

Dataset Extraction and Information Retrieval for Historical Conflicts

Dinis Sousa
FEUP

Porto, Portugal
up202006303@edu.fe.up.pt

Gabriel Ferreira
FEUP

Porto, Portugal
up201906072@edu.fe.up.pt

João Matos
FEUP

Porto, Portugal
up202006280@edu.fe.up.pt

João Pinheiro
FEUP

Porto, Portugal
up202008133@edu.fe.up.pt

ABSTRACT

This paper describes a Historical Conflicts Search Engine project that aims to provide a comprehensive platform for exploring and comprehending historical events. Our key goals include constructing a coherent dataset and developing a user-friendly search engine, and we are focused on obtaining, analysing, and displaying material connected to significant historical battles. This study's initial steps prioritise dataset extraction and refinement, showing crucial findings such as concentrated events in Europe, dataset recency bias, and repeating themes. The following phase looks into the technical components of information retrieval, making use of Apache Solr and Extended DisMax query processing. Precision-recall curves, precision at 10, and average precision are all evaluation criteria. The project narrative is completed with the focuses on the creation of the user interface and enhanced search capabilities, as semantic search and learn to rank. This stage improves access to historical data by giving users a simple and strong tool for investigating historical conflicts. Collectively, our work contributes to the development of an advanced Historical Conflicts Search Engine, bridging the gap between historical data and user accessibility.

CCS CONCEPTS

• **Information systems** → **Web searching and information discovery**; *Data management systems*; *Information retrieval query processing*.

KEYWORDS

historical conflicts, historical events, search engine, dataset, data preparation, data analysis, information, retrieval, processing

1 INTRODUCTION

Navigating the wealth of historical data, especially related to conflicts, is a challenge for researchers and enthusiasts. This challenge is accentuated in the realm of Information Retrieval (IR), where efficiently extracting and presenting relevant historical information is crucial.

The Historical Conflicts Search Engine project addresses this challenge by leveraging advanced IR techniques to create an intuitive platform for retrieving and exploring historical events. We opted to use Wikidata [7] as our primary data source, focusing on historical events that are defined as "particular incident in history that brings about a historical change" [8].

In the context of IR, our task involves extracting structured data and abstracting rich-text information to provide nuanced insights into historical conflicts. This paper marks the first and second phases, focusing on the extraction, preparation and refinement of a dataset followed by an exploration of querying techniques and

results. Our ultimate goal is to contribute to IR by providing a specialized tool that facilitates meaningful exploration of historical conflicts.

Milestone I

Milestone I

2 DATA SELECTION

Our initial step is to choose the datasets we will use. The only requirement was that the selected data should have both structured data and unstructured data.

We decided to explore datasets on wars and conflicts.

Some datasets, such as the GDELT (The GDELT Project) [2] or the UPPSALA Conflict Data Program (UCDP) [6], had too many entries and little to no rich text to explore.

We settled with using data from Wikidata, selecting the wanted data using the Wikidata Query Service, and using Wikipedia[9] articles to get the rich text.

After analyzing the different Wikidata parameters, we settled on using the following query to retrieve the main dataset used (containing 9134 results):

```
1 SELECT ?event (SAMPLE(?date_) as ?date) ?label (SAMPLE(?
2 image_) as ?image) ?article WHERE {
3
4 ?event (wdt:P31/(wdt:P279*)) wd:Q13418847;
5 wdt:P585 ?date_.
6 ?event rdfs:label ?label.
7 FILTER((LANG(?label)) = "en")
8 OPTIONAL {?event wdt:P18 ?image_.}
9 ?article schema:about ?event;
10 schema:isPartOf <https://en.wikipedia.org/>.
11 }
12 GROUP BY ?event ?label ?article
13 LIMIT 10000
```

Listing 1: SPARQL query to retrieve historical events

This dataset has the event id, the event date, a label (name of the event), an image and the URL for the Wikipedia page.

The dataset is in a .json file format.

3 DATA PROCESSING

In this section, we will describe the steps to collect the data and then the data refinement process.

3.1 Data Collection

With the primary dataset in hand, we completed and treated the information as described in the annexed pipeline (Figure 10).

Iterating through the events dataset, we complete each event's information with: 1) a Wikipedia summary, and 2) other statements that Wikimedia had on the event.

Other statements include: list of participants, location, coordinate location, part of (if the historical event is part of a bigger event, a battle is part of a war), time period, number of deaths, number of injured, etc...

After filtering out events that had no summary we ended up with 8998 events, mainly armed conflicts.

3.2 Data Refinement

The process of refining our historical conflicts dataset involves a systematic approach to ensure the quality and relevance of the information. After the initial extraction, we embarked on a multi-step refinement process to enhance the usability and coherence of the dataset.

3.2.1 Completeness Analysis. To address missing or incomplete information, we conducted a thorough analysis of the completeness of the dataset. Columns with a high percentage of missing values (more than 85%) were excluded to streamline the dataset and improve its overall quality. This step ensures that the retained information is robust and contributes meaningfully to the search engine's capabilities.

3.2.2 Attribute Selection. To further refine the dataset, we selected attributes based on their relevance and completeness. Attributes with substantial missing values were excluded, ensuring that the retained information aligns with the project's objectives. This process not only improves data quality but also streamlines the dataset for a better user experience in the search engine.

Everything described up to this point is represented in the data pipeline, shown in Figure 10.

3.3 Conceptual Model

In our conceptual model (see Figure 1), we define several key entities that form the foundation of our historical conflicts dataset:

- **Complex Conflict:** This entity represents a bigger conflict, such as a war. A Complex Conflict exists as a standalone entity and incorporates various Historical Conflicts that are part of it. It serves as a high-level container for historical conflict data.
- **Historical Conflict:** Historical Conflicts are components of Complex Conflicts. These represent individual conflicts within the broader context of a Complex Conflict. Historical Conflicts provide detailed information about specific events, including their participants, locations, and dates.
- **Place:** The Place entity represents various location types, including countries, regions, cities, rivers, and more. It serves as a critical element for specifying where historical conflicts occurred. Each historical conflict is associated with a Place to indicate its geographical context.
- **Participant:** Participants in historical conflicts are classified into different types, including individuals, groups of people, and organizations. This entity helps categorize and identify the key actors involved in Historical Conflicts, providing insights into the composition of these events.

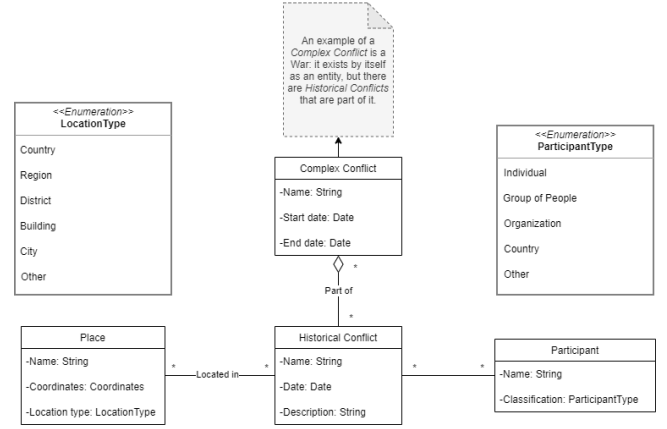


Figure 1: Conceptual Model

The conceptual model defines the fundamental entities and their relationships within our historical event dataset, enabling a structured and comprehensive representation of historical conflicts and their contextual information.

