International Journal of Geo-Information

MDPI

Article

Context-Aware Search for Environmental Data Using Dense Retrieval

Simeon Wetzel * and Stephan Mäs

Chair of Geoinformatics, Technische Universität Dresden, Helmholtzstraße 10, 01069 Dresden, Germany; stephan.maes@tu-dresden.de

* Correspondence: simeon.wetzel@tu-dresden.de

Abstract: The search for environmental data typically involves lexical approaches, where query terms are matched with metadata records based on measures of term frequency. In contrast, dense retrieval approaches employ language models to comprehend the context and meaning of a query and provide relevant search results. However, for environmental data, this has not been researched and there are no corpora or evaluation datasets to fine-tune the models. This study demonstrates the adaptation of dense retrievers to the domain of climate-related scientific geodata. Four corpora containing text passages from various sources were used to train different dense retrievers. The domain-adapted dense retrievers are integrated into the search architecture of a standard metadata catalogue. To improve the search results further, we propose a spatial re-ranking stage after the initial retrieval phase to refine the results. The evaluation demonstrates superior performance compared to the baseline model commonly used in metadata catalogues (BM25). No clear trends in performance were discovered when comparing the results of the dense retrievers. Therefore, further investigation aspects are identified to finally enable a recommendation of the most suitable corpus composition.

Keywords: IR; information retrieval; GeoAI; SDI

Citation: Wetzel, S.; Mäs, S. Context-Aware Search for Environmental Data Using Dense Retrieval. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 380. https://doi.org/10.3390/ijgi13110380

Academic Editors: Wolfgang Kainz, Mara Nikolaidou, Christos Chalkias, Marinos Kavouras and Margarita Kokla

Received: 13 September 2024
Revised: 21 October 2024
Accepted: 28 October 2024
Published: 30 October 2024

1. Introduction

Information retrieval (IR) in spatial data infrastructures (SDIs) is commonly handled by metadata catalogues. These catalogue services provide a search interface in the frontend and are connected to a database and an inverted search index in the backend. Traditional IR approaches in geospatial metadata data catalogues commonly rely on keyword-based (lexical) search approaches using bag-of-words (BOW) retrieval models [1] (e.g., BM25 [2]). Therefore, the search terms of a query (i.e., keywords) are compared to the terms contained in different fields of the metadata records (e.g., title, keywords, abstract). This approach has certain benefits:

(1) It is set up quickly: BM25 is already implemented in established Lucene-based search indexes such as ElasticSearch [3] or Apache SOLR [4].
(2) It is performant and efficient: efficiency tests by [5] showed that BM25 outperforms other retrieval methods in terms of retrieval latency and the required index size.

However, a lexical search suffers from several limitations, such as the vocabulary mismatch problem [6], wherein queries containing synonyms, homonyms, acronyms, or misspelled terms fail to retrieve relevant metadata records. For instance, if a user enters the search term "*precipitation data*", only records that contain the word "*precipitation*" will be retrieved. Metadata records containing synonyms of "*precipitation*" such as "*rain*" or "*rainfall*" will not be retrieved unless a synonym register or ontology is configured to match these terms. The need for an effective geospatial data search becomes even more critical in domains like climate adaptation, where the responsible policymakers may not be familiar with the domain-specific terminology used in the metadata descriptions of research data and still need to retrieve the data or information effectively.

In the context of geospatial metadata catalogues, methodologies of Semantic Web and Linked (Open) Data (e.g., ontologies, knowledge graphs) have been commonly used to minimise the potential lexical gap between queries and the retrieved metadata records [7–9]. However, widely used metadata catalogues such as GeoNetwork [10] or CKAN [11] do not offer an out-of-the-box integration of ontologies or knowledge graphs. Therefore, both the integration and the required curation of ontologies result in additional efforts.

Another drawback of lexical retrieval in the context of geospatial metadata retrieval is the dependence on accurate and complete metadata [12]. Although the geospatial community has put a lot of effort in standardisation (c.f. [13,14]) and increasing metadata quality [15], presenting consistent and complete metadata in publicly available metadata catalogues is still an issue [16,17].

Aside from a lexical search, IR approaches that use dense retrieval models, and, in particular, transformer models such as BERT-based models [18], have become increasingly popular [5]. Pre-trained language models are used to obtain a meaningful semantic and contextual representation of queries and documents. The models generate dense vectors—i.e., numerical representation of a text's semantics (commonly called embeddings)—for queries and documents that are used to rank retrieved documents with a similarity score [19]. This IR approach is referred to as a neural search [20]. In contrast to BOW approaches, the order and context of each word is taken into consideration. Further, synonyms, abbreviations, and misspelled words can be automatically handled to a certain extent. However, this requires that the language model has been trained on a corpus that includes these terms and their particular semantic context and use in the domain. In the past few years, a substantial number of models have been published that are pre-trained on general or on various domain-specific corpora (e.g., SciBERT for understanding scientific texts [21] or SpaBERT for geo-entity representation [22]). To the best of our knowledge, no models exist that are trained to support geospatial metadata retrieval.

Such domain adaption is usually relatively costly and therefore not always effective. However, recent works such as [23,24] have made domain adaptation more feasible by demonstrating that it is possible to achieve superior performance without the need for a labelled corpus. In this context, "labelling" means the assignment of ranking or similarity scores between queries and the documents within the corpus, which requires considerable effort when creating a new corpus. This addresses the need for more self-supervised training methods in the domain of GeoAI (geospatial artificial intelligence) to enhance scalability and reduce the reliance on labelled data. This requirement was already mentioned in 2019 in the work of [25], which provides an overview of progress and future research directions in the field of GeoAI. Recent studies, such as [26,27], demonstrate that this requirement continues to be relevant.

The methodology proposed in the following sections demonstrates the design of an unlabelled corpus required for the domain adaption of a pre-trained BERT model for geospatial metadata retrieval with a special focus on climate-related data. The retrieval model is fine-tuned with various corpus compositions to compare and assess the best performing corpus configuration. The resulting fine-tuned retrieval models are primarily optimised to retrieve thematically relevant records. For environmental datasets, retrieving spatially relevant data is crucial, as users often search for datasets of specific geographic areas. To address this requirement, this study proposes a method for re-ranking the results obtained from the retrieval models fine-tuned with the aforementioned corpora. A prototypical implementation of a neural search in a metadata catalogue demonstrates context-aware search's feasibility and advantages.

2. Related Work

Several prior studies have addressed the improvement of IR in the context of geodata catalogues with diverse solutions. Works by Lacasta et al. [28,29] have addressed an often occurring mismatch between the users' query demands and the returned metadata records. Specifically, ref. [29] describes that retrieved datasets often do not entirely cover the area of

interest that is requested by the user. To improve this, a method is proposed to semantically cluster metadata records and then return these aggregated results. Ref. [30] introduced a method to measure the similarity of geospatial metadata records with neural networks that could be combined with the approach proposed by [29].

Many prior studies have highlighted the importance of semantic-based retrieval approaches using, for example, knowledge bases or ontologies [31–34]. Domain ontologies provide relevant concepts and definitions. Therewith, they support the harmonisation of keywords contained in metadata records as well as the query formulation and query refinement based on the hierarchical structure of the concepts [31]. For instance, queries can be extended by terms that are semantically similar according to the domain vocabulary (query expansion) [35], or terms with multiple meanings can be resolved (termed disambiguation resolution) [32]. Linking metadata records to concepts of an ontology or knowledge graph allows for a more effective search. Therewith, all records that are semantically similar to the query can be fetched and a higher recall can be expected [8,34]. Additionally to thematic query expansion, it is also feasible to extend the query terms with spatially explicit context, as proposed by [36]. However, the semantic-based approaches also have some disadvantages. The authors of [32] noted the constant need to update ontologies as the domain evolves and the concepts alter. Ref. [37] developed a custom ontology for geospatial data and identified additional challenges related to ontologies, such as the absence of metrics to evaluate their quality, completeness, or accuracy, as well as difficulties in integrating custom ontologies with pre-existing ones.

