Abstract. Domain-based learning and research are important applications driving the development of exploratory search systems. A wealth of historical information about events from around the world resides within documents on the web, yet contemporary search engines do not take advantage of the closely integrated temporal and spatial information found within these web pages for indexing and design of search user interfaces. This gap limits the use of the web as a resource for historical and geohistorical information seeking. In this paper we propose chronotopic information interaction as a new interaction concept for web search that explicitly links temporal and spatial entities to keywords using a space-time grid index and a paired search user interface. The space-time grid index allows different modes of interaction between spatial, temporal, and keyword-based views in the search user interface. We demonstrate use of the space-time grid index and chronotopic information interaction concept with the development of Pteraform, a prototype of a search engine that enables users to explore information in the English version of Wikipedia through a geo-historical lens.

Keywords: exploratory search, web search, historical information retrieval, geographic information retrieval, information seeking, information interaction

1 Introduction

Historical thinking and inquiry is a crucial component of active citizenship in a civil society [30,52]. While doing history is a complex endeavour of critical thinking that involves many tasks beyond the discovery of content, information retrieval systems could do a much better job to help facilitate the process. The web is a tremendous resource for historical information with millions of pages that describe rich and varied knowledge about events and processes that have occurred throughout human history. This historical information is present within the text of documents and provides an implicit structure for indexing and presenting search results. In the cataloguing systems of both digital and physical libraries, resources are occasionally organized historically or based on world geography. However, neither existing libraries nor web-based search engines provide
a systematic means for tapping into the historical and geographic content in the vast majority of documents and books that are not already organized in that way. Mining this structure from digital documents we can build search engines that facilitate historical research and learning through the search process, providing new ways of exploring and finding connections between historical events and places that are mentioned in heterogeneous web collections.

In this paper we propose a new concept of temporal and geographic information search called **chronotopic information interaction**, after the concept of the chronotope [2]. M. M. Bakhtin introduced the idea of the chronotope (time-space in Greek) in literary theory for analyzing different categories of how integrated concepts of time and space are configured in narrative texts. We have adopted this notion to the idea of using the integrated connections between time and space in web documents to build an index for interactive search. We can frame chronotopic information information in the context of exploratory search systems, where the goal is not primarily one of fact-finding, but rather to develop search tools that support constructive learning activities [41]. The purpose then of the search engine is to facilitate integrating, synthesizing, comparing, and general discovery of historical and geohistorical information in iterative search sessions.

Time and space are fundamental dimensions of information that are referenced using a diverse set of entities found in many kinds of texts across myriad subject areas. Figure 1 illustrates the spatial and temporal references in one such text, the English Wikipedia article for the *History of Montreal*. Temporal references can include dates as well as event entities, and spatial references include named places and other real locations on the Earth. A vast number of documents—from literary texts to non-fiction books, scientific articles, newspaper articles, encyclopedia entries, and special interest websites—contain these kinds of spatial and temporal references, especially at historical and geographic scales. Unlike the example in Figure 1, the documents do not need to be primarily historical in nature. For example, it can be fictional novel (e.g., *Cryptonomicon* or *Anna Karenina*) that references real world locations at different times in history. Or it can be a primary source that describes a record of historical events as they happened. The temporal and spatial references can be present at any point within the text of the document, and they provide a context for comparison with other documents on the web. Thus, space-time provides an implicit, crosscutting structure to a document collection that can enable us to find relevant documents and discover relationships based on geographic and historical context.

In this paper we present a new space-time grid indexing data structure to support the integrated search of information along temporal, spatial, and thematic dimensions. We follow with the system design for Pteraform[1], a fully integrated geo-historical search engine that allows one to explore a document corpus using the implicit references to places, times, and events in documents. The prototype was developed using the English Wikipedia data set, and the system design can be extended to other data sets.

[1] https://pteraform.csse.canterbury.ac.nz
Chronotopic Information Interaction

Fig. 1. A sample web document about the History of Montreal from Wikipedia that illustrates the kinds of temporal and spatial references that can be found within a text. Green highlighted words are temporal references and red highlighted words are spatial references. Note, that these references vary a great deal in terms of their spatial and temporal granularity.
2 Related work

Over the last 25 years, geographic information retrieval (GIR), as a sub-field of information retrieval, has been concerned with developing systems that can leverage geographic references in texts to help organize information [29]. The first fully-integrated spatial and text-based (thematic) index in GIR was developed for the SPIRIT project and has been subsequently extended [24,14,31,25]. More recently, the Frankenplace prototype introduced a discrete global grid system and indexing scheme that combined notions of spatial and textual granularity, but the index has no temporal components [4].

