphs.
This only concerns aspects that are an integral part of the (textual) artistic realization.
Translations, revised editions, or adaptations of the same work are considered different expressions.

Record A: {record_a}
Record B: {record_b}

Do these records belong to the same expression?
Answer only Yes or No:

## Demonstrative examples used for works:

record A: COL taal VAL ned COL jaar VAL 1974 COL titel VAL Een @handvol rogge COL hoofdauteur vermelding VAL Agatha Christie COL 2e auteur vermelding VAL [geaut. vert. uit het Engels] COL hoofdauteur VAL Christie, Agatha Christie 1890-1976 COL editie VAL [Herdr.] COL plaats VAL Leiden COL uitgever VAL Sijthoff COL omschrijving VAL Balans COL vertaling van VAL A pocket full of rye. - 1953
record B: COL ISBN VAL 9021800500 COL taal VAL ned COL jaar VAL 1970 COL titel VAL Een @handvol rogge COL hoofdauteur vermelding VAL Agatha Christie COL 2e auteur vermelding VAL [geaut. vert. uit het Engels] COL hoofdauteur VAL Christie, Agatha Christie 1890-1976 COL editie VAL 2e dr COL plaats VAL Leiden COL uitgever VAL Sijthoff COL vertaling van VAL A pocket full of rye. - 1953

Behoren deze boeken tot hetzelfde werk? Antwoord Ja of Nee:
reactie: Ja

record A: COL taal VAL ned COL jaar VAL 1828 COL titel VAL @Eustachius, of De zegepraal van het christendom COL ondertitel VAL eene geschiedenis der vroegere christelijke eeuw COL hoofdauteur vermelding VAL door H.C. Schmid COL 2e auteur vermelding VAL [vert. uit het Duits] COL hoofdauteur VAL Schmid, Johann Christoph Friedrich von Schmid 1768-1854 COL plaats VAL Amsterdam COL uitgever VAL Ten Brink en De Vries COL omschrijving VAL Saakes 9 (1829), p. 16 COL vertaling van VAL Eustachius, eine Geschichte der christlichen Vorzeit. - 1828
record B: COL taal VAL fra COL jaar VAL 1843 COL titel VAL @Eustache COL ondertitel VAL episode des premiers temps du christianisme COL hoofdauteur vermelding VAL Christoph von Schmid COL 2e auteur vermelding VAL Louis Friedel [transl.] COL hoofdauteur VAL Schmid, Johann Christoph Friedrich von Schmid 1768-1854 COL 2e auteur VAL Friedel COL plaats VAL Tours COL uitgever VAL Mame

Behoren deze boeken tot hetzelfde werk? Antwoord Ja of Nee:
reactie: Ja

record A: COL ISBN VAL 9022905349 COL taal VAL ned COL jaar VAL 1976 COL titel VAL De @Saint aan het stuur COL hoofdauteur vermelding VAL Leslie Charteris COL 2e auteur vermelding VAL vertaling [uit het Frans]: Maarten Beks COL hoofdauteur VAL Charteris, Leslie Charteris 1907-1993 COL 2e auteur VAL Beks, Martinus Arnoldus Maria Beks 1929-2001 COL plaats VAL Utrecht COL uitgever VAL A.W. Bruna & Zoon COL vertaling van VAL Le Saint au volant. - Paris : Fayard, 1961

uitgever VAL Sijthoff
record B:COL taal VAL ned COL jaar VAL 1959 COL titel VAL De @Saint en de tijger COL hoofdauteur vermelding VAL Leslie Charteris COL 2e auteur vermelding VAL vertaling [uit het Engels]: Havank COL hoofdauteur VAL Charteris, Leslie Charteris 1907-1993 COL 2e auteur VAL Havank, Havank COL plaats VAL Utrecht COL uitgever VAL A.W. Bruna & Zoon COL vertaling van VAL Meet the tiger. - 1935