4 DATA CHARACTERIZATION

After collecting and processing all the necessary data, our next step was to conduct a comprehensive analysis of the data.

4.1 Descriptive Analysis

We assessed the dataset for specific characteristics such as the number of events with available images and the number of events that included details about participants. These statistics are crucial for understanding the richness and completeness of our dataset.

Our analysis revealed that out of the total events, 5664 included images, providing visual context for a significant portion of the dataset. Additionally, 3240 events contained information about participants, indicating the extent to which historical records specify the key actors in these events.

4.2 Spatial Distribution Analysis

To gain insights into the geographic distribution of these events, we utilized an interactive map (see Figure 2) to visualize their locations on a map. The geographical coordinates of each event were extracted using the "coordinate location" of each entry, and we employed the Python module 'folium' [1] to plot these events on a world map.

Our initial hypothesis was the distribution of historical events would exhibit an asymmetric pattern, with a concentration of events in Europe. This hypothesis was based on historical records and expectations regarding the frequency of events across different regions.

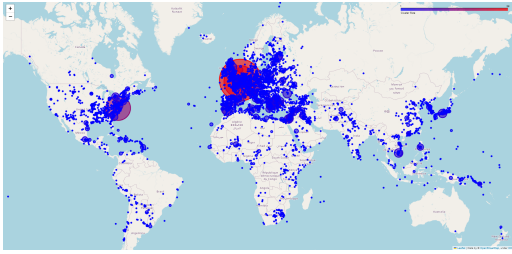


Figure 2: Geographic Distribution of Historical Conflicts

The visualization on the interactive map allowed us to confirm and quantify this hypothesis, providing a clear representation of the distribution of historical events across the world.

4.3 Temporal Distribution Analysis

To understand where in time these events happened, we explored in which centuries the events occurred. Our dataset clearly has a recency bias, with a lot more conflicts reported in the last 5 centuries, peaking in the 20th century. This was expected because the information on more recent events is more well-documented.

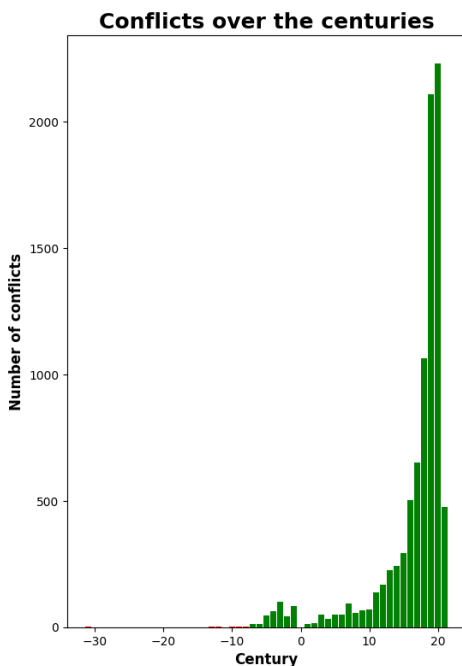


Figure 3: Temporal Distribution of Historical Conflicts

4.4 Analysis of Word Clouds

To acquire a better understanding of the key terms and themes within our dataset of historical events, we generated a word cloud

the summaries of the events. These word clouds visually represent the frequency of words, with larger words indicating higher frequency. Our analysis reveals essential patterns and recurrent themes in the dataset.

4.4.1 Historical Conflict Summaries Word Cloud, see Figure 4. The word cloud for historical event summaries offers a deeper insight into the content of these events. Key takeaways from the summary word cloud analysis include:

- The word 'Battle' remains a dominant term, reflecting its significance in historical narratives.
- The summary word cloud exhibits a broad diversity of terms, including European countries, types of locations, and terms within the same word family as 'war', indicating a range of themes and historical contexts.



Figure 4: Word Cloud for Historical Conflict Summaries

The word cloud offers valuable visual representations of the most frequent terms in event summaries, highlighting key themes and recurring words. The prominence of 'Battle' underscores its historical significance. The summary word cloud showcases the diversity of terms surrounding historical events, providing a nuanced understanding of the dataset.

The spatial distribution analysis we have seen before, along with the characterization of dataset attributes, forms the foundation for our search engine’s capabilities in providing users with a rich and diverse set of historical information. It enables users to explore and understand the geographical and contextual aspects of historical events in an interactive and informative manner.

5 INFORMATION NEEDS

In order to develop a search engine for historical events, it is essential to identify the information needs of potential users. Here, we outline four specific information needs related to historical conflicts, each of which corresponds to a different analytical aspect of the search engine's functionality:

- (1) **Battles by a River (1700-1775):** Users may seek information about battles that took place near a river between 1700 and 1775. This analytical query allows users to explore the geographical and temporal aspects of conflicts associated with rivers during this period.
- (2) **Destructive Battles in Europe (World War I):** Users interested in the impact and scale of warfare during World War I may want to know about the most destructive battles in Europe. This analytical query provides insights into the intensity and consequences of conflicts in the European theater during this historic period.
- (3) **Portuguese Participation in Battles (1300-1800, Non-Main Participant):** Users focused on Portuguese military history may inquire about battles in which Portugal participated between 1300 and 1800 where Portugal was not the main participant. This analytical query allows users to explore Portugal's involvement in conflicts from a nuanced perspective.
- (4) **Battles that are part of Revolutions that have Economic Consequences:** Users interested in the economic impact of revolutions may want to identify battles that had several economic consequences. This analytical query enables users to explore the intersections between military conflicts and their broader economic repercussions.

To address these information needs, our search engine utilizes the dataset sourced from Wikidata and Wikipedia, which contains structured data, rich text, and images related to historical events. The data has been processed and organized to facilitate effective Information Retrieval.

6 SEARCH DOCUMENT

In our search result document, we present a comprehensive overview of a historical conflict, including the following elements:

- **Event Name:** The title or name of the historical conflict, provides a concise identifier for the event.
- **Summary:** A brief summary or description of the event, providing key details and context. The summary may be truncated for readability.
- **Image:** An image or visual representation associated with the historical event, offering visual context and insights.
- **Map:** A geographical map displaying the location of the event. The map helps users understand where the event took place and its spatial context. Location information may include country, region, district, or city.
- **Timeline:** A timeline indicating the date and period during which the historical event occurred. It provides a temporal context, allowing users to understand when the event took place in history.

The search result document combines these elements to provide users with a comprehensive and informative representation of the historical event they are interested in. Users can gain insights into the event's name, summary, visual aspects, location, and temporal context, facilitating a deeper understanding of the historical narrative.