Document geocoding remains an active research area within the GIR literature, and this includes spatio-temporal information [35,50]. The advent of mobile search has shifted the focus of research in GIR toward developing systems that allow us to search for things in the world [13] as opposed to the original ambition to develop geographically-based search for text. In contrast to the growth in mobile search applications in industry, the map-based spatial search systems developed in research have not seen the same degree of uptake by commercial web companies.

However, the growth of the spatial humanities, the spatial turn in digital humanities and the history of sciences, has led to efforts to create large-scale databases of geographic references in historical documents [18,54,23]. Furthermore, spatial search has been proposed as a way of organizing and discovering scientific research objects [28]. A significant body of research has also focused on modeling locations with web text data, especially shorter microblog text, and ranking locations given a query [45,63,27,13].

Despite this research activity surprising little lasting advancement has been made toward implementing fully-operational web search user interfaces that exploit geographic content within documents. Furthermore, spatial and temporal references are often intrinsically related in texts—important events occur in places at specific times and spatial and temporal references co-occur often in texts [2]. Thus, the utility of geographic and historical information retrieval combined in one system has not been adequately explored; the temporal dimension has largely been ignored in GIR research in terms of indexing, relevance ranking, and search user interface design.

The utility of extracting temporal information from documents to support ad hoc search, exploratory search, and top result clustering has motivated research on temporal information retrieval [5]. Visualization of historical information within a document corpus has been explored to differing degrees. Pfoser et al. [39] developed a database of information in history textbooks using only thematic metadata (not a true text index). A conceptual model of temporal, geographic, and thematic search using ontologies was developed by Mata and Claramunt [32]. The TimeTrails project used time and space to visualize documents within a corpus but does not use a text index [49]. Chasin et al. [12] developed methods for visualizing the spatio-temporal entities found within the texts. Temporal clustering of search results based on temporal expressions in document has also been explored by Alonso et al. [6].
There is strong indication of positive outcomes from pairing digital technologies with historical learning in classroom settings \[51, 8\]. In the remainder of this paper we describe the development of a complete spatio-temporal (geohistorical) information retrieval system that combines work on geographic and temporal parsing through to indexing and search user interface development, with the research contribution emphasis on the latter two components. Such a system can support learning by putting web resources in an historical frame. The prototype was built using information from the English Wikipedia but the methods are not particular to that data set and can be applied to any collection. It is novel with respect to the strong interconnection between time and space in the development of both the indexing methods and the design the user interface.

3 Space-time grid index

In order to efficiently search for information in large text collections we need to index the documents. An inverted index that maps words and phrases to documents is a commonly used data structure in information retrieval, but a traditional inverted index does not capture any spatial and temporal structure of the information \[43\]. In this section we introduce a space-time grid index that is a data structure designed to organize document content along spatial and temporal dimensions at multiple spatial and temporal scales. It is capable of returning scores for space-time grid cells given an ad hoc keyword query.

A space-time grid is defined as a set of space-time cell pairs. A space-time cell pair, \(c\), is a tuple \(< g, t >\), where \(g\) is a cell in a two or three-dimensional spatial grid and \(t\) is a cell in a temporal grid. This can be a 2+1 or 3+1 representation of space-time \[20\]. A regular space-time grid is one where the spatial grid is a regular spatial tessellation and the temporal grid consists of fixed-sized temporal cells. A regular space-time grid for organizing information that is world-wide and covering history since the 1500s, for example, might be created using a hexagonal tessellation of the surface of the Earth and a fixed-sized timeline at the granularity of decades.

A space-time grid system is a set of space-time grids that are organized in a two-dimensional matrix of coarser-to-finer spatial and temporal grids. For global data, a hierarchical discrete global grid system can be used to specify spatial grids at multiple granularities. Irregularly shaped regions, such as administrative units, can easily be recovered from the more precise global grid \[1\]. Meanwhile, the temporal grids are defined as one-dimensional segmentations that are progressively decomposed by centuries, decades, years, months, and so on depending on the underlying temporal characteristics of the data.

