Visualisation of 700 years of Labour Conflicts in the Netherlands

Kalliopi Zervanou[*], Vyacheslav Tykhonov[§], Antal van den Bosch[*], and Marien van der Heijden[§]

[*] Radboud University Nijmegen, The Netherlands
{KZervanou,A.vandenBosch}@let.ru.nl
[§] International Institute of Social History, The Netherlands
{vty,mvh}@iisg.nl

Abstract

Information about historical events typically lies in written sources, such as the newspapers of the time, and in the works of historians, where this information is summarized, discussed, and analysed. Although these data sources are increasingly available online, they are mostly scattered across various archives, libraries and collections. Even when this data is accessible, the sheer amount of information is no longer suited for conventional manual inspection and the people seeking this information online are not necessarily history experts. The system presented in this work attempts to address these challenges. The *Strikes in the Netherlands* system is a web-based, interactive system, which provides of a visual overview of linked aggregated data from primary and secondary historical resources, the Database of Labour Actions and the National Library of the Netherlands Databank of Daily Digital Newspapers. This browseable data overview is intended to support historians and non-specialists in exploring and retrieving information available and in spotting significant data trends across time and space that may in turn lead to new insights about historical events originally scattered across disparate sources.

1 Introduction

Primary sources of historical information, such as letters and newspaper articles, and secondary historical sources, the products of historical research such as biographies and research publications, constitute the essential evidence about facts and events in historical research. Large digitisation efforts undertaken by libraries, archives, museums and other cultural heritage institutions have been not only gradually transforming the conventional historical research methods, but have also been expanding historical information access to non-specialists, ranging from amateur historians, to history novices and casual information seekers.

However, although historical data sources are increasingly available in digital form, they are typically not easy to access and comprehend. Both specialists and non-specialists face the problem of finding these sources in pools of information scattered across various archives, libraries and collections, often lacking metadata annotations relevant to historical research and the means to associate all relevant sources relating to a given event. In addition to this data access challenge, the availability of large datasets in digital form makes imperative the need to provide more intuitive data visualisations, suitable for various information seekers, both historians and non-specialists. Such visualisations are required to support new knowledge discovery and understanding by highlighting, for example, interesting data trends and distributions, or salient associations and interdependencies in datasets that are too large for detailed manual examination.

In our *Strikes in the Netherlands* system, we attempt to address these challenges by implementing an online visualisation interface, which combines a methodology for historical data access with different types of linked data overviews. Our aim is twofold:

(i) to improve information access by refining search and by linking primary to secondary historical sources, and

(ii) to enhance data understanding and potentially provide new insights about information originally scattered along various sources, by exploiting visualisations of aggregated data trends in time and space.

Currently long-term data collections on historical events related to labour unrest, such as strikes, exist for very few countries [7]. These data collections are mostly manually compiled by historians who investigate various written sources, such as newspaper articles [7, 12]. Our methodology for linking and associating primary to secondary historical sources is motivated by the need to provide information systems for this purpose. It stems from ongoing research in text mining complex historical events denoting social unrest, such as automatically distinguishing strikes from strike threats in the news [10], identifying specific strike event references across different documents [2], discovering strike information possibly not mentioned in existing historical research [14], and, finally, retrieving news article evidence about specific strike events reported in existing historical research [14]. In the system presented in this work, we have implemented the news article retrieval and association method, for linking evidence found in the historical Dutch daily press to an existing database of labour actions in the Netherlands.

In the remainder of this paper, we begin in Section 2 with a description of the digital historical sources used in our system and, subsequently, in Section 3, we discuss in detail our system architecture, the design motivations and the implementation of the visual interfaces. We conclude in Section 4 with our observations and suggestions for potential extensions of this work.

2 Historical Resources

The historical resources used in this work are of two types. The first is the database of labour actions in the Netherlands [12], which is a secondary historical source, the result of research in compiling all information related to Dutch labour actions. The second is a collection of primary sources, an online collection of historical Dutch newspapers, provided by the National Library of the Netherlands [3]. In this section we describe these sources in greater detail.