These limitations of traditional semantic-based approaches and the ability of dense retrieval approaches to match queries and documents based on their semantic context motivated a dense retrieval approach. In this context, there are previous works that demonstrated the fine-tuning of models for scientific information retrieval. For instance, ref. [38] introduced a semantic search for COVID-19-related publications. Their proposed search engine used a hybrid approach, combining a dense retriever (SBERT model [39]) and two sparse retrievers (TF-IDF and BM25) to first generate document embeddings (context-aware vector representations) and subsequently index these. During inference, query embeddings were produced using all three models. These embeddings were then employed to identify the most proximate document embeddings to the query embeddings by computing retrieval scores. These scores were combined to create a ranked candidate list. Finally, the retrieved documents were parsed into a question-answering model. Ref. [40] demonstrated a novel approach to enhance BERT-based dense retrievers by incorporating spatial context awareness. The proposed method considers the potential geographic distances between spatial entities (e.g., location names) contained in queries and documents. This spatial context heuristic is applied during fine-tuning on a subset of the well-established MS-MARCO benchmark dataset [41]. The authors identify documents that semantically match the query but include spatial entities that are distant from the ones in the query. These documents are considered hard-negative training samples. Hard-negative samples are negative examples that closely resemble the True positive results. These samples are particularly challenging for the model because they are similar to the correct answers, making it difficult to distinguish them from the actual positive cases. The resulting models are then considered to have acquired spatially explicit knowledge, enabling spatial ranking without relying on external sources such as gazetteers. Such models can then be effectively employed for question-answering tasks that include geographic aspects (i.e., questions about places).

Parts of the aforementioned dense retrieval approaches can be incorporated into the method proposed in this paper. However, the method of [38] is optimised for question-like queries and the retrieved documents are scientific publications. The spatial ranking introduced by [40] offers an innovative alternative to traditional spatial filtering methods, which typically rely on the spatial extent (bounding box) specified in metadata records. However, such an approach has limitations when it comes to geographic entities (i.e., regions or place names) that are under-represented in the training dataset. Achieving spatial ranking in

dataset retrieval necessitates a model that can effectively encode the spatial context of each location entity within queries and metadata records, regardless of whether the entities occur in the training data (zero-shot learning). Other recent advancements, such as the spatially aware encoding model introduced by [36] (which utilises a subset of DBPedia with a spatial extent limited to the U.S.), or GeoBERT by [42] (a BERT-based model that learns point-of-interest representations from Chinese city data) may encounter similar constraints. The necessity of developing models that are capable of generalising effectively across diverse geographical contexts was identified as a key challenge for the field of GeoAI research [25]. The review paper from [43] discusses several location encoding techniques that are applicable in the context of GeoAI applications. However, these methods require input features with geometries, such as points, polylines, or raster data. In metadata catalogues, these geometries might be present in the metadata as spatial extents (bounding boxes), but the queries of full-text searches include only textual features such as place names. Therefore, the location encoding methods are not suitable for ad hoc metadata retrieval. Instead, we propose applying traditional techniques for incorporating the spatial context of the queries, such as named entity recognition (NER) and geocoding such as that proposed by [44], and ranking results based on the geocoded entities in the query and the bounding boxes present in the metadata records. This study recommends a two-stage process: an initial retrieval stage using the models described in the following sections, without spatial ranking, followed by a spatial re-ranking stage employing the aforementioned method. This approach allows for a more nuanced and accurate integration of spatial information in the search process.

3. Materials Methods

The following section describes the steps to create a domain-adapted model for spatial metadata IR. Specifically, the steps include the design of the corpora and the refinement of the model using a training algorithm along with the corresponding corpora. The spatial re-ranking process mentioned earlier is also briefly described. A prototype demonstrating the integration of the model into a realistic metadata catalogue setup utilising a CKAN catalogue is presented. Finally, methods for evaluating the presented IR system are elaborated upon.

3.1. Domain Adaptation

3.1.1. Corpus Design

The domain adaptation of a dense retrieval model requires a corpus consisting of passages that are representative of the target domain. As mentioned, ref. [23] introduced a powerful method called GPL (Generative Pseudo Labelling), which does not require a labelled corpus. GPL was used for domain adaptation in this study. Therefore, 33,314 text passages considered relevant for the geospatial and climate-related data search were collected. The passages include an average of 202 words per passage (total number of words included in the collection: 6,715,384). The passages originate from the following sources:

(I) **Dataset descriptions** from openly accessible metadata catalogues, namely the *EEA geospatial data catalogue* [45], *United Nations FAO Map Catalogue* [46], *Copernicus Data Store* [47], and *Data portal of the European Commission* [48] (10573 metadata records).

(II) **Ontology concepts** from the *General Multilingual Environmental Thesaurus* (GEMET) [49]. A subset of 187 concept definitions were selected. Concepts related to the GEMET themes "climate" [50] and "natural dynamics" [51] as well as concepts assigned to the GEMET group "ATMOSPHERE (air, climate)" [52] were used for this study.

(III) **Scientific literature** from open access *Copernicus journals* [53] and scientific textbooks (passages from two basic literature sources were parsed: [54,55]). Thematically relevant journals (a selection of relevant Copernicus journals: *Atmospheric Chemistry and Physics* (ACP), *Atmospheric Measurement Techniques* (AMT), *Advances in Statistical Climatology, Meteorology and Oceanography* (ASCMO), *Earth System Dynamics* (ESD), *Hydrology and Earth System Sciences* (HESS), and *Natural Hazards and Earth System*

Sciences (NHESS)) have been selected from the existing Copernicus publications and all available online abstracts were downloaded (21.618 abstracts and 2 textbooks).

One goal of this study was to explore how various domain-specific text sources influence the effectiveness of the retrieval. Specifically, the aim was to compare the performance of the retrieval model when fine-tuned with a single domain-specific text type (e.g., only dataset descriptions) or if and how the results improved when a more diverse domain-specific corpus was utilised (e.g., dataset descriptions and ontology concepts).

Table 1 shows how the passage collection was used to compose different corpora. Corpus-(1a) only included dataset descriptions. This represents the knowledge a conventional search is based on. Corpus-(1b) combined the dataset descriptions with GEMET concepts and their respective definitions. The aim was to incorporate more domain-specific comprehension into the model. The third corpus (2) included only the scientific literature. This corpus is intended to test whether the literature is sufficient for a domain adaptation, although the structure of these text passages differs from the target passages of the search in metadata catalogues. The last corpus (3) combines all the previous passages. This is to check whether a model trained with a more extensive and heterogeneous corpus performs significantly better.

Table 1. Different corpus compositions used for the domain adaptation of four dense retrievers.

Corpus Compositions	Contents	Number of Text Passages	Average Number of Words per Passage	Number of Words
(1a)	Dataset descriptions	10,573	92	975,659
(1b)	Dataset descriptions + ontology concepts	10,760	96	1,038,174
(2)	The scientific literature	22,137	255	5,645,063
(3) = (1b) + (2)	The scientific literature + dataset descriptions + ontology concepts	33,314	202	6,715,384

3.1.2. Training Method

The unsupervised domain adaptation method GPL [23] was used for the domain adaptation of the retrievers. GPL uses a text corpus and generates synthetic queries, tailored to the contents of the corpus, using a docT5query text generation model [56]. These queries are then utilised by the training algorithm to form training samples, comprising query-positive and query-negative pairs. A cross-encoder model ("ms-marco-MiniLM-L-6-v2") [57] fine-tuned on passage retrieval is then used to generate similarity scores (labels) for all query–document pairs. Thereafter, the pre-trained (not domain-adapted) model is fine-tuned using these training pairs with the generated labels employing MarginMSE [58] as the loss function. In this study, the pre-trained model DistilBERT-base ("distilbert-base-uncased", introduced by [59]) is used. The reason for choosing this model is that it retains most of the capabilities of the original BERT-base model ([18]) while being faster and more lightweight. It is well suited for use as a retriever in a neural search, offering efficient performance even with limited computing resources (e.g., without GPU acceleration), or in large-scale catalogue applications involving large volumes of metadata records. The queries generated by docT5query are formulated like questions. However, the queries in the target system (metadata catalogue) are expected to be based on listed search terms. Therefore, keyword-based queries were generated instead of using the docT5query model. A BERT-based keyword extraction method called KeyBERT [60] was used to generate queries. The remaining GPL algorithm was applied as introduced in the original paper [24] including the proposed model checkpoints and hyper-parameters (1 epoch, with 140 k training steps and a batch size of 32). Finally, the fine-tuning of the corpora described above resulted in four domain-adapted checkpoints of DistilBERT-base.

For the domain adaptation, the models were fine-tuned on the High-Performance Computing (HPC) cluster of the TU Dresden using AMD "Rome" Processors (AMD EPYC CPU 7352) with NVIDIA A100-SXM4 Tensor Core-GPUs. The training jobs took around 10–12 h depending on the corpus size and the respective computing load and capacity of the HPC cluster.

Analogous to the corpus names (see Table 1), the domain-adapted models are named after the respective corpus with which the model was fine-tuned: corpus-1a, corpus-1b, corpus-2, and corpus-3.

3.2. Spatial Re-Ranking Method

While dense retrieval models are primarily fine-tuned for thematic relevance, they do not inherently consider spatial aspects. This limitation becomes apparent when dealing with environmental datasets where geographical relevance is crucial. To address this issue, we propose an extension of the dense retrieval approach through a subsequent spatial re-ranking stage.