Milestone II

Milestone II

7 INFORMATION RETRIEVAL

The field of Information Retrieval involves handling the organization, storage, analysis, searching, and retrieval of information. The objective is to create a system that meets these requirements and includes an analysis of the developed system. Like any other Information Retrieval system, the primary aim is to retrieve all items relevant to a given information need while minimizing the retrieval of non-relevant items. The system will specifically target a common retrieval task known as "ad-hoc search", which entails returning information resources related to a user's information need translated into a natural text query. The focus is on developing a system that efficiently addresses this specific retrieval task.

7.1 Information Retrieval Tool

After an initial search, we chose the Information Retrieval tool for this project, Apache Solr [5]. This tool is open-source and based on Lucene[3] library tools, offering a great variety of features, such as advanced full-text Search, near real-time indexing, high availability, and flexible and adaptable configuration.

8 INDEXING AND SCHEMA

Indexing is one of the most important tasks when setting up a search system for Information Retrieval. We will explore the different attributes, and point out which ones we selected as being relevant for indexing.

8.1 Custom field type

Solr gives some types as default. We used some of these default types, such as pdate (for dates), pint (for integers), text_en (for text in English), string, and location (for coordinates). However, we felt the need to create other types. We created the 'caseInsensitiveString' type for strings we wanted to perform searches on to be case insensitive, 'urlString' for the event URL (so a person can find the conflict by the URL or a part of a URL also), and the 'richText' for the richest intensive attributes, so we could extract the most information out of them for that reason we used filters to ignore stopwords, analyze synonyms and get only the minimal stem of words for better matching with the queries. In Table 1 we can see the filters applied to each one of the types.

Type	Filters
caseInsensitiveString	ASCIIFoldingFilterFactory LowerCaseFilterFactory
urlString	ASCIIFoldingFilterFactory LowerCaseFilterFactory
richText	StopFilterFactory ASCIIFoldingFilterFactory EnglishMinimalStemFilterFactory LowerCaseFilterFactory SynonymGraphFilterFactory

Table 1: Filters applied for each one of the custom types

All these three types are instances of the Solr default 'TextField' class, having both 'caseInsensitiveString' and 'richText' Solr's Standard tokenizer and the 'urlString' type the Path hierarchy tokenizer, as the URL is divided by '/'.

8.2 Indexing

Table 6 displays the attributes of Conflicts, indicating which ones are indexed.

We indexed the 'summary' and 'label', which are the fields that contain the bulk of the text used for rich text analysis. 'event' is indexed so that a user can search for the event URL. 'date' is included for ordering and filtering purposes. The 'participants' and 'participant_count', to search, boost, and filter by participants and their number. 'country', 'location', and 'coordinate_location' for location-based searches. 'part_of' and 'instance_of', so users can search for broader conflict categories such as wars, conflicts, and invasions. The other attributes can be relevant to present as a search result, as they complement the information retrieved, but not as a search index. A comprehensive summary of the types of attributes and whether they are indexed or not can be found in Table 6.

9 RETRIEVAL PROCESS

The retrieval process is a critical phase in the Information Retrieval system, reflecting upon the execution of queries and the extraction of relevant documents. In the context of our system using Apache Solr, the retrieval process unfolds in two primary stages: query parsing and parameter selection.

9.1 Query Parsing

We used the Extended DisMax (eDisMax) query parser, which is an extended version of the DisMax parser. This parser adds relevant parameters compared to Solr's Standard query parser, providing better overall support.

9.2 Parameter Selection

We experimented with different query functionalities to get familiarized and better decide what parameters to produce queries for our information needs. Here are some examples:

- (1) `qf:label^2` summary is useful to give more weight to the label fields;
- (2) `bq:summary:musket^3` summary:weapon gives preference to the word "musket" in the "summary" field;
- (3) `bq:helicopter^20` boost the word "helicopter" by a factor of 20 in all fields;
- (4) `fq:!geofilt & pt:40.7128,-74.0060 & sfield:coordinate_location & d:10`. This can be used to retrieve conflicts that occurred within a distance of 10km from New York City;
- (5) `q:"battle located"` 5 can match, for example, both "this battle was located" and "the battle of Waterloo was located";
- (6) `q:bomb*`. To allow the search of "bombardment(s)", "bombing", "bombarded", etc.;
- (7) `q:located 2`. To allow the search of "location".

These queries and parameter configurations were designed to explore the capabilities of the search engine and to optimize the retrieval of information for specific scenarios.

The parameters we used in our search were the following (description from the Solr documentation page [4])

- `q` (query) - The `q` parameter defines the main "query" constituting the essence of the search. The parameter supports raw input strings provided by users with no special escaping.
- `op` (query operator) - Specifies the default operator for query expressions, overriding the default operator specified in the Schema. Possible values are "AND" or "OR".
- `fq` (filter query) - Apply a filter query to the search results.
- `fl` (field list) - Limits the information included in a query response to a specified list of fields.
- `df` (default field) - Specifies a default field, overriding the definition of a default field in the Schema.
- `qf` (query fields) - Specifies the fields in the index on which to perform the query. If absent, defaults to `df`.
- `bq` (boost query) - Specifies a factor by which a term or phrase should be "boosted" in importance when considering a match.

These parameters collectively contribute to the precision and recall of our search results, allowing us to tailor the search engine behavior to our specific use case.

9.3 Systems Setup

To test and, in the next step, evaluate how we can retrieve documents we set up two systems: a "base" system and a "boosted" system. Both use the same schema already described, however, in the boosted system we give weights to certain fields across queries and even terms. In a real implementation of our systems, there would be the need for some preprocessing of the search input so as to be able to identify dates and locations to adequately filter by the relevant fields. Usually, this is done explicitly by the user in filter fields. More than that, to attribute weights to certain words, the boosted system would need to identify keywords or score each word of the search input. The choice of using the same schema for both systems is meant to highlight the significance of boosting fields and queries, as we can reduce our comparisons to those aspects. Not only that but initial testing with a more minimalist schema without as many filters and custom types produced unsatisfactory results.

10 EVALUATION

Having set up and adapted the retrieval process, we now ought to choose different queries and evaluate their usefulness in fulfilling our information needs. The evaluation of the Information Retrieval system is crucial to assess its performance and effectiveness in meeting user requirements.

In order to get a set of relevant documents per query to be performed, we analyzed the different documents aided by a Python script. The program performs basic filters by "date", "location", "part_of" or other fields and then it searches for keywords that we considered to be of the greatest relevance to each specific query. The program would then display the parts of the text containing the used keywords and other attributes for each fetched document so that we could accept or reject them. In the end, we had a file for each query with the IDs of relevant conflicts and the portion of the summary containing a relevant keyword, for future reference. The

sets of relevant documents are essential to compute the evaluation metrics used to measure performance.

10.1 Evaluation Metrics

Before jumping into specific queries and parameter configurations, it is imperative to outline the metrics that will be employed when evaluating the performance of our Information Retrieval system. The chosen metrics provide a comprehensive understanding of the system's precision, recall, and overall effectiveness.