A space-time grid index is word-level inverted index that matches space-time grid cells to terms, so that keyword searches can provide ranking scores for each space-time cell. The rankings can then be represented using a variety of interactive spatio-temporal visualisations, such as integrated maps and timelines. To build the index we need to define a chronotopic mapping, which is a mapping from a segment of the text within a web document to a set of space-time
grid cells. A document is defined as an ordered set of words \(< w_1, w_2, \ldots, w_m >\). A chronotopic mapping is an ordered triple \(< W, C, E >\), where \(W\) is a set of terms, \(C\) is a set of space-time cell pairs, and \(E\) is a set of edges, where \(E \subseteq \{ \{x, y, \gamma\}\} | (x, y) \in W \times C, \gamma \in \mathbb{R}\). The choice of what words from a document are mapped to a given space-time cell (i.e., how the document is decomposed into segments), as well as the choice of the weighting, \(\gamma\), will be dependent on the algorithm used to match terms within the document to spatial and temporal references or on other semantic metadata associated with the document, such as time of creation or generic information about the era or places described within the text. Ultimately, this depends on the application for the index—the underlying corpus and the kinds of queries that the system should support.

4 Data preparation

Prior to building a space-time grid index, we must first identify the place names and dates within the texts and match them to space-time grid cells. Then we need a mechanism for deciding what words are associated with those places and dates, that is, to establish the chronotopic mappings. In this section we describe how we performed this requisite step for the corpus used in the Pteraform prototype.

4.1 Pre-processing

The source data for Pteraform is the English Wikipedia dump from July 20, 2017. Each of the Wikipedia articles have been pre-processed and stored in a database with each paragraph of each article stored in a separate row. This is because the paragraph-level is the granularity we have chosen for doing chronotopic mappings. All of the data has been processed to extract geographic references and temporal references. For this we used a number of open data sets, including geographic linked data from DBpedia, Wikidata, and digital gazetteers; and temporal parsing software, including Heideltime [10][48][17][1].

The workflow involved in verifying and cleaning the results of the entity recognition tools was extensive. An initial set of named places was identified using DBpedia and Wikidata, however, the spatial data in those sources is poor (point data primarily), so the spatial representation was enriched by matching to named places in the Wāhi discrete global grid gazetteer [1]. This gave us a starting point to know what articles in Wikipedia are about places, and when other articles link to those pages about places. However, in Wikipedia, only the first reference to another Wikipedia page concept is explicitly linked in the data, so we needed to match additional mentions to named places throughout the remainder of the document. Furthermore, many articles make no explicit links to place articles at all, even when references to places exist in the text. We performed further entity recognition on the articles to make those additional links explicit in the database. We used a variety of methods in an iterative manner to perform this entity recognition starting with syntactic matching on all named places from the digital gazetteer. For ambiguous place names we then utilized
multiple open source geoparsing tools \cite{22,16,26}, choosing only places where the tools agreed, prioritizing precision over recall. Finally, we performed some sanity checks on many of the more common place names from around the world. The manual curation involved in these steps should not be understated. This process worked well for Wikipedia; however, other corpora such as collections of primary sources with historical place names, for example, would likely have required even more manual intervention. Drawing in tailored sources of geographic linked open data and using semantic annotation tools to generate new data (e.g., Recogito\cite{47}) could play an important role in those cases.

Compared with the procedure used to identify place references, identifying dates within the articles was relatively simple. We ran the Heideltime temporal parser on all the documents to match references to centuries, decades, years, months, and days \cite{48}. With both the geographic and temporal parsing there were cases of incorrectly matched places and dates, respectively. Throughout the development of the prototype system the data has been manually cleaned as errors are discovered in the data. It is important to note that the decisions made during data collection and cleaning processes, which help to define the chronotopic mappings that underlie the space-time grid index, are important design decisions when building any chronotopic search engine.