Behoren deze boeken tot hetzelfde werk? Antwoord Ja of Nee:
reactie: Nee

record A: COL taal VAL ned COL jaar VAL 2013 COL titel VAL De @vreugde van het leven COL hoofdauteur vermelding VAL Catherine Cookson COL 2e auteur vermelding VAL vertaling [uit het Engels]: Annet Mons COL hoofdauteur VAL Cookson, Catherine Anne Cookson 1906-1998 COL 2e auteur VAL Mons, Annet Mons COL plaats VAL Amsterdam COL uitgever VAL Boekerij COL vertaling van VAL My beloved son. - London : Bantam Press, 1991
record B: COL taal VAL ned COL jaar VAL 2013 COL titel VAL De @drempel van het leven COL hoofdauteur vermelding VAL Catherine Cookson COL 2e auteur vermelding VAL vertaling [uit het Engels]: Annet Mons COL hoofdauteur VAL Cookson, Catherine Anne Cookson 1906-1998 COL 2e auteur VAL Mons, Annet Mons COL plaats VAL Amsterdam COL uitgever VAL Boekerij COL vertaling van VAL The rag nymph. - New York : Bantam, ©1991

Behoren deze boeken tot hetzelfde werk? Antwoord Ja of Nee:
reactie: Nee

## Demonstrative examples used for expressions:

record A: COL taal VAL ned COL jaar VAL 1961 COL titel VAL @Van de boze koster! COL hoofdauteur vermelding VAL door W.G. van de Hulst COL 2e auteur vermelding VAL met tekeningen van W.G. van de Hulst Jr. COL hoofdauteur VAL Hulst, Willem Gerrit van de Hulst 1879-1963 COL 2e auteur VAL Hulst, Willem Gerrit van de Hulst 1917-2006 COL editie VAL 15e dr., 142e-153e duiz COL plaats VAL Nijkerk COL uitgever VAL G.F. Callenbach
record B: COL ISBN VAL 9026642520 COL taal VAL ned COL jaar VAL 1978 COL titel VAL @Van de boze koster COL hoofdauteur vermelding VAL W.G. van de Hulst COL 2e auteur vermelding VAL met zwarte en gekleurde tekn. van W.G. van de Hulst Jr COL hoofdauteur VAL Hulst, Willem Gerrit van de Hulst 1879-1963 COL 2e auteur VAL Hulst, Willem Gerrit van de Hulst 1917-2006 COL editie VAL 20e dr COL plaats VAL Nijkerk COL uitgever VAL Callenbach COL taal VAL ned COL jaar VAL 1900 COL titel VAL @Tekstuitgaaf van de varkenshoeder COL ondertitel VAL een sprookje met zang en piano-begeleiding COL plaats VAL Schiedam COL uitgever VAL Roelants

Behoren deze records tot dezelfde expressie?
Antwoord Ja of Nee:
reactie: Ja

record A: COL taal VAL fra COL jaar VAL 1920 COL titel VAL @Sans famille COL hoofdauteur vermelding VAL par Hector Malot COL hoofdauteur VAL Malot, Hector Henri Malot 1830-1907 COL plaats VAL Paris COL uitgever VAL Fayard
record B: COL taal VAL fra COL jaar VAL 1954 COL titel VAL @Sans famille COL hoofdauteur vermelding VAL Hector Malot COL 2e auteur vermelding VAL Marianne Clouzot [ill.] COL hoofdauteur VAL Malot, Hector Henri Malot 1830-1907 COL 2e auteur VAL Clouzot COL plaats VAL Paris COL uitgever VAL Hachette

Behoren deze records tot dezelfde expressie?
Antwoord Ja of Nee:
reactie: Ja

record A: COL ISBN VAL 3855397937 COL taal VAL ned COL jaar VAL 1987 COL titel VAL De @varkenshoeder COL ondertitel VAL een sprookje van Hans Christian Andersen COL hoofdauteur vermelding VAL vert. [uit het Duits naar de oorspr.