For spatial ranking, we assume that the records stored in the environmental metadata catalogue include a spatial extent property in the form of a bounding box. The following workflow was then applied in the experiments described in Section 4.2.

(1) Spatial context parsing: A geocoding service was developed to parse the spatial context from the query. The service uses the following:
   (a) A BERT model fine-tuned on named entity recognition [61] to extract location entities from a query (e.g., "*Berlin*" from the query "*climate data berlin*").
   (b) An open-source geocoding service [62] to generate a query bounding box (e.g., "*[13.088345, 52.3382448, 13.7611609, 52.6755087]*" for "*Berlin*").

(2) Calculating a spatial similarity metric: To calculate the spatial similarity metric between the query bounding box and the bounding boxes of candidates retrieved by the dense retriever, the Hausdorff distance is an appropriate measure, as proposed by [28,63]. The Hausdorff distance takes into account both the size and position of the geometries (polygons in this case). Unlike other metrics such as area of overlap, which can yield a value of zero when there is no overlap between the bounding boxes, the Hausdorff distance provides a more robust measure for re-ranking as it accounts for the spatial proximity and geometry even without a direct spatial overlap.

(3) Re-ranking: the final step is to re-rank the results according to the descending Hausdorff distances.

One drawback of spatial re-ranking is the risk of promoting results with high spatial relevance but low thematic relevance, which are initially ranked low. To address this, it is essential to limit the re-ranking process to a smaller subset of the top-k retrieved candidates. In our experiments (see Section 4.2), we determined that it was beneficial to only re-rank the top 30 candidates retrieved by the dense retriever. This threshold is dataset-dependent. In cases where spatial information is well documented in the dataset descriptions, a different balance between spatial and thematic relevance may be observed, which may require adjustments to the threshold.

3.3. Prototype Architecture

After the domain adaptation (see Section 3.1.2), the models can be used for dense retrieval. Since conventional metadata catalogues only support searches based on BOW representations, the catalogue's search and indexing processes must be adapted to facilitate dense retrieval. Figure 1 illustrates the customised search and indexing workflow. For the prototype, an established open-source cataloguing solution CKAN (Comprehensive Knowledge Archive Network) was used. CKAN is connected to the inverted index Apache-SOLR, which is based on Lucene. By default, BM25 is used for retrieving documents (i.e., metadata records) from the index. Since using dense retrievers was not supported, the following configurations and extensions became necessary:

(1) The SOLR index needed to be configured to store the embeddings produced by the dense retriever. Therefore, an additional field for storing multidimensional vectors was added to the SOLR schema.
(2) Also, SOLR does not inherently support the calculation of scores based on embeddings. There is a plugin available for SOLR (also for ElasticSearch) that supports the score calculations using the query and document embeddings. This plugin, called SOLR Vector Scoring Plugin [64], was installed on the SOLR instance.
(3) Moreover, a mechanism was required to generate embeddings for both queries and documents. For the prototype, a text embedding service was set up using the framework fastAPI [65]. It provides an API that takes text as input and returns embeddings using the SBERT model.
(4) Finally, a mechanism was required that changes the built-in search method of CKAN from BM25 to the custom search approach using the dense retriever. For that purpose, a CKAN extension [66] was developed.

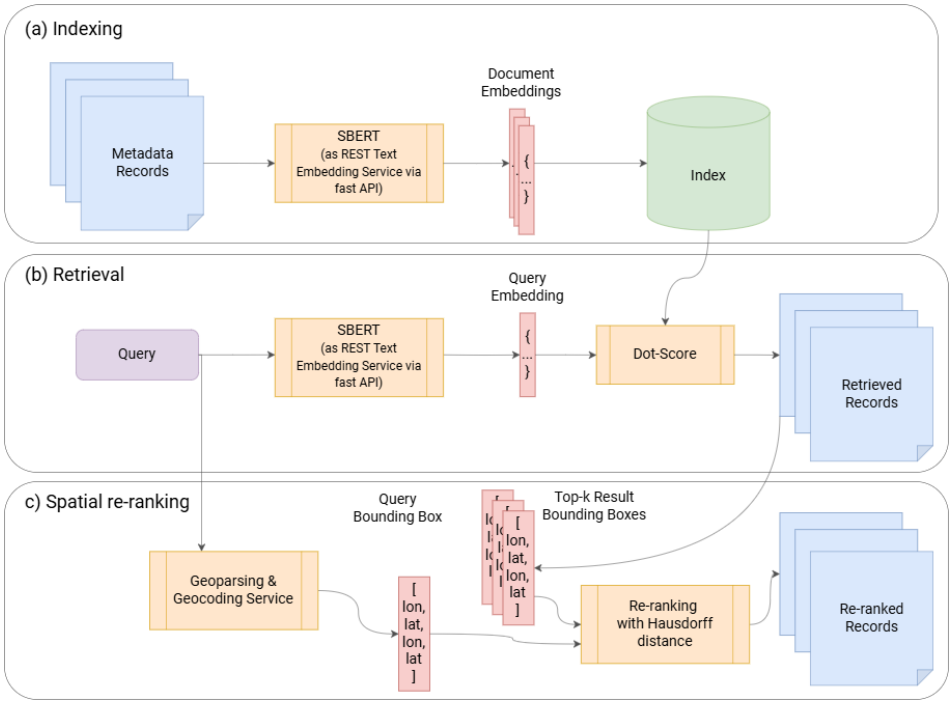

Figure 1. Neural search architecture: (**a**) indexing stage, (**b**) retrieval stage (search), and (**c**) re-ranking stage. The text embedding service using the SBERT model is used in both stages.

The extensions replace the two built-in stages indexing (a) and retrieval (b) and an additional re-ranking stage (c) as shown in Figure 1. For every metadata record that is inserted or updated in the catalogue, a document embedding is generated by passing a concatenated text passage from the title and description field to the text embedding service. The respective document embedding is then stored in the SOLR index. Without the extension, CKAN would only transmit the metadata without the document embedding into the SOLR index. During retrieval stage (b), the built-in search in CKAN passes the search terms and optional search parameters to the SOLR index and fetches retrieved documents ranked by the index. For dense retrieval, it is necessary to generate a query embedding first and then calculate a ranking score between the query embedding and the indexed document embeddings. Therefore, the extension sends the query to the text embedding service, which generates the query embedding. Then, the score (so-called dot-score) between the query embedding and the stored document embeddings is generated by calculating the dot-product between the embeddings. Finally, the dot-scores are used to return the ranked metadata records. By default, the ranking contains all documents. Thus, a threshold value must be set to only return documents with a reasonable dot-score.

During the tests of the prototype, a threshold dot-score of 0.6 appeared to be suitable to retrieve the relevant candidates. However, this value has to be set individually depending on the use case. Influencing factors can be the number of metadata records and the similarity of the records. Further, a maximum number of results has to be defined for the case that too many records are above the threshold. In the prototype, the maximum number of results is set to 1000.

As described in Section 3.2 and tested in Section 4.2, it is essential to only consider the top-ranked (thematically relevant) records for spatial re-ranking. The prototype is configured to include the top 30 hits retrieved by the dense retriever into re-ranking.

To allow us to switch seamlessly between dense retrieval and the standard BM25-based search, the developed CKAN extension integrates into the default search user interface (see Figure 2). The main purpose of this feature is for testing purposes.

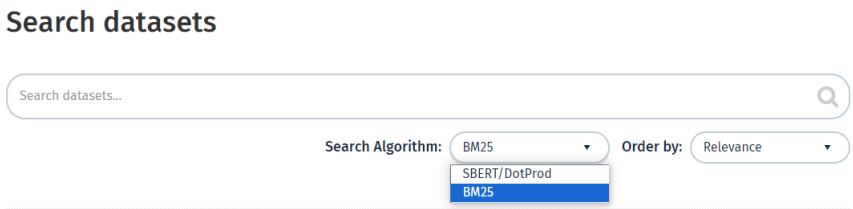

Figure 2. Selection of the search algorithm used in the prototype: Users can either select from the dropdown menu or by passing a URL parameter "*?algorithm=sbert*" or "*?algorithm=bm25*". BM25 is used by default.

3.4. Evaluation Method

3.4.1. Test Collection

Test collections are typically used to assess the effectiveness of an IR system. These collections consist of benchmark datasets with queries (also called topics) and documents that have been carefully annotated to determine their relevance to the given queries. The relevant annotations allow the calculation of evaluation metrics such as *Precision* or *Recall*. The existing domain-unspecific test collections are too generic for evaluating models fine-tuned on the specific task of an environmental data search. On the other hand, there are test collections for specific domains such as COVID-19-related research [67] or the biomedical domain [68], but none fitting for environmental data search. Therefore, it became necessary to create a test collection. A set of 1739 metadata records was harvested from the World Data Centre for Climate (WDCC) [69] provided by the German Climate Computing Center (DKRZ). This archive contains metadata records describing meteorological and climate-related data. In addition to the collected metadata descriptions, test queries were also needed. The metadata records of WDCC provide keywords. As the IR system was tested for matching documents with search term-based queries it seemed straightforward to pick these tags as queries in the test collection. Four keyword pairs, which were frequently used in the metadata of the WDCC platform, were selected for the test collection:

(1) Q1: "climate simulation";
(2) Q2: "greenhouse gases";
(3) Q3: "observation data";
(4) Q4: "aircraft measurement".