10.1.1 Precision-Recall (P-R) Curves. Precision-Recall curves are instrumental in visualizing the trade-off between precision and recall at different decision thresholds. These curves plot precision on the y-axis and recall on the x-axis, offering insights into the system's ability to retrieve relevant documents while minimizing false positives.

10.1.2 Precision at K (P@K). Precision at K, where K represents a specific rank or position in the result set, provides a focused measure of precision. In our evaluation, P@10 will be employed, indicating the precision of the top 10 results. This metric is particularly relevant for scenarios where users often focus on the initial set of retrieved documents.

10.1.3 Average Precision. Average Precision provides a summary measure of precision across different recall levels. It considers precision at each point where a relevant document is retrieved, offering a comprehensive assessment of the system's overall precision across the entire result set.

10.2 Queries to be evaluated

With these evaluation metrics in mind, we will now introduce specific queries and system configurations to assess the Information Retrieval system's performance. The queries will be executed under two scenarios: the simple system and the boosted system. The simple system represents the baseline configuration, while the boosted system incorporates refined parameters, such as boosted queries or additional filters, to enhance retrieval performance.

We have one query for each of our four information needs, already defined in the [information needs section](#). From each of them, we extracted keywords to formulate a query in Solr. Adding to those, every query in both systems has "**df**: summary", and the same fields are being queried: "**qf**: summary participants^2 location^2 label^2" in the case of the boosted system and "**qf**: summary participants location label" in the base system. These fields are the most relevant and diverse and the weights in the boosted system are meant to increase the probability of what is being queried appearing as relevant in the retrieved documents, as many terms that are present in the summary can just appear as a brief mention.

10.2.1 Battles by a River (1700-1775). **Possible search input:** "Which battles took place by a **river** during the period from **1700 to 1775**?"

In this query, "battles" is unspecific so we want to retrieve conflicts that have their date of occurrence inside of the given interval of 1700 to 1755 and have "river" explicit in their location or title. However, mentions of "river" in the summary may still mean the

battle took place near a river in spite of the two other fields. The previously mentioned query fields (qf) take that into account already. Our query is done in Solr using the following parameters:

- q: river
- fq: date:[1700-01-01T00:00:00Z TO 1775-12-31T23:59:00Z]

10.2.2 Destructive Battles in Europe (World War I): **Possible search input:** "Which were the most **destructive** battles in **Europe** related to **World War I**?"

With this query, we want to retrieve battles with destructive impact so we will search for "destruction", some synonyms and related terms. We also want to filter results that occurred during the period of the First World War so we can filter by date and, to be sure, events that explicitly mention the war. There is only interest in conflicts that took place on European soil so filter for "Europe" or the names of European countries. In the end we have these parameters:

- q: destruction ruins bomb* devastat* destroy* damaging
- fq: date:[1914-07-28T00:00:00Z TO 1918-11-11T23:59:00Z]
part_of:("war 1" OR "war I" OR "great war" OR "world war")
OR summary:("war 1" OR "war I" OR "great war" OR "world war")
summary:(europe italy germany france britain united kingdom belgium poland austria hungary russia)
location: (europe italy germany france britain united kingdom belgium poland austria hungary russia)
country: (europe italy germany france britain united kingdom belgium poland austria hungary russia)
- (Boosted system only) bq: fort^3 bombard^5 bridge^5 city^5 terrain^4 village^5 devastated^7 ruins^7 severely^2 enormous^2

The boosted terms (bq) are a way of giving more relevance to words related to a greater level of destruction, e.g. "bombardment", as well as targets for the attacks, e.g. "bridge". The relative weights reflect how infrequent we considered those words to be unrelated to events of great destruction.

10.2.3 Portuguese Participation in Battles (1300-1800, Non- Main Participant): **Possible search input:** "In which conflicts did the **Portuguese** participate **between 1300 and 1800** where **Portugal** was not a main participant?"

This query is meant to retrieve conflicts in which Portugal or the Portuguese participated in the period of the 14th to the 18th century. We can then query "Portugal" and filter by the range of dates provided. We do not force the field "participants" in a filter query since it is not present in most documents.

- q: Portugal
- fq: date:[1300-01-01T00:00:00Z TO 1799-12-31T23:59:00Z]
- (Boosted system only) bq: allied 3^3
participants_count:[3 TO *]^7

The usage of "allied 3" should also catch "ally" and "allies", which together with boosting conflicts that had many participants, can be a sign of Portugal not being a main contender.

10.2.4 Battles that are part of Revolutions that have Economic Consequences: **Possible search input:** "What were the confrontations

that had huge **economic** consequences and were part of a **revolution**?"

In this query, we intend to retrieve documents with mentions of the economy, directly by querying "economy" or indirectly, since synonyms and related terms were included as part of one indexing filter. The only relevant results are the ones part of a revolution so that filter will also be applied. We can translate these intentions to parameters in Solr:

- q: economy
- fq: label:revolution*summary:revolution*part_of:revolution*
- (Boosted system only) bq: part_of:Revolution^3
instance_of:revolution^3
label:Revolution^3 summary: consequence

Equal weights are applied since we are trying to find the same information in all of them. The presence of "revolution" in these fields almost certainly means the retrieved conflict is a revolution and we want to find those first.

10.3 Metrics and P-R curves

In this subsection, we will outline the evaluation process conducted to assess the performance of our Information Retrieval system under different configurations. The evaluation process involves executing specific queries and analyzing the retrieved results computing predefined metrics.

10.3.1 Battles by a River (1700-1775). For this query, 59 documents were retrieved, 37 documents had been considered relevant and the precision-recall curves matched in both systems, which can be seen in Figure 5, since the only differences were the boosts in the query fields which appear to not make any difference, even though one of those fields was the "location". Most documents mentioning rivers actually refer to conflicts that took place near one, and so the curves only start to decrease after fetching most documents. The first 10 results of each system are relevant (Table 7) and so are both metrics as indicated in Table 2.

	AvP (Average Precision)	P@10
Base System	0.905253	1.000000
Boosted system	0.905253	1.000000

Table 2: Query 1 evaluation metrics

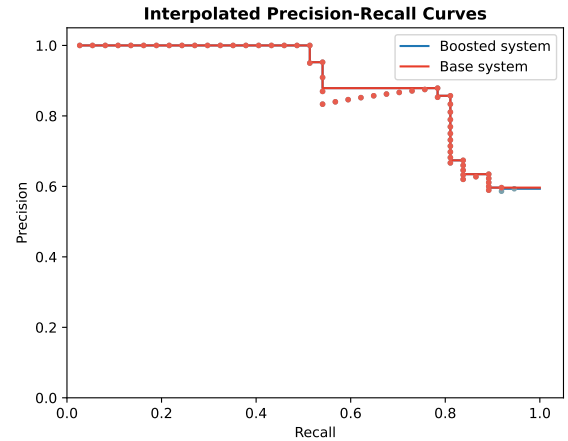


Figure 5: Interpolated Precision-Recall curves for query 1

10.3.2 Destructive Battles in Europe (World War I). For this query, 24 documents had been considered relevant and 75 were retrieved. The first 10 results of each system are not the same in both systems (Table 8) and both metrics are better in the boosted system (Table 3), in which terms related to destruction were highly favored. The precision was generally higher in the boosted system (Figure 6), proving the boosted words do really have relevance for this query. The base system has lower precision since there are many uses of words like "destruction" to refer to killings or in a broader sense, but we were instead looking for the destruction of physical locations or buildings.