Utilizing the spatial representation from the discrete global grid gazetteer \cite{4}, we used a simple heuristic to create mappings between the terms within each paragraph and grid cells in the discrete global grid. More sophisticated natural language processing-based methods could be developed in future to better match the relationships between event references and the surrounding text, which in turn could lead to better chronotopic mappings. However, adopting a heuristic window size of individual paragraphs proved sufficient for the development of the prototype, even though in some cases the connection between a place or date and other terms within the paragraph might be tenuous. For example, in the historical summary article shown in Figure 1 most references to individual years are thematically related only to other words found within the same sentence. However, the paragraphs are roughly written in such a way that, at the century-level granularity, all the words of the paragraph can be lumped together. Thus, an index which contains this article that is built at year-level granularity will likely have some false positive results, whereas at century-level it will have fewer. In Wikipedia, most articles are not historical summaries and do not have this density of individual dates, hence we have settled on the paragraph heuristic for simplicity sake. In other data sources it might be appropriate to associate an entire document with a single date or place based on metadata information.

The space-time grid we implemented is based on an ISEA aperture 4 hexagonal hierarchical discrete global grid system (DGGS) \cite{12}. This DGGS represents a hierarchy of tessellations that increase in resolution by four (the aperture) with each level. For example, at resolution hierarchy of 8 the hexagonal grid has 655,362 equal area cells that cover the Earth. The majority of the hexagonal cells in the grid do not end up contributing to the index, however, because there are many areas of the Earth that do not have any documents associated with
them (e.g., large sections of the oceans). Two temporal grids were used: one at the granularity of centuries and another at the level of individual years. Other granularities are possible and the data described in the previous section has been pre-processed to extract additional temporal entities—such as by decade, year-month, and year-month-day—but indexes were not implemented at those levels.

Pre-processing the data and storing the intermediary data in this format is not a required to build a space-time grid index, but the data has been stored in this manner to facilitate quick development of new iterations of the prototype as well as re-use of the data for other projects. There are very few large scale datasets available for comparative analysis of GIR systems, which has hindered progressive development of new systems that can be easily compared with existing systems. The pre-processed data developed for this study is freely available for download. The data is packaged as a PostgreSQL database using the PostGIS extension and contains the following tables:

- **pages**: page information, including page id, title, and PageRank (in the Wikipedia article graph)
- **sections**: structure of the key sections in the Wikipedia page, including page id, section title (e.g., abstract), and header level.
- **paragraphs**: contains the text of each paragraph and pre-processed information: page id, section id, and extracted entities, including place links, references to days, years, decades, centuries.
- **cells**: hexagon cell ids and geometry in GeoJSON format.
- **cell mappings**: contains mappings between hexagon cell ids, temporal ids, and paragraph ids, with weightings.

The appendix has more information on the schema for each type of database table.

### 4.2 Topics

In addition to the pre-processing required for indexing, we used the Mallet toolkit to perform latent Dirichlet allocation on the entire Wikipedia corpus to generate 1024 topics. After manually cleaning up the topics to remove ‘junk’ topics, a vector containing the topic distribution for each article was stored to be included as an extra field in the document index. The topic vectors for the document results are aggregated by the user interface to generate a list of related terms at search time (see Section 6.2).

### 5 Query scoring and implementation

For the prototype that is described later in Section 6.2, we created a space-time grid index using the ElasticSearch indexing software, which is in turn built on the open source Apache Lucene project. By building off ElasticSearch, we were able to use a mature code base for query parsing and fast parallel search. For the index we implemented four new scoring models which are described in the following sub-sections.
5.1 Space-time cell scoring

The space-time grid index uses an information-based model to score space-time grid cells based on a query \[15\]. The retrieval function \( RSV_c \) is defined in Equation \[1\].

\[
RSV_c(q, c) = \sum_{w \in q \cap c} -x_q^w \log \left( \frac{\lambda_w \gamma_c - \lambda_w}{1 - \lambda_w} \right) \tag{1}
\]

In the retrieval function, \( x_q^w \) is a boosting factor for the word, \( w \), in the query, \( q \). \( t_c \gamma_w \) is a normalized version of the sum of occurrences of the word \( w \) multiplied by the \( \gamma \) weighting in the chronotopic mapping to cell \( c \). H2 term frequency normalization is used: \( tfn = tf \ln \left( 1 + \frac{avg_d}{l(c)} \right) \), which normalizes inversely related to the length \( l(c) \) of the number of words mapped to cell \( c \) [7]. Let \( \lambda_w = \frac{N_w}{N} \), where \( N \) is the number of grid cells indexed and \( N_w \) is the number of cells where word \( w \) occurs. Thus, \( \lambda_w \) is the average number of grid cells where the word \( w \) occurs. This normalization prevents locations and times that are over-represented in the data set from dominating the search result rankings.