2.1 The Database of Labour Actions

The database of labour actions is the product of an ongoing social history study into strikes and other types of labour actions in the Netherlands [12]. It contains manually compiled information about labour actions dating as far back as 1372 and up to 2008[1] [11]. In our visualisation interface, we only consider strike actions. A *strike* is defined as a specific type of labour action conforming to the following three criteria [13]:

(i) it is undertaken by employees only; student and farmer actions are not considered;

(ii) it involves a temporary interruption of work;

(iii) it is a collective action, involving the participation of at least two persons.

The database has a relational structure, consisting of 32 linked tables. The most important attribute fields for each record are illustrated in Figure 1, which depicts a screenshot of the database original online search interface[2]. The database contains information about the type of labour action (e.g., strike, lock-out, demonstration), the dates, duration, companies affected, industrial sectors, number of participants, type of participants (e.g. women, young workers, immigrants), trade unions and chambers of commerce involved, reasons, locations, outcomes and information sources. We should note that database fields do not always contain full information about the respective labour action. For example, there are cases where the exact date is unknown and the database record states only the month or the year. Moreover, information about the action's exact location or profession might be too general (e.g. "nation-wide", "general"), or even unknown. There is also a text field "Report" (*Verslag/Uitleg*), containing notes in free-text form, mainly copied from the original sources of the information, along with some other notes of the historian.

The information on the historical resources used (i.e. "Source" (*Bron*)) refers to a variety of sources, such as trade union meeting minutes, books, or newspapers articles, which are not currently available in digitised form. For this reason we have sought in this work to identify and associate primary historical sources referring to labour actions in the database using as an alternative source the digitised

[1] Data prior to 1810 is fragmentary. Our version of the data is from September 2012.

[2] https://collab.iisg.nl/web/labourconflicts/search-database

collection of daily Dutch newspapers. However, once the original sources referred to in the database become available our method could be used to associate the database information to their original source and display the respective data on our visualisation interface.

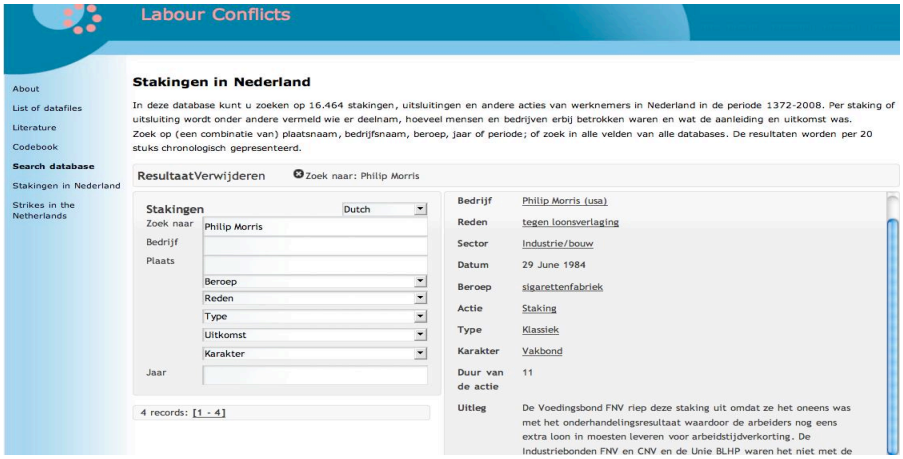

Figure 1: Database of Dutch labour actions existing online search site, with example output

2.2 The Historical Dutch Newspaper Collection

The historical Dutch newspaper collection[3] is an online collection of digitised daily newspapers dating from 1618 to 1995 [3]. It is the result of a digitisation project, the *Databank of Daily Digital Newspapers*[4], which has been initiated in 2006 by the National Library of the Netherlands aiming at digitising and providing online access to eight million pages of daily Dutch newspapers. An 8% of the 8,000 available newspaper titles have been selected for digitisation, based on a variety of importance criteria. The digitised data is in XML format, following Dublin Core [1] metadata standards. The newspapers have undergone OCR processing, and the various newspaper sections have been semi-automatically segmented. The metadata provided include newspaper title, page, section type (e.g., article, advertisement), publication and digitisation dates, author, and publisher.