Deense uitg.] door Ineke Ris COL 2e auteur vermelding VAL met ill. van Dorothee Duntze COL hoofdauteur VAL Andersen, Hans Christian Andersen 1805-1875 COL 2e auteur VAL Ris, Ineke Ris COL 2e auteur VAL Duntze, Dorothee Duntze 1960- COL plaats VAL Den Haag COL uitgever VAL De Vier Windstreken COL vertaling van VAL Der Schweinehirt. - Mönchaltorf : Nord-Süd Verlag, cop. 1987 COL auteur/primair VAL Andersen
record B:COL ISBN VAL 3851951212 COL taal VAL dui COL jaar VAL 1982 COL titel VAL Der @Schweinehirt COL hoofdauteur vermelding VAL H.C. Andersen COL 2e auteur vermelding VAL [ill.] Lisbeth Zwerger COL hoofdauteur VAL Andersen, Hans Christian Andersen 1805-1875 COL 2e auteur VAL Zwerger, Lisbeth Zwerger 1954- COL plaats VAL Salzburg COL uitgever VAL Verlag Neugebauer Press

Behoren deze records tot dezelfde expressie?
Antwoord Ja of Nee:
reactie: Nee

record A: COL taal VAL ned COL jaar VAL 2014 COL titel VAL @Omdat ik zoveel van je hou COL hoofdauteur vermelding VAL Guido van Genechten COL hoofdauteur VAL Van Genechten, Guido Van Genechten 1957- COL editie VAL Zesde herziene druk COL plaats VAL [Amsterdam] COL uitgever VAL Clavis
record B: COL taal VAL ned COL jaar VAL 2016 COL titel VAL @Omdat ik zoveel van je hou COL hoofdauteur vermelding VAL Guido van Genechten COL hoofdauteur VAL van Genechten, Guido Van Genechten 1957- COL editie VAL Eerste druk editie klein formaat COL plaats VAL [Amsterdam] COL uitgever VAL Clavis

Behoren deze records tot dezelfde expressie? Antwoord Ja of Nee:
reactie: Nee

## B Metadata fields used including pica+ fieldcode

”010@a”: ”taal”,
”011@a”: ”jaar”,
”011@e”: ”oorspronkelijk jaar”,
”019@0”: ”land”,
”021Aa”: ”titel”,
”021Ad”: ”ondertitel”,
”032@a”: ”editie”,
”033Ap”: ”plaats”,
”033An”: ”uitgever”,
”037Aa”: ”omschrijving”,
”004A0”: ”ISBN”,
”004Z8”: ”trefwoord”,
”020Aa”: ”annotatie”,
”039Dc”: ”vertaling van”,
”137Aa”: ”lokale annotatie”,
”021Aj”: ”2e auteur vermelding”,
”021Ah”: ”hoofdauteur vermelding”

We additionally use fields 039B and 039E which denote relations between titles, very important thus. For fieldname we use the description of the relation (039Ba and 039Ea), and for field value the corresponding title (039Bc and 039Ec). We further use fields 150C 028C and 028A to create the main and second author field. Short expression string excludes the following fields: ”ISBN”, ”omschrijving”, ”annotatie”, ”lokale annotatie”, ”plaats” and ”jaar”.

Short work string excludes the following fields: ”taal”, ”ISBN”, ”omschrijving”, ”annotatie”, ”lokale annotatie”, ”uitgever”, ”plaats”, ”oorspronkelijk jaar”, ”jaar”, ”2e auteur vermelding” and

”hoofdauteur vermelding”.