Finally, the harvested metadata records of the test collection had to be manually checked for their relevance to the above queries (1–4) and the relevance annotations had to be stored. These annotations categorise a metadata record either as "relevant" or as "irrelevant", with no indication regarding the degree of relevance. These annotations are essential for the calculations of the evaluation metrics described in the following section. Although the metadata contained the keyword pairs used for the queries, all records were double-checked for relevance to ensure that no relevant record had been missed.

3.4.2. Evaluation Metrics

To evaluate the effectiveness of the presented IR system, the following standard metrics were used: *Precision (P)*, *Recall (R)* and *(Mean) Average Precision (MAP/AP)*. The metrics *Precision* and *Recall* were calculated for different ranking levels k, where k refers to the number of top results retrieved by the models for a query. For instance, if k is set to 10, the evaluation considers the top 10 results for each query. However, *Precision* and *Recall* do not take into account the ranking order, whereas the AP value is sensitive to the order of the results and offers an impression of the ranking quality of an IR system. The metrics are defined as follows [70–72]:

$$P = \frac{TP}{TP + FP} \text{with } P_k \text{ as } P \text{ at rank } k \tag{1}$$

$$R = \frac{TP}{TP + FN} \text{with } R_k \text{ as } R \text{ at rank } k \tag{2}$$

where TP, FP, and FN are True Positive, False Positive, and False Negative search results. Table 2 specifies the meaning of these terms in the context of IR.

Table 2. Confusion matrix for True/False Positive/Negative documents in IR.

	Retrieved	**Not Retrieved**
Relevant	True Positive (TP)	False Negative (FN)
Irrelevant	False Positive (FP)	True Negative (TN)

The *Average Precision (AP)* is defined as

$$AP_k = \frac{1}{n} \sum_{k=1}^{n} P_k * rel_k \tag{3}$$

with the relevance function:

$$rel_k = \begin{cases} 1 & \text{if the item at rank } k \text{ is a relevant document} \\ 0 & \text{if the item at rank } k \text{ is an irrelevant document} \end{cases} \tag{4}$$

Here, n is the number of all available documents, k is the rank and P_k is the *Precision* at rank k. Table 3 illustrates an example where $n = 5$, $document_1$, $document_4$ and $document_5$ are relevant while the remaining documents are irrelevant:

Table 3. Example calculation of AP_k.

Rank k	1	2	3	4	5
Relevance of $document_k$	relevant	irrelevant	irrelevant	relevant	relevant
rel_k	1	0	0	1	1
P_k	1/1	1/2	1/3	2/4	3/5
$P_k * rel_k$	1	0	0	1/2	3/5
R_k	1/3	1/3	1/3	2/3	3/3
AP_k	1	1/2	1/3	0.375	0.42

The *Mean Average Precision* MAP_k is defined as the mean of the AP_k values over all queries q of the test collection:

$$MAP_k = \frac{1}{q} \sum_{i=1}^{q} AP_{i,k} \tag{5}$$

While the metrics Precision and Recall only provide insights into the ability to retrieve relevant documents, AP takes into account the ranking quality, emphasising the importance of ranking relevant documents higher in the list.

4. Results and Discussion

4.1. Dense Retrieval

The four domain-adapted dense retrievers (corpus-1a, corpus-1b, corpus-2, and corpus-3) were evaluated using the metrics defined in the previous chapter. We expect users to be most interested in search results within the first few result pages. Therefore, only the top 100 ($k \leq 100$) retrieved records (i.e., the first five result pages in the CKAN protype) were considered for the calculation of the evaluation metrics. The results were compared to the performance of the baseline model BM25 and the initial DistilBERT-base (the model before domain adaptation).

Comparing the MAP_{100} values (Table 4 in the right column), as a summarising evaluation metric considering all queries, the domain-adapted dense retrievers outperformed the baseline BM25 model. Further, all retrievers, including BM25, clearly outperformed the DistilBERT-base model. This emphasises the need for a domain adaptation of dense retrievers.

Table 4. AP_{100} and MAP_{100} values calculated for all retrievers and queries (the bold values mark the best performing model for each query).

	AP_{100}				
Retrieval Model	**Q1**	**Q2**	**Q3**	**Q4**	$\boldsymbol{MAP_{100}}$
Relevant Items	***1384***	***975***	***285***	***139***	
BM25	0.8726	0.8432	0.5760	0.9358	0.8069
DistilBERT-base	0.4216	0.1732	0.2237	0.2440	0.2656
corpus-1a	**1**	0.9670	0.9491	0.9180	**0.9585**
corpus-1b	0.9986	0.9603	**0.9894**	0.7973	0.9364
corpus-2	**1**	0.9677	0.8066	**0.9707**	0.9362
corpus-3	**1**	**0.9680**	0.7177	0.8917	0.8943

The AP_{100} values provide a more differentiated basis for the comparison of the four domain-adapted dense retrievers and their responses to the four test queries. As the AP_{100} values in Table 4 document, the domain-adapted dense retrievers showed a better ranking performance than BM25, except for query Q4 ("aircraft measurement") where BM25 performed better than three of the dense retrievers (corpus-1a, corpus-1b, and corpus-3). Further, one could expect that the dense retriever with the largest corpus (corpus-3: the corpus composition combining all contents) leads to the best results. This was not verified by the AP_{100} values. Corpus-3 showed particularly good results for Q1 and Q2, but here, all adapted dense retrievers performed very well with only minor differences. For Q3 and Q4, the corpus-3 dense retriever performed significantly worse than at least one of the other retrievers. This is a general finding verified by the AP_{100} values: the domain-adapted dense retrievers do not necessarily perform better when they are trained on a larger corpus. There is also no dense retriever that significantly outperforms the others. Each retriever has at least one query with weaker results:

- For Q3 ("observational data"), the dense retrievers fine-tuned with corpora including text passages from the scientific literature performed weaker.
- Corpus-1b (containing dataset descriptions and ontology concepts) showed a relatively weak performance for Q4 ("aircraft measurement").

A further evaluation regarding the *Precision* and *Recall* values revealed a similar trend (see Figure 3). Please note that as *Recall* is the ratio of retrieved relevant results to the total number of relevant documents, the slope of the *Recall* graphs in Figure 3 depends on the total number of possible relevant documents for each query. Consequently, the curves of different queries are not directly comparable. The *Recall* curves for Q1 and Q2 are less

steep due to significantly more relevant results than the other two queries. To improve readability, the y-axis of the *Recall* graphs was scaled differently for Q1 and Q2.