The digitised newspapers are accessible online at the National Library website[5], where a search interface is provided. Moreover, the archive is also accessible via a *Search/Retrieval via URL* (SRU) interface [9]. The data accessible via the National Library online interface include all metadata, text and article image, while via the SRU, the user may have access to metadata and text only. In this work, we use the SRU interface to send our automated queries to the collection. SRU is a standard XML-based search protocol for internet search queries which exploits the HTTP GET method for message transfer [4]. The queries are formed in CQL (Contextual Query Language), a query language designed not only to be intuitive and human readable, but also more powerful than boolean search (Google-type) engine languages [4]. The SRU search interface allows for keyword-based search in newspaper segments classified as articles in the collection and published at a specific date, or date range. It supports regular expression patterns, where for example, the pattern `stak?n*` may match in the article, various words, such as *staking, staken, stakingacties, stakingen*. Finally, it allows for Boolean logical operators, such as `TermA AND TermB`, or `TermA OR TermB`. An example of a basic SRU query is illustrated in Figure 2.

SRU by definition constrains the search to be keyword-based. While this form of querying is quite powerful, it poses also certain constraints on the complexity of the query made. In particular, it is difficult to constrain and refine the results, based e.g., on a number of optionally matched terms. Moreover, it does not allow for more refined querying, such as one that would include term weighting,

---

[3] Historische Kranten: http://kranten.kb.nl/

[4] Databank Digitale Dagbladen

[5] http://kranten.kb.nl/

or one that would be sensitive to domain-specific semantic categories such as “profession” or “trade union”, as the latter are not included in the collection general metadata.

```
http://jsru.kb.nl/sru/sru?version=1.2&maximumRecords=1&
operation=searchRetrieve&startRecord=1& recordSchema=ddd&
x-collection=DDD_artikel&
query=(dc.type=artikel and date within "07-01-1980 13-01-1980" and
stak?n*)
```

Figure 2: SRU query for articles published in the second week of January 1980 with term pattern stak?n*

3 The Strikes in the Netherlands system

The Strikes in the Netherlands system is a web-based, interactive system, which provides a visual overview of linked data from the aforementioned primary and secondary historical resources, the *Database of Labour Actions* and the *National Library of the Netherlands Databank of Daily Digital Newspapers*. This data overview is intended to support historians and non-specialists in exploring and retrieving information available and in spotting significant data trends across time and space that may lead to new insights about historical events originally scattered across various sources. For this purpose, as illustrated in Figure 3, the system operates in two stages: the *Retrieval and Association* stage, where historical information is retrieved from disparate sources and associated to the event they refer to, and the second stage, the *Visualisation,* where interactive visual interfaces provide aggregated data overviews and access.

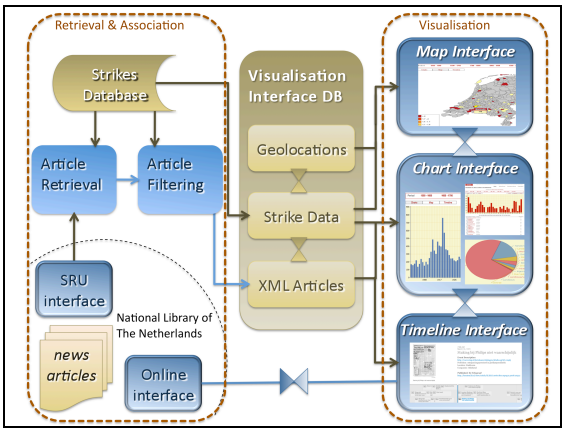

Figure 3: Strikes in the Netherlands system architecture

3.1 Retrieval and Association stage

The Retrieval and Association stage comprises of the system backbone processes for the visualisation front-end. Their objective is to retrieve, refine and enrich information about complex historical events originally located in disparate sources and store this information in a mySQL database for easy access, the *visualisation interface database*. The Retrieval and Association stage operates whenever an update of the visualisation interface database is required. It applies an information retrieval and association method and enriches the retrieved data with references to a known strike event and spatio-temporal information.