	AvP	P@10
Base System	0.303171	0.200000
Boosted system	0.488829	0.500000

Table 3: Query 2 evaluation metrics

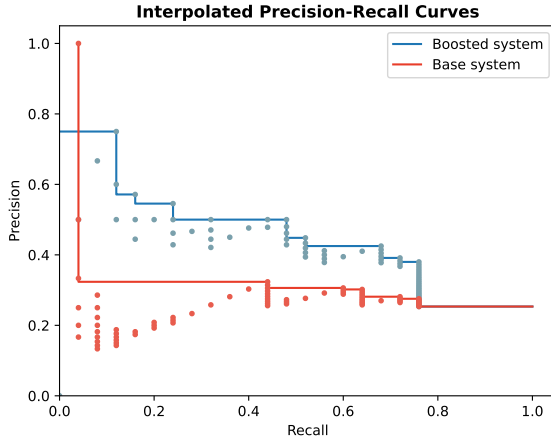


Figure 6: Interpolated Precision-Recall curves for query 2

10.3.3 Portuguese Participation in Battles (1300-1800, Non- Main Participant). For this query, 144 documents were retrieved but only 25 documents had been considered relevant. The performance of the base system is very poor with 0 relevant documents in the first 10 retrieved (Table 9) and a very low average precision (Table 4). It is difficult to correctly select such nuance just by querying "Portugal". The boosted system managed to achieve a better curve (Figure 7) by boosting terms which increased the chance of retrieving relevant documents earlier but still achieved a low average precision over the results set (Table 4).

	AvP	P@10
Base System	0.120492	0.000000
Boosted system	0.368246	0.500000

Table 4: Query 3 evaluation metrics

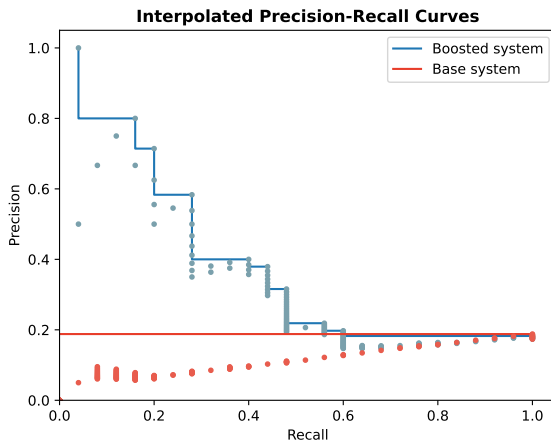


Figure 7: Interpolated Precision-Recall curves for query 3

10.3.4 Battles that are part of Revolutions that have Economic Consequences. For this query, 50 documents were retrieved and 39 documents had been considered relevant. Although the documents were not retrieved by the same order, in both cases 8 out of 10 were relevant (Table 10). The values of the metrics are also similar (Table 5) but we can see from the curves in Figure 8 that the boosted system performed slightly better, due to the boost given to "revolution".

	AvP	P@10
Base System	0.740397	0.800000
Boosted system	0.782353	0.800000

Table 5: Query 4 evaluation metrics

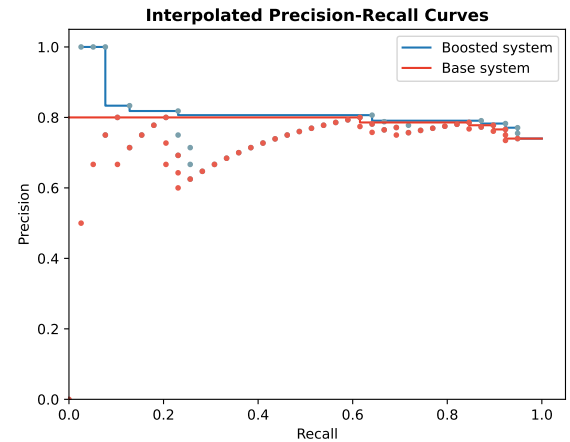


Figure 8: Interpolated Precision-Recall curves for query 4

10.3.5 Systems Comparison. In Figure 9 we observe that our boosted system managed to perform better than the base system by achieving a precision equal or better than the base system for every recall value.

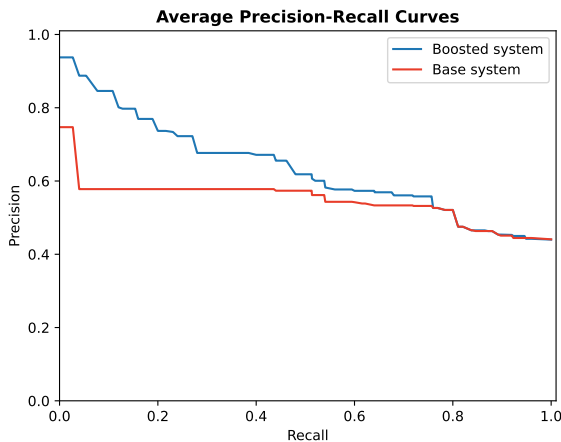


Figure 9: Average Interpolated Precision-Recall Curves of Both Systems

Milestone III

Milestone III

11 IMPROVEMENTS

We refined our search system in the most recent phase of our project, outperforming earlier iterations. Our emphasis has been on new innovations, with a special emphasis on Semantic Search. This advanced feature allows our system to understand context and meaning, improving the accuracy of information retrieval. Furthermore, a detailed evaluation of the results and a comparison with the previous version highlight.

11.1 Semantic Search

Semantic Search enhances the system’s ability to understand user queries in a more contextually meaningful way, going beyond traditional keyword matching. In this phase, our focus is on leveraging semantic analysis to extract deeper insights from historical conflict data. By incorporating advanced natural language processing techniques, the search engine can recognize relationships, context, and nuances within the dataset. This enables more accurate and relevant results, particularly beneficial for users with complex or nuanced queries.

Key aspects of our Semantic Search implementation include:

- **Contextual Understanding:** Enhancing the system’s grasp of contextual information, allowing it to interpret the meaning behind user queries more accurately.
- **Relationship Recognition:** Improving the engine’s ability to identify relationships between historical events, participants, and locations, providing a more comprehensive view of conflicts.
- **Nuanced Query Interpretation:** Enabling the system to interpret nuanced queries, such as those involving historical terminology, specific time periods, or interconnected events.

- **Enhanced Relevance:** Delivering search results with a higher level of relevance and context, enhancing the overall user experience.