## C Baseline algorithm

### Work Matching Rules

1. **Fuzzy Match on Title and Author**

$$\text{If Fuzzymatch}(r_1.\text{title},\ r_2.\text{title}) \text{ and Fuzzymatch}(r_1.\text{author},\ r_2.\text{author}),$$

$$\text{then WorkMatch}(r_1, r_2) = \text{true}$$

2. **Translation Field Contains Other Title**

$$\text{If } r_1.\text{translation field contains } r_2.\text{title or vice versa},$$

$$\text{then WorkMatch}(r_1, r_2) = \text{true}$$

### Expression Matching Rules

1. **Do Not Match if the languages are different**

$$\text{If } r_1 \text{ or } r_2 \text{ is a translation, then ExpressionMatch}(r_1, r_2) = \text{false}$$

2. **Do Not Match if Edition Mentions "herzien" or "herz."**

$$\text{If } r_1.\text{edition or } r_2.\text{edition contains "herzien" or "herz.",}$$

$$\text{then ExpressionMatch}(r_1, r_2) = \text{false}$$

3. **Match on either Exact Title, Author, or Publisher match**

$$\text{If } r_1.\text{title} = r_2.\text{title, or } r_1.\text{author} = r_2.\text{author, or } r_1.\text{publisher} = r_2.\text{publisher},$$

$$\text{then ExpressionMatch}(r_1, r_2) = \text{true}$$

This rule-based baseline is intentionally simple and designed to simulate a matching system with minimal time and resource investment.

## D Examples of dataset

Here there are some more example positive and negative pairs.

### Positive work pairs

COL taal VAL fra COL jaar VAL 189X COL titel VAL @Sans Familie COL hoofdauteur vermelding VAL par Hector Malot COL 2e auteur vermelding VAL ill. par É. Bayard COL hoofdauteur VAL Malot, Hector Henri Malot 1830-1907 COL 2e auteur VAL Bayard, Émile Antoine Bayard 1837-1891 COL plaats VAL Paris COL uitgever VAL Librairie Hachette
COL taal VAL fra COL jaar VAL 1879 COL titel VAL @Sans famille COL hoofdauteur vermelding VAL par Hector Malot COL hoofdauteur VAL Malot, Hector Henri Malot 1830-1907 COL editie VAL 11e éd COL plaats VAL Paris COL uitgever VAL E. dentu COL omschrijving VAL Ouvrage couronné par l'académie Francaise

COL taal VAL ned COL jaar VAL 1964 COL titel VAL @Clubhuis de Crocus COL hoofdauteur vermelding VAL door Freddy Hagers COL 2e auteur vermelding VAL geïllustreerd door Rudy van Giffen COL hoofdauteur VAL Hagers, Freddy Hagers COL 2e auteur VAL Giffen, Rudy van Giffen 1929-2005 COL plaats VAL Alkmaar COL uitgever VAL Kluitman COL omschrijving VAL Voor jongens en meisjes tot 10 jaar

COL taal VAL ned COL jaar VAL 1939 COL titel VAL Het @clubhuis ”De Crocus” COL hoofdauteur vermelding VAL door Freddy Hagers COL 2e auteur vermelding VAL geïllustreerd door Rie Reinderhoff COL hoofdauteur VAL Hagers, Freddy Hagers COL 2e auteur VAL Reinderhoff, Marie-Louise Reinderhoff 1903-1991 COL plaats VAL Alkmaar COL uitgever VAL Gebr. Kluitman COL omschrijving VAL Een vrolijk boek voor meisjes

COL ISBN VAL 3891062389 COL taal VAL dui COL jaar VAL 1995 COL titel VAL @Blinker und der Blaue Morgenstern COL hoofdauteur vermelding VAL Marc de Bel COL 2e auteur vermelding VAL aus dem Niederländischen von Silke Schmidt COL hoofdauteur VAL Bel, Marc De Bel 1954- COL 2e auteur VAL Schmidt, Silke Schmidt COL plaats VAL Weinheim COL uitgever VAL Anrich COL vertaling van VAL Blinker en het BagBag-juweel. - Leuven : Davidsfonds/Infodok ; Amsterdam : Infodok, 1991
COL ISBN VAL 3407783558 COL taal VAL dui COL jaar VAL 1999 COL titel VAL @Blinker und der blaue Morgenstern COL ondertitel VAL Abenteuer-Roman COL hoofdauteur vermelding VAL Marc de Bel COL 2e auteur vermelding VAL aus dem Niederländischen von Silke Schmidt COL hoofdauteur VAL Bel, Marc De Bel 1954- COL 2e auteur VAL Schmidt, Silke Schmidt COL plaats VAL Weinheim [etc.] COL uitgever VAL Beltz & Gelberg COL vertaling van VAL Blinker en het BagBag-juweel. - Leuven : Davidsfonds/Infodok ; Amsterdam : Infodok, 1991