Database of Labour Actions processing: At this stage the information about strike labour actions is selected from the *Database of Labour Actions* based on the labour action type information. For these labour actions, a selection of relational information categories is retrieved. These categories, illustrated in Table 1, were selected on the basis of their potential significance in highlighting interesting strike data trends. In addition to these categories, temporal information is selected and location information is enriched with geocode ontology references to geographical coordinates. The results of this processing are stored in the *visualisation interface database*.

DB Field	Description
Sector	industrial sector affected
Bedrijf	companies or other working place involved
AantalBetrokkenBedrijven	number of companies involved
Beroep	profession involved
Bond	trade union involved
Voornamelijk	workers group involved (women, immigrants, etc.)
AantalStakers	number of strike participants
Duur	strike duration in calendar days
GestaakteDagen	man hours lost
Eisen	reason of the strike
Uitkomst	result of the strike (successful, unsuccessful, settled)
Karakter	strike character (supported by a union or spontaneous)

Table 1: Strike attribute categories used for visualisation of strike trends in time

Article Retrieval: This component uses the selected *Database of Labour Actions* strike records to form queries for relevant news articles to the *National Library of the Netherlands*. Then it parses the XML results to retrieve the full article.

For automated query formation, the component uses strike record information referring to *strike date, striker profession, location, province, chamber of commerce*, and *trade unions* involved, as query terms to the *Digital Newspapers* SRU interface. If any of these values is missing from the labour actions database, the respective information category is ignored and the query is formed with the remaining values. In transforming the database category values into query terms we ignore terms which are too general for performing a query, namely terms, such as *algemeen* (general), *arbeiders* (workers), *personeel* (personnel) and *landelijk* (nation-wide), which do not specify a particular location or profession. The date range used in the query spans from a week before to a week after the strike date mentioned in the database. Finally, the term pattern `stak?n*` is added to the database record terms.

The results of these queries to the *National Library of the Netherlands* SRU interface consist of references in XML to articles found in the historical newspaper collection and associated metadata, such as publication date, unique ID references to the article scanned image and OCR text, newspaper publisher, title and page, and other metadata information. The Article Retrieval component parses the XML results and uses the respective information to retrieve the full article text and the metadata references.

The full article XML is then enriched with information resulting from the SRU query process and the XML article reference results metadata. The SRU query process information comprises of the SRU query submitted, the reference to the specific strike record ID for which the query was formed, and the query terms found in the article text. The XML article reference results metadata that is added to the full article XML, consists of selected article metadata that is found in the results but not on the article full text and comprises of the article publication date, newspaper and page, and the article ID reference for accessing the scanned image file and other metadata, if required.

Article Filtering: As discussed in Section 2.2, SRU by definition constrains article retrieval to a keyword-based search, thus not allowing for results refinement based on term weighting or term semantic category. In the Article Filtering component we attempt to address this issue by post-filtering the retrieved articles by the SRU query process. Experiments conducted on a subset of the retrieved articles using the Article Retrieval method and discussed in detail in [14] indicate that strike articles constitute a small amount, well below 2% of the total articles published in a given time frame. The correct association of articles referring to a given strike in the database poses an additional challenge. Our experiments indicated that the correlation of retrieved strike articles to the database records may be up to 50%, with 100% precision and with recall ranging from 20% to 30%, when subsequent term weighting and filtering is applied to the initially retrieved articles [14]. In our implementation of the Article Filtering method for the *Strikes in the Netherlands* system, we have opted for high-precision term weighting, so as to ensure that the results displayed on the interface are as accurate as possible, even though this means we miss about 70% to 80% of all relevant articles. The Article Filtering component identifies matched query terms in the articles and term respective semantic categories in

the Labour Actions database, and selects articles matching at least two semantic categories. The resulting articles XML text and metadata are then stored in the *visualisation interface database*.

3.2 Visualisation stage

In designing our visualisation interface we had to consider two main issues: data overview and data access. The first consideration is related to creating effective and intuitive visual overviews for aggregated data, so that both historians and non-specialists can easily spot points of particular interest in the datasets. The second consideration is related to providing the means to the user to access the data points of particular interest and inspect the data, both for validation as well as for additional information purposes. With regards to data overviews, the obvious parameters requiring visualisation in a historical event type of research are the area where an event took place and the time, followed by any other attribute associated with the specific event, such as participants and actors.