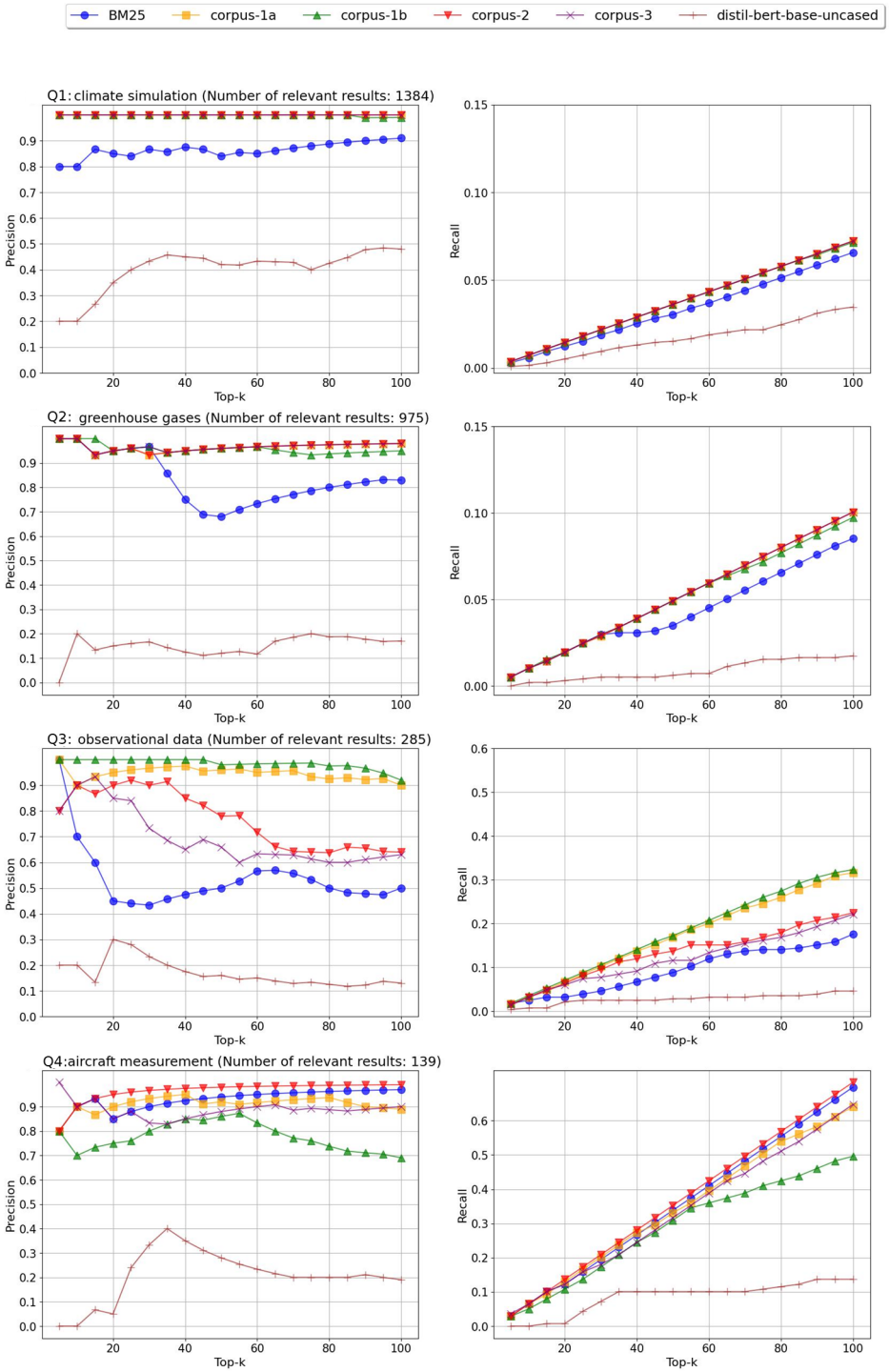

Figure 3. Comparison of the results of the dense retrievers for each test collection query. The left graphs show the *Precision* and the right shows the *Recall* values for varying top-k levels (5–100).

Due to the black box character of the domain-adapted dense retrievers, it is relatively difficult to uncover possible reasons for weak performances. In our analysis, we compared the top 50 hits of the best performing retriever with those of the weaker performer, specifically focusing on the additional False Positive documents. The identified False Positive documents were analysed for recognisable semantic patterns or conspicuous

features that could explain the retrieval errors. Further, the training corpus was analysed for query–document pairs that might have influenced the retrieval of the False Positives. However, we could not find robust evidence that could help to explain the differing results of the retrievers.

One relatively obvious source of error was that some retrievers did not fully comprehend the semantics of the provenance of the derived data. Specifically, for query Q3 ("observational data"), some retrievers (in particular corpus-2) could not differentiate the observational data from the data, which had been derived from the observational data. In the context of climate data, *observational data* refers to datasets sourced from atmospheric or oceanographic measurements, typically collected by meteorological stations. These observations can then be used to evaluate or calibrate climate models [73]. Such provenance linkages to data sources are commonly described in the abstracts of the data descriptions [74]. Compared to corpus-(1b), corpus-(2) did not provide sufficient knowledge for the retriever to distinguish between observational data and data derived from observational data.

We also tried to improve the performance of the model by adding specific text passages. For this purpose, the following experiment was conducted: The dense retriever based on corpus-(1b) performed weaker than the other dense retrievers at Q4 ("aircraft measurement") (see Table 4). Therefore, the corpus-(1b) was extended with a passage, providing context for the topic "aircraft measurement". We re-trained the model with the modified corpus-(1b) to test if an improvement for Q4 could be observed. Contrary to expectations, the opposite effect occurred, and the retrieval quality decreased. As illustrated in Figure 4, the retriever with the modified corpus-(1b) shows weaker Precision, especially for top-k levels below 20 compared to the original corpus-1b retriever. It is noteworthy that corpora-(2) and -(3) also contained this text passage but performed significantly better for Q4 than the modified version of corpus-1b (see Table 3 and Figure 3). Obviously, corpus-(2) and corpus-(3) also contained other relevant passages that could balance the negative effect.

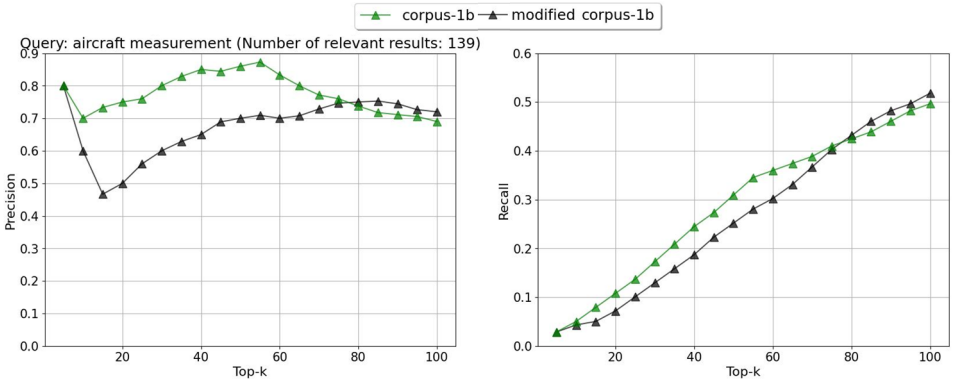

Figure 4. Testing a re-trained version of DistilBERT-base using the modified corpus-(1b). The graph shows the Precision and Recall values for Q4 ("aircraft measurement") for both models.

4.2. Spatial Re-Ranking

To test the spatial re-ranking, Q2 ("greenhouse gases") was extended by the search term "Italy" as a spatial context. Since not all records in the test collection had a spatial extent property (bounding box), a subset of 663 records was taken that included only records with this spatial property. As with the preceding experiment (see Section 4.1), the top 100 results were retrieved using a dense retriever. For this experiment, the model corpus-3 was selected as it has performed best on Q2 (c.f. Table 4). As described in Section 3.2, it could be expected that the overall search quality would decrease if the re-ranking was carried out with all hits returned in the retrieval stage. To evaluate this effect, we performed the re-ranking in this experiment once with all 100 search results and once with only the top 30 search results. The graphs in Figure 5 confirm this assumption with

significantly better values for Precision and Recall for re-ranking with the top 30 results (lower two graphs).

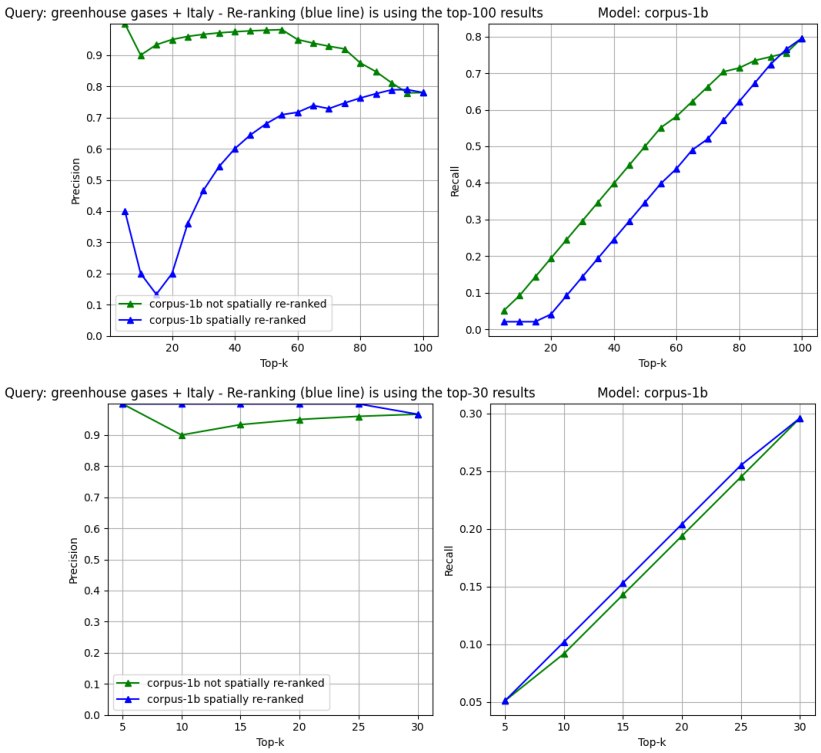

Figure 5. Spatial re-ranking effects on Precision and Recall for the "greenhouse gases" query (Q2) extended with "Italy" using corpus-1b.