## Negative work pairs

COL taal VAL ned COL jaar VAL 2012 COL titel VAL De @ring van de hartstocht COL hoofdauteur vermelding VAL Danielle Steel COL 2e auteur vermelding VAL [vert. uit het Engels: Conny van Manen] COL hoofdauteur VAL Steel, Danielle Fernande Steel 1947- COL 2e auteur VAL Manen, Conny de Kleuver-van Manen COL editie VAL 8e dr COL plaats VAL Amsterdam COL uitgever VAL Poema Pocket COL vertaling van VAL The ring. - London : Hodder & Stoughton, 1980
COL ISBN VAL 9024549515 COL taal VAL ned COL jaar VAL 2003 COL titel VAL Een @tijd van hartstocht COL hoofdauteur vermelding VAL Danielle Steel COL 2e auteur vermelding VAL [vert. uit het Engels: Suzanne Braam] COL hoofdauteur VAL Steel, Danielle Fernande Steel 1947- COL 2e auteur VAL Braam, Suzanne Braam COL plaats VAL Amsterdam COL uitgever VAL Poema Pocket COL vertaling van VAL Season of passion. - Cop. 1979

COL taal VAL ned COL jaar VAL 1957 COL titel VAL @Maigret in het Wilde Westen COL hoofdauteur vermelding VAL Georges Simenon COL 2e auteur vermelding VAL [vert. uit het Frans door V.H. Uurbanus-Harbrink Numan] COL hoofdauteur VAL Simenon, Georges Joseph Christian Simenon 1903-1989 COL 2e auteur VAL Uurbanus-Harbrink Numan, V.H. Uurbanus-Harbrink Numan 1925-1960 COL plaats VAL Utrecht COL uitgever VAL Bruna COL vertaling van VAL Maigret chez le coroner. - 1950
COL taal VAL fra COL jaar VAL 1969 COL titel VAL @Maigret et le tueur COL ondertitel VAL roman COL hoofdauteur vermelding VAL Georges Simenon COL hoofdauteur VAL Simenon, Georges Joseph Christian Simenon 1903-1989 COL plaats VAL Paris COL uitgever VAL Presses de la Cite

COL ISBN VAL 9041011951 COL taal VAL ned COL jaar VAL 2002 COL titel VAL @Dribbel een dagje thuis COL hoofdauteur vermelding VAL Eric Hill COL 2e auteur vermelding VAL [vert. uit het Engels] COL hoofdauteur VAL Hill, Eric Hill 1927-2014 COL plaats VAL Houten COL uitgever VAL Unieboek/Kitt COL vertaling van VAL Spot my day at home. - London : Warne/Ventura, cop. 2002
COL taal VAL ned COL jaar VAL 2007 COL titel VAL @Dribbels dagje uit COL hoofdauteur vermelding VAL Eric Hill COL 2e auteur vermelding VAL [vert. uit het Engels: Unieboek] COL hoofdauteur VAL Hill, Eric Hill 1927-2014 COL plaats VAL Houten [etc.] COL uitgever VAL Van Holkema & Warendorf COL omschrijving VAL Omslag omvat geluidseffectentableau COL vertaling van VAL Spot’s day out. - London : Warne, cop. 2006