In order to address these requirements, the visualisation stage comprises of three main interface components:

(i) a *chart interface* for interactive two-dimensional view of linked aggregated data in terms of time;

(ii) a *map interface* for interactive three-dimensional view of linked aggregated data in spacio-temporal terms;

(iii) a *timeline interface* for interactive timeline view and access to/retrieval of linked aggregated data.

As discussed in more detail below, all interfaces are interconnected. As indicated in Figure 3 of the system architecture, the *Timeline interface* is also connected to the online interface of the National Library of the Netherlands, where the user can have full access to the article image and all metadata.

Chart Interface: The *Chart interface* has been implemented using one-dimensional and two-dimensional visualisations of the strikes data in the interface database (articles and labour action descriptions). Apart from amounts of strike articles and strikes retrieved from the Labour Actions database, a detailed description of the information categories used in visualisations of trends and frequency overviews is illustrated in Table 1. The user may select the type of variable s/he wishes and may zoom in on the time scale, ranging from centuries to specific months and days. For the day display the user may choose to view either the respective database record details, as illustrated in Figure 6, or a timeline view of the articles from the National Library of the Netherlands, as illustrated in Figure 9, where again the user may read the newspaper article of interest (by a link to the respective Library page).

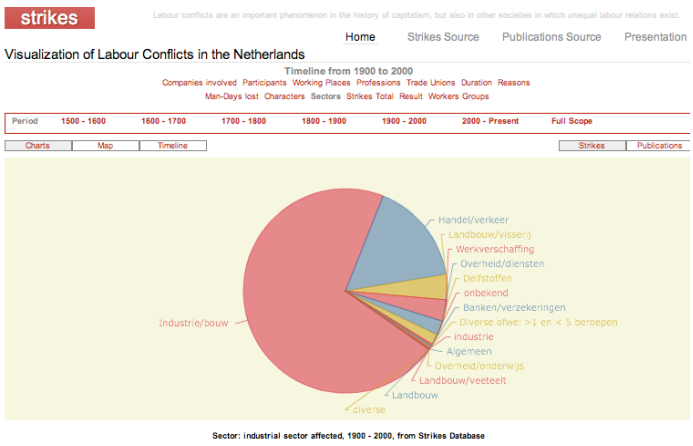

Figure 4: Pie chart of activity sectors where a strike occurred in the period 1900-2000 (based on DB data)

Information categories having a fixed set of predefined nominal values, such as the *strike outcomes* category (victory, loss, settlement, etc.) are visualised using one-dimensional pie chart implementations. An example for activity sectors where a strike occurred is illustrated in Figure 4.

We can see that in the period 1900-2000 the majority of strike actions occurred in the industrial construction (*Industrie/bouw*) and the commercial transportation (*Handel/verkeer*) sectors.