The results of our Semantic Search were individually evaluated and compared with the previous version of the system, as described in the *Evaluation* section.

11.2 Learn To Rank

12 GUI

We prioritised a user-friendly experience while smoothly integrating complex capabilities in our user interface design. The graphical user interface (GUI) enhances the user’s interaction with the historical conflicts search system by providing straightforward navigation and a visually appealing layout.

12.1 Search Page

The search page has a simple and well-organized interface (see Figure 11) where users can enter queries to retrieve historical conflict materials. Each document result prominently displays a title, date of occurrence, a brief synopsis, conflict type (e.g., combat), and coordinates with a link to a map viewer for spatial context (see Figure 12). This reduced layout ensures that search results are easily understood.

In addition, for those seeking more precise results, we offer an advanced search function (see Figure 13). This tool allows users to personalise their queries, narrowing searches based on certain criteria to obtain information more precisely.

12.2 Document Page

Users are routed to a detailed document page after selecting a document (see Figure 14). This article contains detailed information, including links to Wikipedia and Wikidata for more context, as well as information on participants, location, and the connected main fight.

13 EVALUATION

13.1 Metrics

13.2 Did our search engine improved?

General Conclusion and Annexes

General Conclusion and Annexes

14 CONCLUSION

In this section, we summarize the key findings and outcomes of our project. We reflect on the challenges encountered, the solutions implemented, and the overall significance of our work. Through this conclusion, we aim to provide a comprehensive overview of the project’s achievements and contributions.

14.1 Main takeaways from the project

In this paper, we covered two milestones of the same project. In the *first milestone*, we started by selecting, collecting, and processing a dataset of historical conflicts, characterizing the data, and analyzing its spatial and temporal distribution, as well as word clouds that conceived the main keywords in the text. We finished by identifying

information needs a user of a historical conflicts search engine may have, that could be answered by our data.

In the course of developing our rich-text, coherent dataset of historical conflicts, several key takeaways have emerged:

- (1) Utilizing Wikidata as a data source has proven to be a valuable approach. The combination of structured data and rich text from Wikipedia articles has allowed us to create a robust dataset for historical events.
- (2) The spatial distribution analysis of historical events has provided insights into their geographic concentration, with a notable focus on Europe. This information is essential for understanding the global distribution of historical conflicts.
- (3) The generation of word clouds for event titles and summaries has revealed recurring themes and terms, such as the prominence of 'Battle.' This analysis provides users with an immediate visual representation of key historical themes.
- (4) Our dataset's inclusion of images and participant information is critical for enhancing users' understanding of historical events. Images offer visual context, while participant details help identify the key actors involved.

Moving on to the **second milestone**, our main aim in this milestone was to use an Information Retrieval tool to build a search engine for the data previously collected. We started by choosing this Information Retrieval tool (as previously stated we used Solr), indexing our data for search, and adding field types that could represent more accurately our data, we explored the different parameters Solr had to offer and chose the ones we found more fitting, and lastly conducted an evaluation, to see if our information needs (pointed at the end of the first milestone) were being met by the search engine, and using the parameters Solr had to offer to boost and improve the search results.

In the course of developing our retrieval process, several key takeaways have emerged:

- (1) Solr is a powerful Information Retrieval tool, from where we can extract much to improve our Information Retrieval process.
- (2) In the context of Information Retrieval, is extremely important to evaluate the results we achieved. We had our focus on the Precision at 10 (P@10), Average Precision (AvP) metrics, and Precision-Recall curves (P-R curves), as these give us a better understanding of if the information needs of the user are being satisfied, and on the direction of change needed for improving.
- (3) The results varied across information needs, as some were more straightforward than others and so it was easier to find queries to retrieve relevant documents.
- (4) Our boosted system performed consistently better overall than the base system, for the four information needs.

14.2 Future Directions

Having completed the development of our user interface and integrated semantic search and learn-to-rank capabilities into our Historical Conflicts Search Engine, we look ahead to future enhancements:

- Continuing to refine and expand the search engine's query capabilities to accommodate a broader spectrum of user queries, especially addressing complex historical research questions.
- Exploring Solr Geospatial search capabilities to retrieve conflicts that occurred in proximity to the user's current location, providing a personalized historical context.
- Ongoing work on improving the interface and interactivity of the search engine to ensure a seamless and informative user experience, enhancing accessibility and engagement.

In conclusion, this project has evolved into a robust Information Retrieval process, leveraging Wikidata and Wikipedia to create a valuable resource for history enthusiasts, researchers, and educators. The completion of this third milestone marks the end of the project and a significant step towards a comprehensive Historical Conflicts Search Engine.

REFERENCES

- [1] Folium. 2023. Retrieved 2023-10-12 from <https://python-visualization.github.io/folium/latest/>
- [2] GDELT. 2023. Retrieved 2023-10-12 from <https://www.gdeltproject.org/>
- [3] Apache Lucene. 2023. Retrieved 2023-11-15 from <https://lucene.apache.org/>
- [4] Solr. 2023. Retrieved 2023-11-15 from https://solr.apache.org/guide/6_6/the-standard-query-parser.html
- [5] Apache Solr. 2023. Retrieved 2023-11-15 from <https://solr.apache.org/>
- [6] Uppsala Universitet. 2023. Retrieved 2023-10-12 from <https://ucdp.uu.se/>
- [7] Wikidata. 2023. Retrieved 2023-10-12 from <https://www.wikidata.org/wiki/?variant=zh-tw>
- [8] Wikidata. 2023. Retrieved 2023-10-07 from <https://www.wikidata.org/wiki/Q13418847>
- [9] Wikipedia. 2023. Retrieved 2023-10-12 from <https://www.wikipedia.org/>