5.2 Map cell scoring

The space-time cell score is a score for a specific place and time. A map cell score is an aggregation of space-time cell scores based on a fixed spatial (hexagon) cell, across one or more temporal units. Let \( S_c^g \) be a set of space-time cell, score tuples \(< c, s >\) based on \( RSV(q, c) \), and \( T \) be a set of temporal units corresponding to cells in the temporal grid of the index (e.g., the range from 18th to 21st century, or the years 1941 and 1942). The retrieval function \( RSV_m \) for map cell, \( g \), is defined in Equation \[2\].

\[
RSV_m(S_c^g, g) = \sum_{t \in T, t \cap S_c^g} RSV(q, < g, t >) \tag{2}
\]

5.3 Timeline unit scoring

Similarly, the timeline unit score is calculated as aggregation of space-time cell scores based on a fixed timeline unit, across one or more spatial grid cells. Letting \( G \) be a set of spatial grid cells, the retrieval function \( RSM_t \) for a timeline unit, \( t \), is defined in Equation \[3\].

\[
RSV_t(S_c^g, t) = \sum_{g \in G, g \cap S_c^g} RSV(q, < g, t >) \tag{3}
\]
5.4 Document scoring

Relevance scores for document segments are calculated using a separate, standard inverted index of words to document segments as defined by the chronotopic mappings. The retrieval function filters based on a set of selected spatial grid cells, \( G \) and temporal grid cells, \( T \), and the resulting scores are then aggregated by document id to generate scores for individual documents. Let \( RSV(q, p) \) be a relevance score value for a query and document segment, \( p \) (any scoring mechanism can be used here, such as information-based, divergence from randomness, or language modeling).

\[
RSV_d(q, d, T, G) = \sum_{t \in T \cap d, g \in G \cap d, p \in d} RSV(q, p) \tag{4}
\]

6 Search user interface

In this section we introduce a set of chronotopic interaction view dependencies based on the holistic presentation of spatial, temporal, and textual information. We follow with a description of an implemented prototype of one of the view dependencies.

6.1 Chronotopic interaction paradigms

There are effectively three dimensions of information that define the state of our chronotopic search user interface: keyword input, temporal selection, and spatial selection. The state of these inputs work in tandem to set the views for three main components of the search user interface: the map (M), the timeline (T), and the document search results (K).

The **map** is an interactive web map that consists of three main components. The first is a base map that shows the geographic frame of reference, which helps to both contextualize the search in space and understand better the geographic distribution of the search over space. The second is a hexagonal grid that represents the spatial grid cells that match the current search. The third is a heatmap overlay derived from a score for each cell.

The **timeline** is an interactive timeline that shows a bar graph that corresponds to the scores for the temporal grid cells that match the current search. It also has interaction tools to allow the user to zoom the timeline (i.e., switch between century and year granularity).

The **document search result** is a top-k list of documents for the current search (Wikipedia articles in the prototype). Additional information such as related searches can be found here as well.

These three views have dependencies, which influence the input options in the other views and thus the type of visual information seeking that we want the user to perform. In other words, these view dependencies are different model-based presentations of navigational cues for users to efficiently discover and explore knowledge in the web corpus [19]. The choice of dependency dictates what view
operates to create an overview of search results, and where the other views allow
the user to hierarchically explore contextual details on demand. For example, a
selection on the timeline can change the view on the map, or a selection on the
map can change the view on the document search result.

**K:M, K:T**—The top-k results shown in the document search result are based
purely on the document index. Selection events on the search results update the
map and the timeline independently according to the places and dates that are
referenced within the selected document. The information seeking behavior that
this dependency provides is to give the user an understanding of the temporal
and spatial context of the top-k results. Thus, it allows the user to learn about
spatial and temporal content of specific documents, but not at the corpus-level.

**M:K, T:K**—The map and timeline serve to provide an overview of the
results for a keyword search, and selection events on both the map and the
timeline affect the top-k results shown in the document search result. However,
the map does not affect the timeline view, or vice-versa. This approach uses the
map and timeline to provide a visual overview of how the keyword is represented
across space and time within the full corpus.

The following two variants of **M:K, T:K** introduce dependencies between the
map and timeline views, which allows the user to use these views to successively
refine the search and drill down to detailed results.