In addition to improving thematic relevance, the spatial relevance of the search results significantly improved. After re-ranking the top 30 results from the retrieval phase, the top 10 final results were analysed. The average Hausdorff distance between the bounding boxes of the top 10 results decreased from about 156 (before re-ranking) to about 44. The maps in Figures 6 and 7 illustrate these findings.

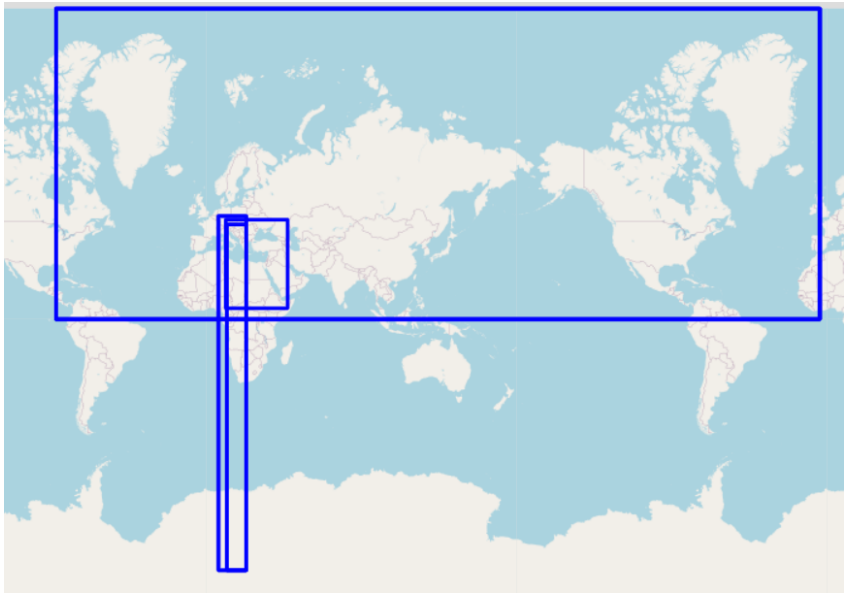

Figure 6. Spatial extent (blue boxes) of the top 10 search results before spatial re-ranking.

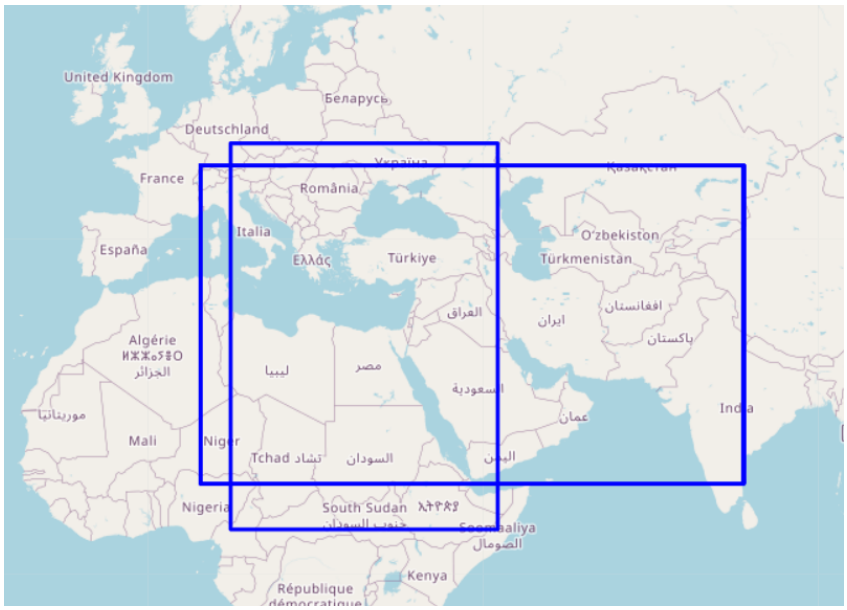

Figure 7. Spatial extent (blue boxes) of the top 10 search results after spatial re-ranking with the spatial extent of Italy.

5. Future Works

5.1. Extension and Improvement of the IR Approach

The defining attribute of spatial data is their inherently multi-dimensional nature, including temporal, spatial, and thematic dimensions. This is also reflected in the way users search [75]. The authors of ref. [29] describe users' search demands as queries for "concept at a location in time". The presented work aims to take the initial step of integrating a contextual search into metadata catalogues using dense retrieval. However, the current spatial re-ranking stage is effective but it can only serve as a provisional solution. As retrieval models evolve to better understand and integrate spatial and temporal contexts directly, re-ranking might no longer be required in the future. Moving towards models capable of natively identifying and interpreting spatiotemporal aspects in queries would reduce the reliance on external techniques like geocoding and named entity recognition (NER), which may introduce errors or fail to accurately extract the spatial scope from queries. Beyond detecting location context, we also identify additional elements of the spatiotemporal context that future retrieval models should be capable of interpreting:

(1) Scale and resolution: Ensuring alignment between the spatial scale of the query context and the document context is essential. For instance, if data for "Europe" are queried, metadata records on the continental scale and a corresponding resolution should be ranked best.
(2) Spatial relations: Some queries might contain relative spatial descriptions based on topological (e.g., within) or directional (e.g., north of) relations. This aspect has already been addressed in the study by [76].
(3) Temporal aspects: Geospatial and especially climate data often contain a temporal dimension, such as "past" (e.g., observation data) or "future" (e.g., climate projections), a certain time period (e.g., climate reference period, decades) or temporal relations (e.g., before). Dense retrievers could especially be fine-tuned with keywords that indicate time-related context.

As the results of the study demonstrated, the specific semantics of datasets and the relations among them are not sufficiently covered by the corpora so far. This could be addressed by adding specific ontologies like PROV-O [77] to include the data provenance relations in the corpora, for instance.

5.2. Improvement of the Evaluation

The necessity to establish an evaluation test collection led to the creation of a dataset with limitations in terms of its size and reliability. The manual annotation of a substantial number of documents is time-consuming and does not scale when undertaken by a limited number of annotators. The development of a larger test collection typically involves major campaigns, such as the Text REtrieval Conference (TREC) [78], in which experts develop relevant topics (queries) and make relevance judgments for query–document pairs (*qrels*) [79]. Although methods like "*pooling*" (c.f. [80]) can reduce the effort required to create the *qrels*, these test collections typically involve several months of work and a large team of experts. Moreover, the binary metric, which determines the relevance of a document as either query-positive or query-negative, further restricts the assessment. To obtain a more comprehensive understanding of the quality of the output ranking, a graded relevance metric like nDCG [81] could be used. Given these constraints, we decided to build a smaller test collection for this study while working in parallel on a larger TREC-compliant test collection in a separate effort [82]. A forthcoming study is planned to engage additional expert annotators providing graded relevance labels. The involvement of a greater number of annotators would also lead to a more extensive and diverse document collection, accompanied by an increased number of sample queries.

6. Summary and Conclusions

This study highlights the importance of exploring diverse IR approaches beyond relying solely on traditional BOW-based models to improve environmental data searches. In particular, the utilisation of dense retrieval through pre-trained transformer models is presented. The objective was to demonstrate a pragmatic methodology for preparing pre-trained models by employing existing domain adaptation techniques and subsequently integrating them into the search architecture of a standard metadata catalogue.

This work is groundbreaking since there is hardly any prior literature focusing on dense retrieval for research data searches or providing methods for designing suitable corpora for this task. The used domain adaptation method proposed by [24] requires an unlabelled text corpus with domain-specific vocabulary. The lack of benchmark corpora for vocabulary related to environmental research data made it necessary to create a custom corpus. Various sources, including public geospatial data portals, ontologies, and the scientific literature, were used to compile domain-specific corpora. The DistilBERT-base model was then domain-adapted using four distinct corpora.

The evaluation revealed the superior performance of the dense retrievers compared to the BM25 baseline and the original DistilBERT-base checkpoint. Notably, no significant differences between the different domain-adapted dense retrievers were observed during the evaluation. The domain-adapted dense retrievers do not necessarily perform better when they are trained on a larger corpus. Nevertheless, the identified limitations in the evaluation process require a more extensive assessment, involving a larger test collection and additional metrics such as nDCG. Generally, the dense retrievers performed well in retrieving thematically relevant documents. However, the retrievers did not fully comprehend the semantics of relations among datasets and the provenance of derived data. Further, the retrieval did not incorporate the spatio-temporal dimension of geospatial data and respective queries. To meet this requirement, which is inherent in an environmental data search, an approach was proposed that includes a spatial re-ranking stage following the initial retrieval stage. This allowed us to refine results based on geocoded entities and bounding boxes. While effective, this method is a provisional solution until future retrieval models can directly handle spatial and temporal contexts.

In conclusion, this study not only advances the understanding of effective IR approaches in the context of an environmental data search but also contributes a practical methodology for incorporating dense retrieval into existing metadata catalogues. The identified limitations pave the way for future research, emphasising the need for a more comprehensive evaluation and exploration of the spatio-temporal dimension.