## Positive expression pairs

COL ISBN VAL 9021413310 COL taal VAL ned COL jaar VAL 1973 COL titel VAL @Total loss, weetjewel COL hoofdauteur vermelding VAL Miep Diekmann COL 2e auteur vermelding VAL met tekeningen van The Tjong Khing COL hoofdauteur VAL Diekmann, Maria Hendrika Jozina Diekmann 1925-2017 COL 2e auteur VAL The Tjong Khing, Thé Tjong-Khing 1933- COL plaats VAL Amsterdam COL uitgever VAL Querido

COL ISBN VAL 9021431211 COL taal VAL ned COL jaar VAL 1991 COL titel VAL @Total Loss, weetjewel COL hoofdauteur vermelding VAL Miep Diekmann COL 2e auteur vermelding VAL met tek. van The Tjong Khing COL hoofdauteur VAL Diekmann, Maria Hendrika Jozina Diekmann 1925-2017 COL 2e auteur VAL The Tjong Khing, Thé Tjong-Khing 1933- COL editie VAL 6e dr COL plaats VAL Amsterdam COL uitgever VAL Querido

COL taal VAL ned COL jaar VAL 1973 COL titel VAL @Oorlog & vrede COL hoofdauteur vermelding VAL Leo Tolstoj COL 2e auteur vermelding VAL bewerking [naar het Russisch]: T. Knape COL hoofdauteur VAL Tolstoj, L.N. Tolstoj 1828-1910 COL 2e auteur VAL Knape, T. Knape COL plaats VAL Amsterdam COL uitgever VAL Amsterdam Boek COL vertaling van VAL Vojna i mir. - Moskou, 1937. - Oorspronkelijke uitgave: 1868-1869
COL ISBN VAL 9064075999 COL taal VAL ned COL jaar VAL 2001 COL titel VAL @Oorlog en vrede COL hoofdauteur vermelding VAL Lev Tolstoj COL 2e auteur vermelding VAL met ill. van Christian Wilhelm Faber du Faur COL 2e auteur vermelding VAL [bew.: T. Knape COL 2e auteur vermelding VAL vert. uit het Russisch] COL hoofdauteur VAL Tolstoj, L.N. Tolstoj 1828-1910 COL 2e auteur VAL Knape, T. Knape COL plaats VAL Amsterdam [etc.] COL uitgever VAL The Reader's Digest COL vertaling van VAL Vojna i mir. - 1876-1869 COL auteur/primair VAL Tolstoj

COL taal VAL eng COL jaar VAL 1905 COL titel VAL The @tale of the Mrs. Tiggy-Winkle COL hoofdauteur vermelding VAL by Beatrix Potter COL hoofdauteur VAL Potter, Helen Beatrix Potter 1866-1943 COL plaats VAL London COL plaats VAL New York COL uitgever VAL Frederick Warne
COL taal VAL eng COL jaar VAL 1905 COL titel VAL The @tale of Mrs. Tiggy-Winkle COL hoofdauteur vermelding VAL Beatrix Potter COL hoofdauteur VAL Potter, Helen Beatrix Potter 1866-1943 COL plaats VAL London [etc.] COL uitgever VAL Warne