**T:M:K**—Selections on the timeline affect the map display, and the combi-
nation of selections on both the timeline and the map change the top-k results.

**M:T:K**—Similar to T:M:K, but in this case the selection on the map updates
the timeline view.

All of these view dependencies can be supported by the same space-time grid
index using different aggregated map cell and timeline unit scores.

### 6.2 Prototype

Pteraform is a chronotopic search engine prototype developed using space-time
grid and document indexes of the English Wikipedia, and utilizes a **T:M:K** view
dependency model. The indexes exist on a web server and are used to generate
the space-time cell scores and document scores at query time, via a web socket
connection from the browser client. The aggregate map cell and timeline scores
are generated on the client, which enables real-time interactivity.

Figures 2–7 show the basic layout of the Pteraform system with the following
main components: 1) a search box at the top for ad hoc queries, 2) a dynamic map
view, 3) a timeline representation, and 4) the top-k results window showing the
most relevant documents given the current state of the system. Four versions
of the search query *roman empire* are shown based on different states. The
results shown in these figures are built from the top-50,000 space-time grid cells
(hexagonal aperture 4 level 8 DGGS) that match the query. The total number
of unique space-time grid cell pairs that exist in the index (and the resulting
index size) depends on the spatial and temporal granularity and the density of
spatial and temporal references with the documents. For example, for the century
granularity, the index consists of 375,026 space-time grid cell pairs (29.4 GB on
Fig. 2. The heatmap shows a geographic overview of “Roman empire” references in Wikipedia without any filter on dates (spanning from 3000 BCE to present). The timeline shows a temporal overview of the same references. The document results (shown in the upper right) are based on the users’ map selection on the Balkan coast.
Fig. 3. The green circle overlay on the heatmap shows all “Roman empire” locations (i.e., grid cells) that also contain a reference to a date in the 1st century. The user makes this selection by hovering the mouse over the 1st century bar in the timeline.

Fig. 4. After making a selection on the 1st century (by clicking and dragging) the heatmap is updated to reflect only those locations that contain a reference to a date in the 1st century. The document results are likewise updated.
Fig. 5. Zooming into a local area shows the underlying hexagonal grid cells.

Fig. 6. Clicking and selecting a grid cell renders a set of related terms overlaid on the map. The terms are derived from the topics associated with the 6 document results. These terms reflect the intersection of the keyword (“battle”), the selected location (off the coast of Newcastle, England), and date selection (8th to 12th century).
Fig. 7. By changing the date selection to 18th to 21st century, 22 documents are returned and the related terms overlaid on the map change.

In Figure 2 the timeline shows the timeline unit scores in a bar graph format. The default view shows a range of centuries from the 10th century BCE to the 21st century. The map view is dependent on the timeline, so the map cell scores for all hexagons in the view window are based on aggregated values across all the centuries. The map cell scores are visualized with a density surface (a.k.a. heatmap) to give a geographic overview of “roman empire” references in Wikipedia. Map interaction then allows the user to select a single hexagonal grid cell (in this case on the Balkan coast). The top-k results shown in the document search result are based on the relevance scores given the timeline selection and map selection. 50 results are shown (the maximum number in the current version). The top result is an article about Julius Nepos, the ruler of Roman Dalmatia in the 5th century.

Following the visual information seeking mantra (“overview first, zoom and filter, then details on demand”) [46], interaction with the timeline creates real-time feedback for the user. Moving the mouse over the timeline bars highlights on the map which hexagonal cells have references to the time unit. Figure 3 shows the result of the user highlighting the 1st century with green circles overlaying the heatmap. This allows the user to compare the geographic distribution at the highlighted time unit to the overall distribution from the selected time range. This is quick visual feedback for the user but highlighting a time unit does not alter the top-k results.
However, if the user chooses to select a subset of the timeline (e.g., 1st century as shown in Figure 4), then the heatmap is updated and becomes less spatially distributed, and subsequently the document results are updated with now only 15 results shown. Because the 1st century is selected the top article becomes one on *Illyria*, a geographic region in the Balkans from antiquity, and the article on *Julius Nepos* does not appear as it is no longer relevant. The map visualization also updates based on zooming in and out. Discrete grid cells replace the heatmap as the user zooms to finer grained resolution (Figure 5).