Author Contributions: Conceptualisation, Simeon Wetzel and Stephan Mäs; methodology, Simeon Wetzel.; software, Simeon Wetzel; formal analysis, Simeon Wetzel; investigation, Simeon Wetzel; resources, Simeon Wetzel; writing—original draft preparation, Simeon Wetzel; writing—review and editing, Stephan Mäs; visualisation, Simeon Wetzel; supervision, Stephan Mäs; project administration, Simeon Wetzel; funding acquisition, Stephan Mäs. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the German Minsistry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF) as part of the funding initiative RegiKlim (Regional Information on Climate Action) grant number 01LR2005A.

Informed Consent Statement: Not applicable.

Data Availability Statement: The corpora presented in this study and the test collection are available on request from the corresponding author due to the copyright or licenses of the harvested metadata records. The domain-adapted dense retrieval model corpus-1a presented in this study is available in HuggingFace at https://huggingface.co/simeonw/distilbert-base-uncased-climate, accessed on 29 October 2024. The CKAN extension presented in this study is available in GitHub at https://github.com/simeonwetzel/ckanext-solr-vectorstore, accessed on 29 October 2024.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Hervey, T.; Lafia, S.; Kuhn, W. Search Facets and Ranking in Geospatial Dataset Search. In *11th International Conference on Geographic Information Science (GIScience 2021)—Part I. Leibniz International Proceedings in Informatics (LIPIcs)*; Schloss Dagstuhl—Leibniz-Zentrum für Informatik: Wadern, Germany, 2020; Volume 177, pp. 1–5. [CrossRef]
2. Robertson, S.; Zaragoza, H. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.* **2009**, *3*, 333–389. [CrossRef]
3. ElasticSearch. Available online: https://www.elastic.co/de/ (accessed on 29 October 2024).
4. Apache SOLR. Available online: https://solr.apache.org/ (accessed on 29 October 2024).
5. Thakur, N.; Reimers, N.; Rücklé, A.; Srivastava, A.; Gurevych, I. BEIR: A Heterogeneous Benchmark for Zero-Shot Evaluation of Information Retrieval Models. *arXiv* **2021**, arXiv:2104.08663.
6. Furnas, G.; Landauer, T.; Gomez, L.; Dumais, S. The Vocabulary Problem in Human-System Communication. *Commun. ACM* **1987**, *30*, 964–971. [CrossRef]
7. Lehmann, J.; Athanasiou, S.; Both, A.; Garcia Rojas, A.; Giannopoulos, G.; Hladky, D.; Le Grange, J.J.; Ngonga Ngomo, A.C.; Sherif, M.A.; Stadler, C.; et al. Managing Geospatial Linked Data in the GeoKnow Project. *Semant. Web Earth Space Sci. Curr. Status Future Dir.* **2015**, *20*, 51–78. [CrossRef]
8. Jiang, S.; Hagelien, T.F.; Natvig, M.; Li, J. Ontology-Based Semantic Search for Open Government Data. In Proceedings of the 13th IEEE International Conference on Semantic Computing, ICSC 2019, Newport Beach, CA, USA, 30 January–1 February 2019; pp. 7–15. [CrossRef]
9. Yue, P.; Guo, X.; Zhang, M.; Jiang, L.; Zhai, X. Linked Data and SDI: The Case on Web Geoprocessing Workflows. *ISPRS J. Photogramm. Remote Sens.* **2015**, *114*, 245–257. [CrossRef]
10. Geonetwork. Available online: https://geonetwork-opensource.org/ (accessed on 29 October 2024).
11. CKAN. Available online: https://ckan.org/ (accessed on 29 October 2024).
12. Chapman, A.; Simperl, E.; Koesten, L.; Konstantinidis, G.; Ibáñez, L.D.; Kacprzak, E.; Groth, P. Dataset Search: A Survey. *VLDB J.* **2020**, *29*, 251–272. [CrossRef]
13. *ISO19115*; Geographic Information—Metadata. ISO: Geneva, Switzerland, 2014. Available online: https://www.iso.org/standard/53798.html (accessed on 29 October 2024).
14. Dublin Core. Available online: https://www.dublincore.org/specifications/dublin-core/dces/ (accessed on 29 October 2024).
15. Wagner, M.; Henzen, C.; Müller-Pfefferkorn, R. A Research Data Infrastructure Component for the Automated Metadata and Data Quality Extraction to Foster the Provision of FAIR Data in Earth System Sciences. *AGILE GIScience Ser.* **2021**, *2*, 41. [CrossRef]
16. Schauppenlehner, T.; Muhar, A. Theoretical Availability Versus Practical Accessibility: The Critical Role of Metadata Management in Open Data Portals. *Sustainability* **2018**, *10*, 545. [CrossRef]
17. Quarati, A. Open Government Data: Usage Trends and Metadata Quality. *J. Inf. Sci.* **2021**, *49*, 887–910. [CrossRef]
18. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT), Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186.
19. Zhao, W.X.; Liu, J.; Ren, R.; Wen, J.-R. Dense Text Retrieval Based on Pretrained Language Models: A Survey. *ACM Trans. Inf. Syst.* **2023**, *42*, 89. [CrossRef]

20. Nakamura, T.A.; Calais, P.H.; Reis, D.C.; Lemos, A.P. An Anatomy for Neural Search Engines. *Inf. Sci.* **2019**, *480*, 339–353. [CrossRef]
21. Beltagy, I.; Lo, K.; Cohan, A. SCIBERT: A Pretrained Language Model for Scientific Text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3615–3620. [CrossRef]
22. Li, Z.; Kim, J.; Chiang, Y.-Y.; Chen, M. SpaBERT: A Pretrained Language Model from Geographic Data for Geo-Entity Representation. *arXiv* **2022**, arXiv:2210.12213.
23. Wang, K.; Thakur, N.; Reimers, N.; Gurevych, I. GPL: Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval. *arXiv* **2021**, arXiv:2112.07577.
24. Wang, K.; Reimers, N.; Gurevych, I. TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning. *arXiv* **2021**, arXiv:2104.06979.
25. Hu, Y.; Gao, S.; Lunga, D.; Li, W.; Newsam, S.; Bhaduri, B. GeoAI at ACM SIGSPATIAL. *SIGSPATIAL Spec.* **2019**, *11*, 5–15. [CrossRef]
26. Corcoran, P.; Spasić, I. Self-Supervised Representation Learning for Geographical Data—A Systematic Literature Review. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 64. [CrossRef]
27. Chen, Y.; Huang, W.; Zhao, K.; Jiang, Y.; Cong, G. Self-supervised Learning for Geospatial AI: A Survey. *arXiv* **2024**, arXiv:2408.12133.
28. Lacasta, J.; Lopez-Pellicer, F.J.; Espejo-García, B.; Nogueras-Iso, J.; Zarazaga-Soria, F.J. Aggregation-based information retrieval system for geospatial data catalogs. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 1583–1605. [CrossRef]
29. Lacasta, J.; Lopez-Pellicer, F.J.; Zarazaga-Soria, J.; Béjar, R.; Nogueras-Iso, J. Approaches for the Clustering of Geographic Metadata and the Automatic Detection of Quasi-Spatial Dataset Series. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 87. [CrossRef]
30. Chen, Z.; Song, J.; Yang, Y. Similarity measurement of metadata of geospatial data: An artificial neural network approach. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 90. [CrossRef]
31. Munir, K.; Sheraz Anjum, M. The use of ontologies for effective knowledge modelling and information retrieval. *Appl. Comput. Inform.* **2018**, *14*, 116–126. [CrossRef]
32. Asim, M.N.; Wasim, M.; Khan, M.U.G.; Mahmood, N.; Mahmood, W. The Use of Ontology in Retrieval: A Study on Textual, Multilingual, and Multimedia Retrieval. *IEEE Access* **2019**, *7*, 21662–21686. [CrossRef]
33. Noy, N.; Burgess, M.; Brickley, D. Google Dataset Search: Building a Search Engine for Datasets in an Open Web Ecosystem. In Proceedings of the World Wide Web Conference (WWW) 2019, Francisco, CA, USA, 13–17 May 2019; pp. 1365–1375. [CrossRef]
34. Zrhal, M.; Bucher, B.; Hamdi, F.; Van Damme, M.D. Identifying the Key Resources and Missing Elements to Build a Knowledge Graph Dedicated to Spatial Dataset Search. *Procedia Comput. Sci.* **2022**, *207*, 2911–2920. [CrossRef]
35. Glocker, K.; Knurr, A.; Dieter, J.; Dominick, F.; Forche, M.; Koch, C.; Pascoe Pérez, A.; Roth, B.; Ückert, F. Optimizing a Query by Transformation and Expansion. *Stud. Health Technol. Inform.* **2017**, *243*, 197–201. [CrossRef] [PubMed]
36. Mai, G.; Janowicz, K.; Prasad, S.; Shi, M.; Cai, L.; Zhu, R.; Regalia, B.; Lao, N. Semantically-Enriched Search Engine for Geoportals: A Case Study with ArcGIS Online. *AGILE GIScience Ser.* **2020**, *1*, 13. [CrossRef]
37. Sun, K.; Zhu, Y.; Pan, P.; Hou, Z.; Wang, D.; Li, W.; Song, J. Geospatial Data Ontology: The Semantic Foundation of Geospatial Data Integration and Sharing. *Big Earth Data* **2019**, *3*, 269–296. [CrossRef]
38. Esteva, A.; Kale, A.; Paulus, R.; Hashimoto, K.; Yin, W.; Radev, D.; Socher, R. COVID-19 Information Retrieval with Deep-Learning Based Semantic Search, Question Answering, and Abstractive Summarization. *Npj Digit. Med.* **2021**, *4*, 68. [CrossRef]
39. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 3982–3992. [CrossRef]
40. Coelho, J.; Magalhães, J.; Martins, B. Improving Neural Models for the Retrieval of Relevant Passages to Geographical Queries. In Proceedings of the ACM International Symposium on Advances in Geographic Information Systems, Beijing, China, 2–5 November 2021; pp. 268–277. [CrossRef]
41. MS MARCO. Available online: https://microsoft.github.io/msmarco/ (accessed on 29 October 2024).
42. Gao, Y.; Xiong, Y.; Wang, S.; Wang, H. GeoBERT: Pre-Training Geospatial Representation Learning on Point-of-Interest. *Appl. Sci.* **2022**, *12*, 12942. [CrossRef]
43. Mai, G.; Janowicz, K.; Hu, Y.; Gao, S.; Yan, B.; Zhu, R.; Cai, L.; Lao, N. A Review of Location Encoding for GeoAI: Methods and Applications. *Int. J. Geogr. Inf. Sci.* **2022**, *36*, 639–673. [CrossRef]
44. Syed, M.A.; Arsevska, E.; Roche, M.; Teisseire, M. GeospatRE: Extraction and Geocoding of Spatial Relation Entities in Textual Documents. *Cartogr. Geogr. Inf. Sci.* **2023**, 1–16. [CrossRef]
45. EEA Geospatial Data Catalogue. Available online: https://sdi.eea.europa.eu/catalogue/srv/eng/catalog.search#/home (accessed on 29 October 2024).
46. United Nations FAO Map Catalogue. Available online: https://data.apps.fao.org/map/catalog/srv/ger/catalog.search#/home (accessed on 29 October 2024).
47. Copernicus Data Store. Available online: https://cds.climate.copernicus.eu/#!/home (accessed on 29 October 2024).
48. data.europe.eu. Available online: https://data.europa.eu/en (accessed on 29 October 2024).
49. GEMET. Available online: https://www.eionet.europa.eu/gemet/en/themes/ (accessed on 29 October 2024).