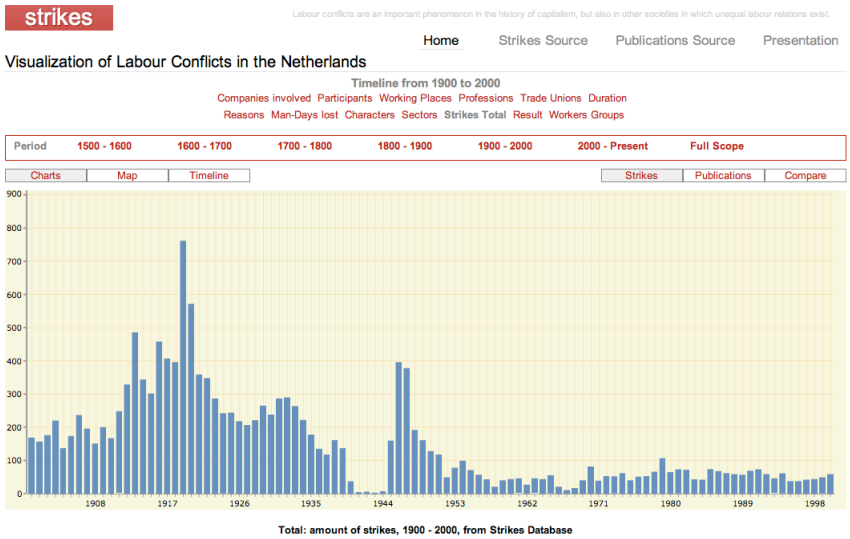

Figure 5: Strikes in the Labour Actions database in the 1900-2000 period

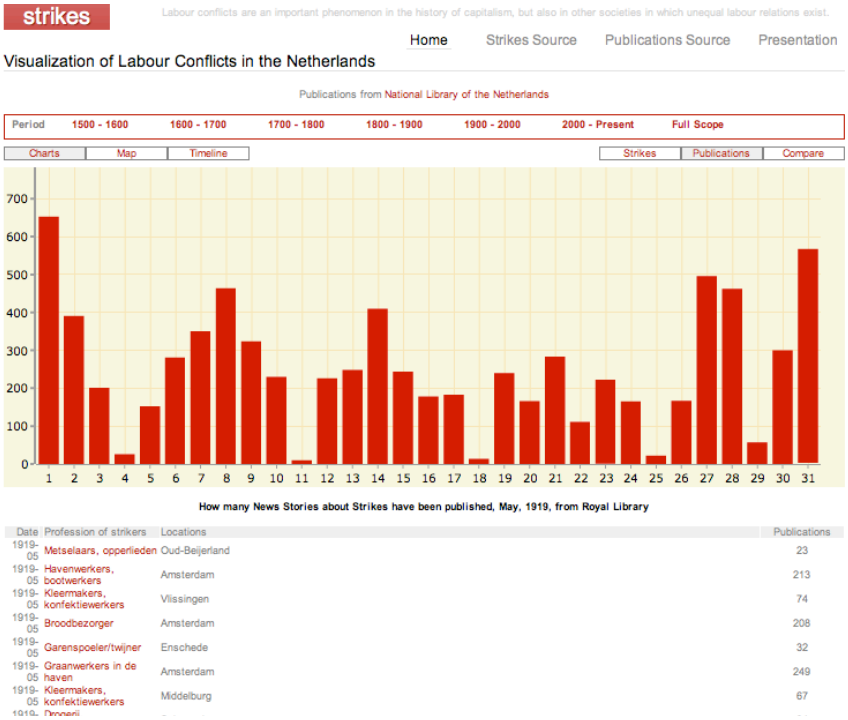

Figure 6: Newspaper articles found reporting about strike events. Time detail: month. The table below lists respective information found in the Labour Actions DB (with links to respective DB record details)

Information categories with numerical values are visualised using two-dimensional bar charts, such as those illustrated in Figure 5, which displays the amount of strikes found in the database of Labour Actions for the 1900-2000 period and Figure 6, which displays the number of strike related articles retrieved from the National Library Newspaper collection for each calendar day of May 1919, with links to the respective Labour Action database data below the chart. Other types of numerical value

information are displayed using trend lines in time, such as in Figure 7, which displays the numbers of strike participants.

The *Chart interface* components have been implemented using Open Flash Chart libraries [6].