## Negative expression pairs

COL taal VAL ned COL jaar VAL 2003 COL titel VAL De @Kleine Zeemeermin COL hoofdauteur vermelding VAL [creation, text and illustraties: A.M. Lefèvre ... et al.] COL 2e auteur VAL Lefèvre, A.M. Lefèvre COL plaats VAL Oegstgeest COL uitgever VAL Boek Specials Nederland COL taal VAL ned COL jaar VAL 2011 COL titel VAL De @kleine zeemeermin COL hoofdauteur vermelding VAL [naar] Disney COL 2e auteur vermelding VAL [verteller Arnold Gelderman COL 2e auteur vermelding VAL vert. uit het Engels Jan Derk Beck] COL 2e auteur VAL Disney, Walter Elias Disney 1901-1966 COL 2e auteur VAL Gelderman, Arnold Gelderman COL 2e auteur VAL Beck, Jan Derk Beck COL plaats VAL Amsterdam COL uitgever VAL Rubinstein
COL ISBN VAL 9026611722 COL taal VAL ned COL jaar VAL 2003 COL titel VAL @Houen jongens! COL hoofdauteur vermelding VAL K. Norel COL 2e auteur vermelding VAL met ill. van Roel Ottow COL hoofdauteur VAL Norel, Klaas Norel 1899-1971 COL 2e auteur VAL Ottow, Roelof Rudolf Ottow 1955- COL editie VAL 12e dr COL plaats VAL Kampen COL uitgever VAL Callenbach COL lokale annotatie VAL Toegevoegd het hoofdstuk: Vijftig jaar later, Tholen en Sint-Philipsland; geschreven door C. Overwater COL lokale annotatie VAL Omslagontwerp: Peter Dees. Layout/dtp: Gerard De Groot COL lokale annotatie VAL De speciale uitgave kwam tot stand i.s.m. de Werkgroep 'Herdenking Watersnoodramp 1953' van de Heemkundekring 'Philippuslandt' in Sint-Philipsland en is aangeboden aan alle basisschoolleerlingen in de gemeente Tholen
COL taal VAL eng COL jaar VAL 1955 COL titel VAL @Stand by, boys! COL ondertitel VAL a true story about Holland's fight against the sea COL hoofdauteur vermelding VAL by K. Norel COL 2e auteur vermelding VAL transl. from the Dutch by Marian M. Schoolland COL hoofdauteur VAL Norel, Klaas Norel 1899-1971 COL 2e auteur VAL Schoolland, Marian M. Schoolland 1902- COL plaats VAL Grand Rapids, Michigan COL uitgever VAL Wm.B. Eerdmans Publishing Company COL vertaling van VAL Houen, jongens!. - Nijkerk : Callenbach, 1953

COL taal VAL ned COL jaar VAL 2014 COL titel VAL @Karlsson van het dak COL hoofdauteur vermelding VAL Astrid Lindgren COL 2e auteur vermelding VAL vertaald [uit het Zweeds] door Rita Törnquist-Verschuur COL 2e auteur vermelding VAL met tekeningen van Georgien Overwater COL hoofdauteur VAL Lindgren, Astrid Anna Emilia Lindgren 1907-2002 COL 2e auteur VAL Törnquist-Verschuur, Marguérite Elisabeth Verschuur 1935- COL 2e auteur VAL Overwater, Georgien Overwater 1958- COL editie VAL Vijfde, herziene druk COL plaats VAL Amsterdam COL uitgever VAL Uitgeverij Ploegsma COL vertaling van VAL Lillebror och Karlsson pa taket. - Raben & Sjögren COL Oorspronkelijke titel VAL Erik en Karlsson van het dak. - Amsterdam : C.P.J.van der Peet, 1959COL lokale annotatie VAL Omslagontwerp: Steef Liefting. - Vormgeving binnenwerk: Studio Cursief, Irma

Hornman

COL taal VAL ned COL jaar VAL 1968 COL titel VAL @Karlsson van het dak COL hoofdauteur vermelding VAL Astrid Lindgren ;omslag en illustraties Ilon Wikland COL 2e auteur vermelding VAL vertaald uit het Zweeds door Rita Törnqvist-Verschuur COL hoofdauteur VAL Lindgren, Astrid Anna Emilia Lindgren 1907-2002 COL 2e auteur VAL Wikland, Ilon Wikland 1930- COL 2e auteur VAL Törnqvist-Verschuur, Marguérite Elisabeth Verschuur 1935- COL editie VAL 2e dr COL plaats VAL Amsterdam COL uitgever VAL Ploegsma COL vertaling van VAL Lillebror och Karlsson pa taket COL Oorspr. titel VAL Erik en Karlsson van het dak