50. GEMET Theme Climate. Available online: http://www.eionet.europa.eu/gemet/theme/7 (accessed on 29 October 2024).
51. GEMET Theme Natural Dynamics. Available online: http://www.eionet.europa.eu/gemet/theme/8 (accessed on 29 October 2024).
52. GEMET Atmosphere (Air, Climate). Available online: http://www.eionet.europa.eu/gemet/group/618 (accessed on 29 October 2024).
53. Copernicus Open Access Journals. Available online: https://publications.copernicus.org/open-access_journals/journals_by_subject.html (accessed on 29 October 2024).
54. Kotamarthi, R.; Hayhoe, K.; Mearns, L.; Wuebbles, D.; Jacobs, J.; Jurado, J. *Downscaling Techniques for High-Resolution Climate Projections: From Global Change to Local Impacts*; Cambridge University Press: Cambridge, UK; New York, NY, USA, 2021. [CrossRef]
55. Spiridonov, V.; Curic, M. *Fundamentals of Meteorology*; Springer: Cham, Switzerland, 2020; pp. 1–437. [CrossRef]
56. Nogueira, R.; Yang, W.; Lin, J.; Cho, K. Document Expansion by Query Prediction. *arXiv* **2019**, arXiv:1904.08375.
57. Cross-Encoder ms-marco-MiniLM-L-6-v2. Available online: https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-6-v2 (accessed on 29 October 2024).
58. Hofstätter, S.; Althammer, S.; Schröder, M.; Sertkan, M.; Hanbury, A. Improving Efficient Neural Ranking Models with Cross-Architecture Knowledge Distillation. *arXiv* **2020**, arXiv:2010.02666.
59. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
60. KeyBERT by Maarten Grootendorst. Available online: https://github.com/MaartenGr/KeyBERT (accessed on 29 October 2024).
61. BERT NER Model (dslim/bert-base-NER-uncased). Available online: https://huggingface.co/dslim/bert-base-NER-uncased (accessed on 29 October 2024).
62. Photon Geocoding API by Komoot. Available online: https://photon.komoot.io/ (accessed on 29 October 2024).
63. Degbelo, A.; Teka, B.B. Spatial Search Strategies for Open Government Data: A Systematic Comparison. In Proceedings of the 13th Workshop on Geographic Information Retrieval, Lyon, France, 28–29 November 2019. [CrossRef]
64. SOLR Vector Scoring. Available online: https://github.com/saaay71/solr-vector-scoring (accessed on 29 October 2024).
65. FastAPI. Available online: https://fastapi.tiangolo.com/ (accessed on 29 October 2024).
66. CKAN Solr VectorStore Extension. Available online: https://github.com/simeonwetzel/ckanext-solr-vectorstore (accessed on 29 October 2024).
67. NIST COVID-19 Track. Available online: https://ir.nist.gov/covidSubmit/index.html (accessed on 29 October 2024).
68. BioASQ. Available online: http://bioasq.org/ (accessed on 29 October 2024).
69. WDC Climate Data Center. Available online: https://www.wdc-climate.de/ui (accessed on 29 October 2024).
70. Derczynski, L. Complementarity, F-score, and NLP evaluation. In Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016, Portorož, Slovenia, 23–28 May 2016; pp. 261–266, ISBN 9782951740891.
71. Zhu, M. *Recall, Precision and Average Precision*; Department of Statistics and Actuarial Science, University of Waterloo: Waterloo, ON, Canada, 2004; pp. 1–11.
72. Robertson, S. A new interpretation of average precision. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, 20–24 July 2008; pp 689–690. [CrossRef]
73. Zumwald, M.; Knüsel, B.; Baumberger, C.; Hirsch Hadorn, G.; Bresch, D.N.; Knutti, R. Understanding and assessing uncertainty of observational climate datasets for model evaluation using ensembles. *Wiley Interdiscip. Rev. Clim. Chang.* **2020**, *11*, e654. [CrossRef]
74. Henzen, C.; Mäs, S.; Bernard, L. Provenance information in geodata infrastructures. In *Lecture Notes in Geoinformation and Cartography*; Springer: Cham, Switzerland, 2013; pp. 133–151. [CrossRef]
75. Jiang, Y.; Li, Y.; Yang, C.; Hu, F.; Armstrong, E.M.; Huang, T.; Moroni, D.; McGibbney, L.J.; Finch, C.J. Towards intelligent geospatial data discovery: A machine learning framework for search ranking. *Int. J. Digit. Earth* **2018**, *11*, 956–971. [CrossRef]
76. Shin, H.; Park, J.;Yuk, D.; Lee, J. BERT-based Spatial Information Extraction. In Proceedings of the Third International Workshop On Spatial Language Understanding (SpLU 2020), Virtual, 19 November 2020; Volume 8, pp. 10–17.
77. PROV-O W3C Recommendation. Available online: https://www.w3.org/TR/prov-o/ (accessed on 29 October 2024).
78. Text REtrieval Conference (TREC) by NIST. Available online: https://trec.nist.gov/ (accessed on 29 October 2024).
79. Sanderson, M. Test collection based evaluation of information retrieval systems. *Found. Trends Inf. Retr.* **2010**, *4*, 247–375. [CrossRef]
80. Buckley, C.; Dimmick, D.; Soboroff, I.; Voorhees, E. Bias and the limits of pooling for large collections. *Inf. Retr.* **2007**, *10*, 491–508. [CrossRef]
81. Wang, Y.; Wang, L.; Li, Y.; He, D.; Chen, W.; Liu, T.Y. A theoretical analysis of NDCG ranking measures. *J. Mach. Learn. Res.* **2013**, *30*, 25–54.
82. CCTC. Available online: https://github.com/simeonwetzel/CCTC (accessed on 29 October 2024).