The view dependency **T:M:K** presents a hierarchical model for visual organization of the search results, and it is possible to overlay additional views on the display. Figures 6 and 7 show how related search terms based on the document search results overlay the map view as the user zooms in on individual hexagon cells. Here the related searches derived from topic modeling are represented as a word-cloud centered on the hexagon cell 11.

### 6.3 Heuristic evaluation

For new kinds of complex user interfaces that are not directly comparable to existing technologies, a heuristic evaluation of the kinds of situations, tasks, and users that the system supports can be preferable to a traditional usability study performed in a controlled setting [37]. The value of search systems that use chronotopic information interactions will only truly become apparent through observation of the complex interactions and learning processes that such systems foster in real-world use cases. Here we focus on highlighting the strengths of our system and what differentiates it from traditional search.

There are two types of users that benefit from the utilization of the chronotopic interaction-based design. First of these are “developers”, where in we also consider information architects, such as librarians, who want to organize large collections of unstructured information. By conceptualizing the information in terms of how it will be presented along map, timeline, and more thematic dimensions and indexing it appropriately, developers are able to develop general-purpose tools for history that can flexibly index and present the collections that they possess. The second group are the “end-users” such as students and researchers who want to learn about geographic and historical information. The importance of designing search interfaces to support historical learning has been detailed above. Yet, no existing general-purpose search interfaces inherently support historical research tasks. Discovering evidence of continuity and change is one key aspect of historical thinking [44]. This can be supported by mechanisms that allow for the quick comparison between places and dates through the lens of a keyword search, something that is easy with a system like Pteraform. The system is generalizable to any kind of historical or geographic scope as well as any kind of raw textual information that might be of interest.

Although we have chosen to implement one prototype of a system with chronotopic information interaction, the space-time grid index data structure is highly flexible in that it supports a number of different ways to present the information and build view dependencies. Thus, the building of one index can
allow the design of multiple front-end solutions that support different kinds of
users performing different kinds of tasks. For example, it would be possible to
extend the Pteraform system to allow the user to select a different view depen-
dency. Likewise, it is extensible in that other kinds of semantic information
can be used to provide additional facets on the search results. In addition, the
map and timeline views provide a background canvas for overlaying all kinds of
additional thematic information.

Finally, a system using map and timeline-based visualization has an expres-
sive match to historical research tasks. Users who bring contextual background
knowledge about history or geography can leverage the user interface to hone
their search parameters. Thus, a user who is familiar with the system, will be
able to utilize the chronotopic features of the system to more efficiently discover
relevant information. Further work on understanding how chronotopic search
is used during real-world research tasks will help us to refine the best ways of
designing the map and timeline based elements of the system.

7 Conclusion

Millions of primary and secondary sources exist online for historical research.
The indexing of these sources using an explicit spatio-temporal model could
revolutionize how people discover information and learn about the past. In this
paper we presented a novel spatio-temporal grid cell index data structure that
can be used to develop a variety of different kinds of geo-historical search engines.
We demonstrated how this index can be used to develop a search user interface
that supports chronotopic information interaction, a technique for using the
integrated temporal and spatial structure in a corpus to interactively navigate
and explore the document space.

Exploratory search systems can be difficult to evaluate because traditional
information retrieval metrics do not measure the kind of performance that ex-
ploratory search systems are meant to optimize, and task or goal-based usability
evaluations must be carefully constructed to reflect appropriate proxy measures
for learning in complex environments [9,41]. In this paper we used an imple-
mented demonstration prototype along with a heuristic evaluation based on the
principles from Olsen [37] to show the potential of the chronotopic information
interaction paradigm.

Future work on chronotopic information interaction will first and foremost fo-
cus on developing appropriate evaluation methods, including how to measure the
efficacy of such a system to support research tasks as well as critical and creative
learning. Furthermore, the development of finer-grained chronotopic mappings
between web page content and temporal and geographic entities is dependent on
innovation in natural language processing methods such as extraction of event
entities, co-reference resolution, and narrative analysis. We will explore making
chronotopic mapping methods more robust across a variety of web sources beyond
Wikipedia toward the end goal of developing chronotopic information interaction
possible across large web collections, digital libraries, and other online cultural
resources. Developing infrastructure to ingest structured linked open data into a chronotopic indexing pipeline would help streamline the development of bespoke applications based on published data. In addition, although our motivation has been indexing full text document collections, chronotopic information interaction could be applied to other kinds of scientific and humanities data sets that contain references to places and dates, and have unstructured text fields—the interaction of these three dimensions is pervasive in data. For example, the view dependencies that we describe could be used to design search engines for browsing the relationships in geographic and historical linked open data.