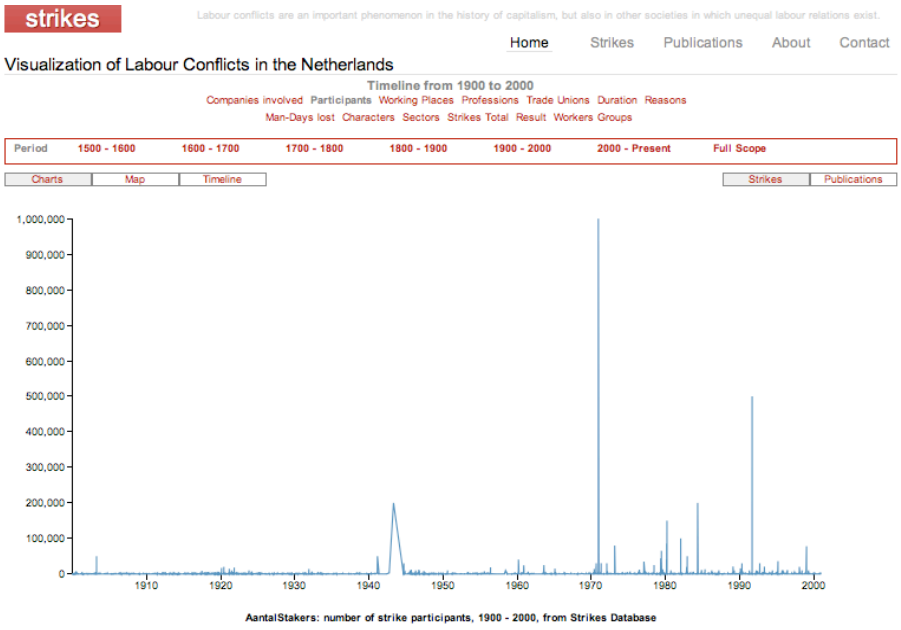

Figure 7: Number of strikers in the period 1900-2000 (based on DB data)

Map Interface: The *Map interface* implements a three-dimensional visualisation, using both space and time information. In the current version of the system it displays in a heat-map style on the map of the Netherlands the strike occurrences in time. As illustrated in Figure 8, the user may select a specific time period from a time bar at the bottom and zoom in specific region strike information by selecting regions with the mouse. The map may also display strike occurrence trends in time in animation mode for a given period.

The *Map interface* component has been implemented using the StatPlanet software [8].

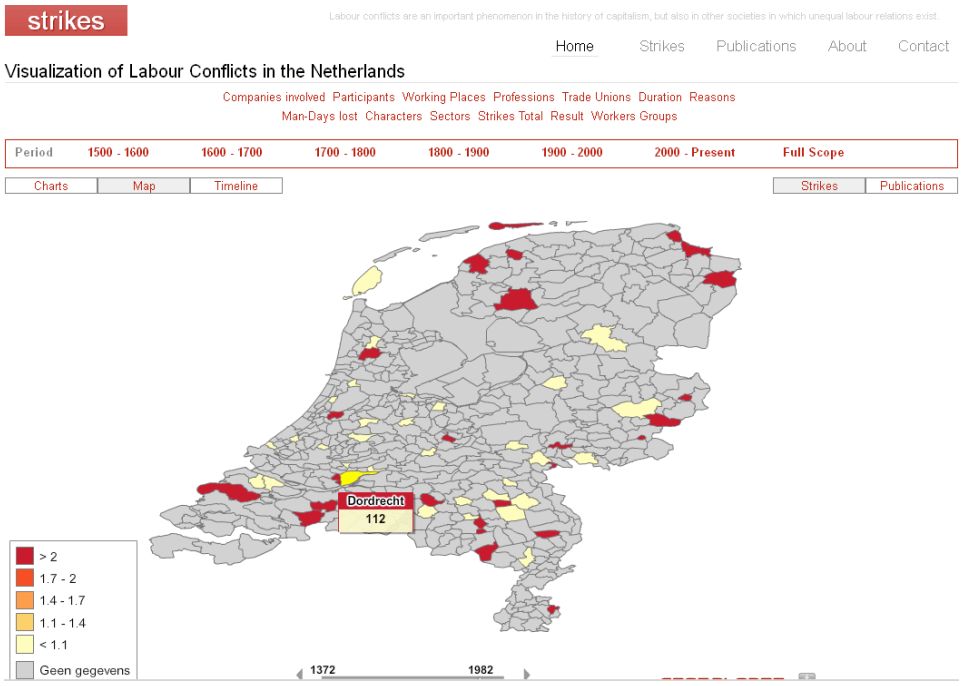

Figure 8: Strikes in 1982 in regions of the Netherlands. Detail: Dordrecht area – 112 strikes (based on DB data)

Timeline Interface: The Timeline interface provides a visualisation in time that allows for exploratory search and access to the specific newspaper text found for a strike reported in the Labour Actions database. In this sense, it combines a temporal data overview with direct fine-grained access to the information in the news articles and the associated strike research data.