A ANNEXES

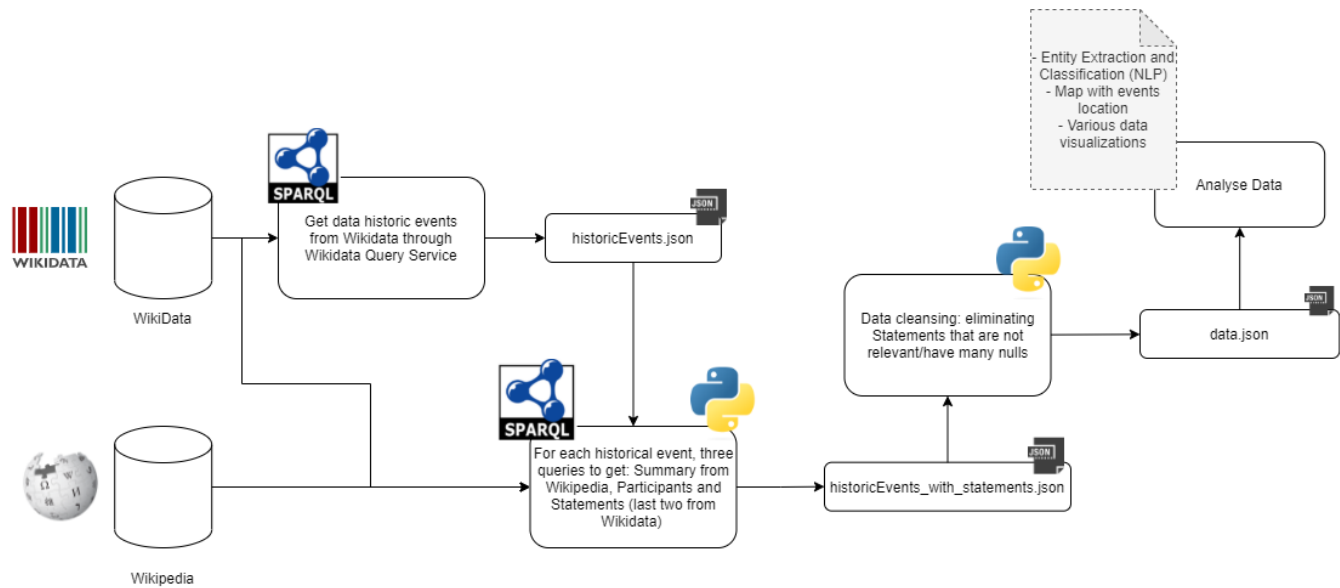


Figure 10: Data pipeline

Type	Field	Indexed
caseInsensitiveString	country	true
	instance_of	true
	location	true
location	coordinate_location	true
pdate	date	true
	end_time	false
	inception	false
	point_in_time	true
	start_time	false
pint	participants_count	true
richText	participants	true
	part_of	true
	summary	true
string	article	false
	day_in_year_for_...	false
	destroyed	false
	located_in_on_physical_...	false
	significant_person	true
	end_time	false
	facet_of	false
	followed_by	false
	follows	false
	has_cause	false
	has_effect	false
	image	false
	in_opposition_to	false
	present_in_work	false
	time_period	false
	topics_main_category	false
	labeled	false
	text	false
urlString	event	true
text_en	label	true

Table 6: Field Information (Ordered by Type)

Rank	Base	Boosted
1	Relevant	Relevant
2	Relevant	Relevant
3	Relevant	Relevant
4	Relevant	Relevant
5	Relevant	Relevant
6	Relevant	Relevant
7	Relevant	Relevant
8	Relevant	Relevant
9	Relevant	Relevant
10	Relevant	Relevant

Table 7: First 10 query 1 results

Rank	Base	Boosted
1	Relevant	Not relevant
2	Not relevant	Relevant
3	Not relevant	Relevant
4	Not relevant	Relevant
5	Not relevant	Not relevant
6	Not relevant	Not relevant
7	Relevant	Relevant
8	Not relevant	Not relevant
9	Not relevant	Not relevant
10	Not relevant	Relevant

Table 8: First 10 query 2 results

Rank	Base	Boosted
1	Not relevant	Relevant
2	Not relevant	Not relevant
3	Not relevant	Relevant
4	Not relevant	Relevant
5	Not relevant	Relevant
6	Not relevant	Not relevant
7	Not relevant	Relevant
8	Not relevant	Not relevant
9	Not relevant	Not relevant
10	Not relevant	Not relevant

Table 9: First 10 query 3 results

Rank	Base	Boosted
1	Not relevant	Relevant
2	Relevant	Relevant
3	Relevant	Relevant
4	Relevant	Not relevant
5	Relevant	Relevant
6	Not relevant	Relevant
7	Relevant	Not relevant
8	Relevant	Relevant
9	Relevant	Relevant
10	Relevant	Relevant