Acknowledgments

I would like to thank the anonymous reviewers for their insightful reviews, which helped to improve the content this paper. The New Zealand eScience Infrastructure (NeSI) high performance computing system was used in the pre-processing stage of this research.

References

8 Appendix: Database tables

The pre-processed Wikipedia data that was used to build the Pteraform indexes can be downloaded as a database dump from https://www.dropbox.com/s/4z22w14fajzkma3/pteraform-enwiki-data-20200730.sql?dl=0. Note, the file size is 15.77 GB. This appendix describes schema for each table.

8.1 Pages table

This table contains the primary key (page_id) for each article in the English Wikipedia (7,955,791 rows). This is data that was extracted directly from the Wikipedia dump file. In addition, the pagerank for each article has been calculated in the context of the Wikipedia article graph. The following columns are defined:

- page_id (integer) – primary key and Wikipedia page id
- original_length (integer) – length of the article before markup removed
- new_length (integer) – length of the article after markup removed
- stub (integer) – 1 if a stub article, otherwise 0
- disambig (integer) – 1 if a disambiguation article, otherwise 0
- category (integer) – 1 if a category page, otherwise 0
- image (integer) – 1 if an image resource, otherwise 0
8.2 Sections table

This table contains all the header information for each of the pages in the database (34,478,680 rows). The header level is 0 if it is the abstract for the article. The following columns are defined.

- id (integer) – primary key
- page_id (integer) – page id of the article containing this section
- section_title (text) – section title (blank if abstract)
- header_level (integer) – header level of the section (increases for each subsection)

8.3 Paragraphs table

This table contains the text for each article divided into paragraphs (one for each row, totalling 97,352,848 rows). The ids for the paragraphs of a given page (or section) are in the same order as in the original text. The words field contains the text for the paragraph, including link information to other Wikipedia pages which are specified by <a> tags. The place_links column is an integer array containing the page ids for all places that are explicitly linked in the paragraph. The stripped_words column contains the words but stripped of punctuation and link tags. The temporal_words column shows the output of the Heideltime tagger on the words as a textual representation of a JSON array of XML snippets using the TIMEX3 format [40]. This output has been further distilled into year_refs (integer array), decade_refs (integer array), century_refs (integer array), year_month_refs references (text array), and year_month_day_refs (text array). The format for these are described in more detail below. The pct_total_words records what proportion of the total number of words for an article are found in this paragraph.

The century references are integers from -99 to 25. 20 stands for 20th century, 19 for 19th, and so on. Negative numbers refer to BCE centuries. Decade references are three-digit numbers such as 197, for 1970’s. All references are collapsed so that a reference to the day May 15, 1978 will be represented as “1978-05-15” in year_month_day_refs, “1978-05” in month_year_refs, 1978 in year_refs, 197 in decade_refs, and 19 in century_refs.
8.4 Discrete global grid cell tables

Each level of a hexagonal discrete global grid system is stored as an individual table in the database. The geometry of the cells are defined using the ISEA aperture 4 hexagonal tessellation and stored using the geography PostGIS type in the table. The table names have the format isea4h*, where the * indicates the resolution hierarchy. Thus, the table that contains the grid cells at level 6 are stored in the isea4h6 table. The table has two columns gid (primary key) and geog (geography(Polygon, 4326)). To recover the GeoJSON form of each hexagonal cell one can execute a query similar to the following: SELECT ST_AsGeoJSON(geog) FROM hexgrid.isea4h6 LIMIT 1;

A similar set of tables exists (isea4h*p) with the same ids for each hexagon and containing the geometry for the centroid of each hexagon in the geog column (geography(Point,4326)).

8.5 Cell mappings tables

Each page in Wikipedia associated with a geographic place (with spatial coordinates) has been mapped to corresponding grid cell ids based on the data from the Wähi discrete global grid gazetteer [1]. A table exists for each level in the global grid system in the form place.page_hexgrid.isea4h*. The table has two columns: page_id (integer primary key) and gids (integer[]), an integer array of grid ids that define the shape of the place.