As illustrated in Figure 9, the display shows the article title, the *Event Description* from the Database of Labour Actions, with a link to the respective entry on the existing online search site (c.f. Figure 1) and a link to the National Library of The Netherlands online interface for accessing all details related to the article (text, OCR image and metadata). The timeline below this information shows other strike articles found during that period.

The *Timeline interface* component has been implemented using Timeline JS [5].

Figure 9: Timeline interface

4 Conclusion

The Strikes in the Netherlands system described in this work combines a methodology for improved retrieval and association of information with data visualisation components that allow for overviews of, and access to historical sources about strikes in the Netherlands in a 700-year period. This system is intended to support both historians and non-specialists in exploring and retrieving information available and in spotting significant data trends across time and space that may lead to new insights about historical events originally scattered across disparate sources.

This work is still in a prototype phase, but we believe that it showcases the potential of the applied methods. Future versions of the system could incorporate methodologies for retrieving unknown strike events, such as for example the chi-square approach [14] and topic clustering approaches [2]. Moreover, the system could be eventually extended to cover other digitised sources about strikes such as trade union archives.

Acknowledgements. This work was partly funded by the NWO/CATCH HiTiME and the Digging into Data ISHER projects. The authors wish to thank Sjaak van der Velden for his support with the Database of Labour Actions in the Netherlands, and Marian Hellema, Anouk Janssen, and René Voorburg from the National Library of the Netherlands for their support in accessing the Dutch Daily Newspaper collection.

References

1. DCMI: Dublin Core Metadata Initiative, http://dublincore.org/
2. Hendrickx, I. Düring, M., Zervanou, K., Van den Bosch, A.: Searching and Finding Strikes in the New York Times. In: Mambrini, F. et al. (eds.) *Proceedings of the Third Workshop on Annotation of Corpora for Research in the Humanities* (ACRH-3). pp. 25–36. Sofia, Bulgaria (2013)
3. Klijn, E.: Databank of digital daily newspapers: moving from theory to practice. *News from the IFLA Section on Newspapers* (19), 8–9 (2009)
4. McCallum, S.: A look at new information retrieval protocols: SRU, OpenSearch/A9, CQL, and Xquery. In: *The World Library and Information Congress: 72nd IFLA General Conference and Council*, Seoul, Korea (2006)
5. Northwestern University Knightlab: TimelineJS, http://timeline.knightlab.com/
6. Open Flash Chart project: Open Flash Chart 2, http://teethgrinder.co.uk/open-flash-chart-2/
7. Silver, B.: *Forces of Labor. Workers' Movements and Globalization since 1870*. Cambridge University Press, New York (2003)
8. StatSilk: StatPlanet - Interactive Mapping & Visualization Software, http://www.statsilk.com/software/statplanet
9. The Library of Congress: SRU – Search/Retrieval via URL, http://www.loc.gov/standards/sru/
10. Van den Hoven, M., Van den Bosch, A., Zervanou, K.: Beyond reported history: Strikes that never happened. In: Darányi, S., Lendvai, P. (eds.) *Proceedings of the First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts*. pp. 20–28. Vienna, Austria (2010)
11. Van der Velden, S.: Database of Dutch labour actions. Available online at: https://collab.iisg.nl/web/labourconflicts/datafiles
12. Van der Velden, S.: *Stakingen in Nederland. Arbeidersstrijd 1830–1995*. Stichting Beheer IISG/NIWI, Amsterdam, The Netherlands (2000)
13. Van der Velden, S.: *Werknemers in actie. Twee eeuwen stakingen, bedrijfsbezettingen en andere acties in Nederland*. Aksant, Amsterdam (2004)
14. Zervanou, K., Düring, M., Hendrickx, I., Van den Bosch, A.: Documenting Social Unrest: Detecting Strikes in Historical Daily Newspapers. In: Nadamoto A. et al. (eds.): *Social Informatics* (SocInfo 2013-Histoinformatics). Springer: LNCS, vol. 8359, pp. 120-133 (2014)