Table 10: First 10 query 4 results

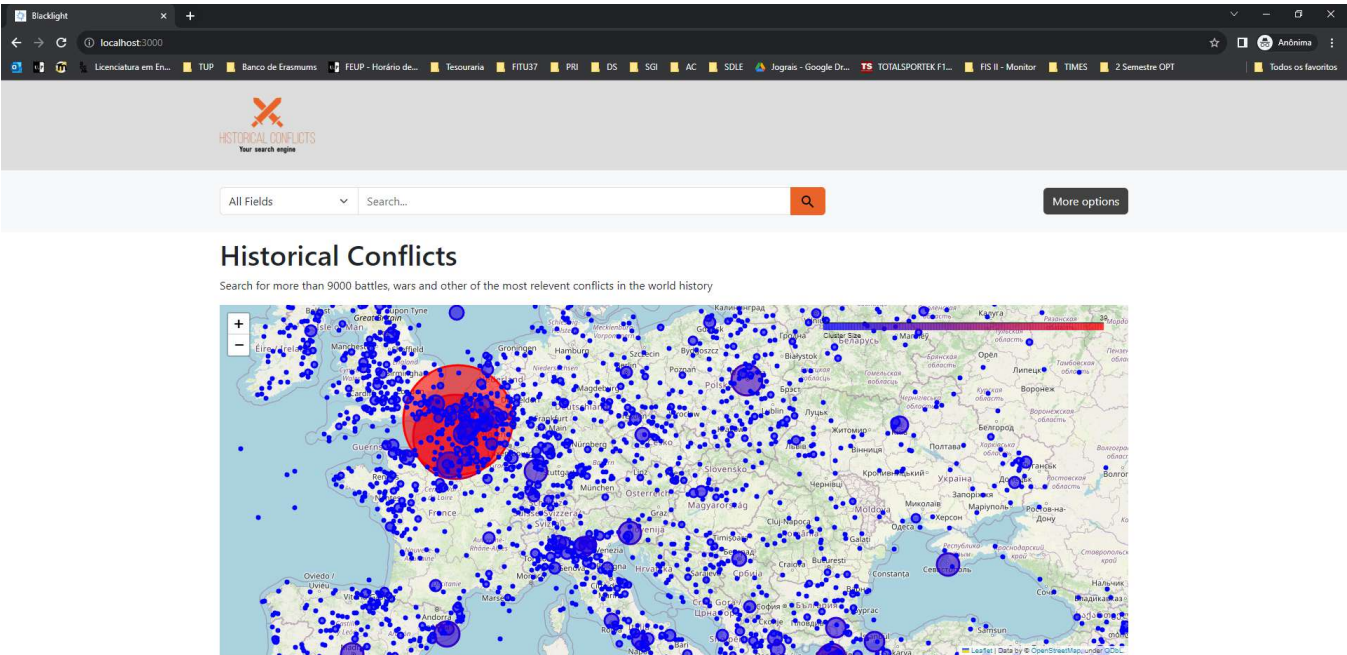


Figure 11: Home Page

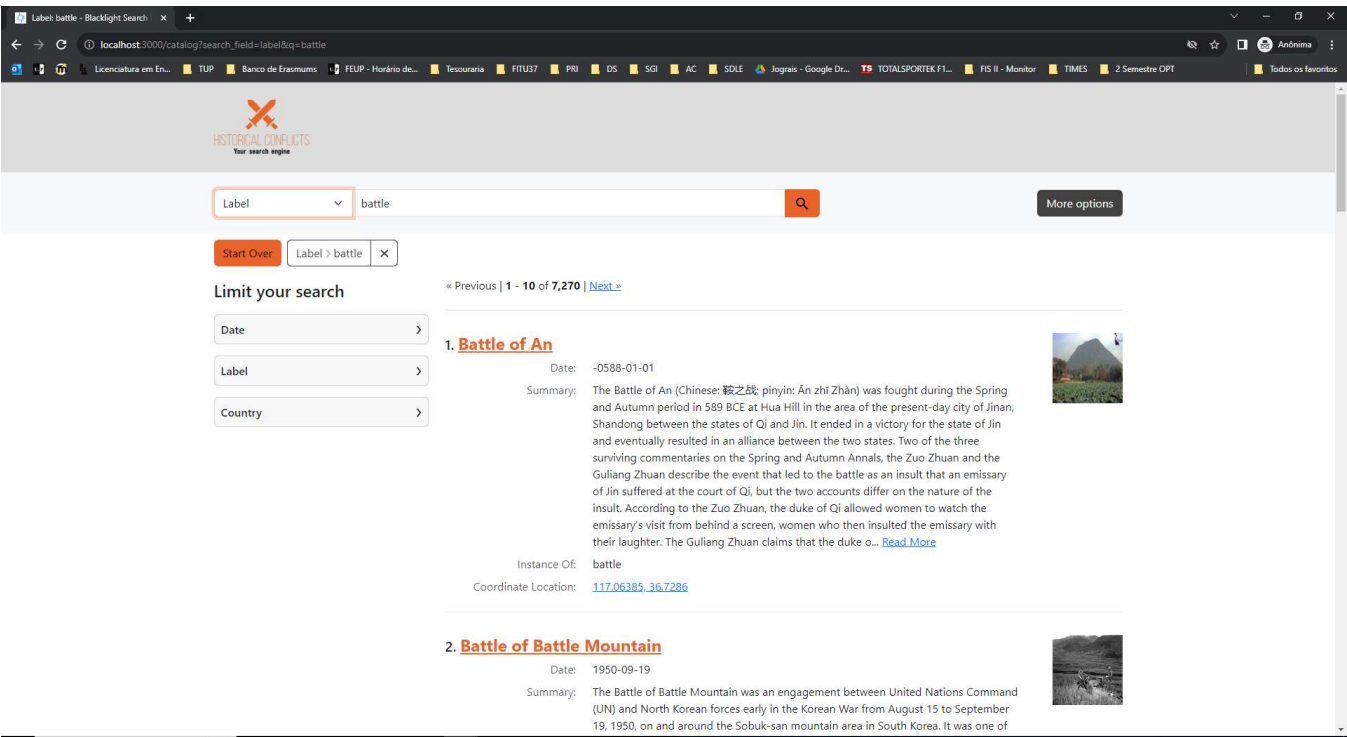


Figure 12: Search Page

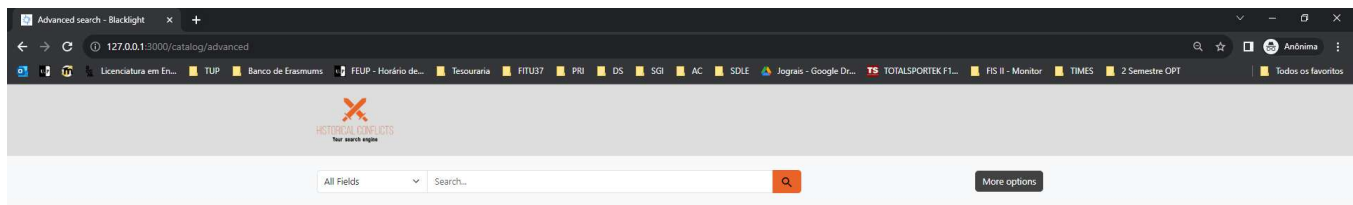


Figure 13: Advanced Search

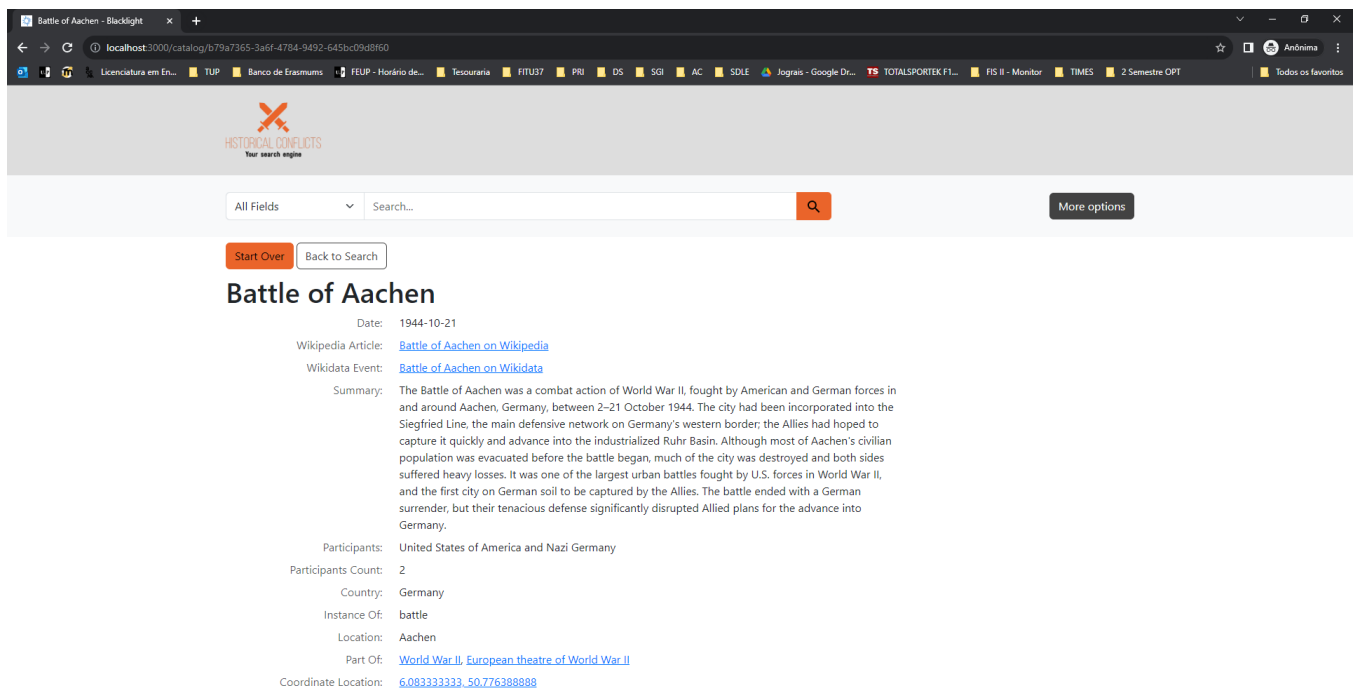


Figure 14: Document Page