Exploring Patient Organization Periodicals with the Topic Timelines Text Visualization Method

Maria Skeppstedt[1,*], Adam Maen[2], Vera Danilova[3], Gijs Aangenendt[3], Andrew Burchell[3] and Ylva Söderfeldt[3]

[1]*Stockholm University, Department of Linguistics, Stockholm, Sweden*
[2]*Uppsala University, Centre for Digital Humanities and Social Sciences, Department of ALM, Uppsala, Sweden*
[3]*Uppsala University, Department of History of Science and Ideas, Uppsala, Sweden*

Abstract

The text visualization technique Topic Timelines offers a compact visualization to represent the evolution and clustering of topics over time, while also providing direct access to the texts in which these topics appear. In this paper, we describe how Topic Timelines was further developed within the ActDisease project, by adding functionality for generating timelines using different types of topic extraction techniques and connecting the visualization to existing interfaces for the close reading of texts. Additionally, we evaluate how the updated temporal topic overview can support corpus exploration.

The experiments were conducted on a digitalized corpus from the ActDisease project, consisting of patient organization periodicals from the Swedish Diabetes Association, published between 1949 and 1990. Timelines were generated based on topics extracted using sentence transformers clustering and integrated with the ActDisease text database interface – a user interface developed for exploring and reading texts digitalized within the project.

Keywords

Text visualization, Automatic topic extraction, Sentence transformers, Patient organizations

1. Introduction

The ActDisease project studies the history of patient organizations in twentieth century Europe, combining traditional modes of analysis with computer-based methods[1]. Periodicals from selected British, French, German and Swedish patient organizations have been digitalized [1] and made searchable through a graphical user interface developed within the project. The interface enables the historians within the project to carry out keyword searches, filter on type of publication and language, as well as browse the scanned magazine issues page by page (see Figures 1 and 2).

In addition to searching and browsing the materials using the interface, the Topic Timelines visualization technique has been developed within the project as an additional method for exploring the patient organization periodicals [2]. By visualizing the evolution of topics within a corpus over time, an overview of the corpus content is provided. However, previous techniques for visualizing topics in temporal text collections [3, 4, 5, 6, 7, 8, 9, 10, 11, 12] typically aim to support quantitative data exploration, and therefore only provide a temporal overview in an aggregated format, without any direct connection to the texts from which the data is extracted. Topic Timelines, in contrast, is specifically designed to support historical research, offering a compact and information-rich overview visualization of topics in a corpus, while also providing direct access to the texts in which these topics appear[2].

Huminfra Conference 2025, Stockholm, 12–13 November 2025.
*Corresponding author.
maria.skeppstedt@ling.su.se (M. Skeppstedt); adam.maen@abm.uu.se (A. Maen); vera.danilova@idehist.uu.se (V. Danilova); gijs.aangenendt@idehist.uu.se (G. Aangenendt); andrew.burchell@idehist.uu.se (A. Burchell); ylva.soderfeldt@idehist.uu.se (Y. Söderfeldt)

[1]https://www.actdisease.org
[2]The code for creating the visualization, as well as examples of how the visualization can be used is found here: https://github.com/CDHUppsala/topic-timelines

This paper discusses the most recent developments of Topic Timelines, focusing specifically on the added functionality for generating timelines using different types of topic extraction techniques as well as the possibility to use the visualization as a method for navigation in existing database interfaces. As a case study for testing and evaluating the updated Topic Timelines, we generated a visualization for one of the ActDisease corpora – periodicals published by the Swedish Diabetes Association between the years 1949 and 1990[3] – and connected it to the graphical user interface for the ActDisease database.

2. Adapting and connecting Topic Timelines

In previous versions [2, 13], Topic Timelines visualizations could only be generated from the output of one specific topic extraction tool, the Topics2Themes tool [14]. To make it possible to apply the visualization technique to different types of topic extraction methods, we updated the Topic Timelines Python module, enabling it to create visualizations based on data provided in a tabular format[4]. There are many different topic extraction methods, each of them with different advantages. The previously used Topics2Themes tool extracts topics based on word co-occurrence patterns, and thereby provides topics that are relatively transparent to the user. In the case study described below, we instead used topic extraction based on clustering of sentence transformers, a technique that, e.g., allows the user to extract topics from multi-lingual corpora.

In addition, we explored the possibility to extend the support for organizing the topics on the timeline. Analyzing and comparing many topics on a timeline is difficult, even on a large computer screen. The Topic Timelines implementation, therefore, provides the user with the possibility to manually determine the order in which the extracted topics appear. We have previously used this functionality to manually group topics that share the same characteristics or themes. In this iteration of Topic Timelines development, we experimented with using topic clustering to instead automatically generate such high-level topic groups.

The third focus of this study was an integration of the timeline visualization with the graphical user interface for the ActDisease database, to investigate whether a thematic exploration of the database contents could complement the traditional ways of searching and browsing. The Topic Timelines Python module makes it possible to link each text visualized to a URL, which can be reached by clicking in the graph. We here linked each text to a corresponding URL in the ActDisease database interface. To make this connection, there must be a naming consistency between the URLs in the graphical user interface and the file names in the text corpus from which topics are extracted. We, therefore, updated the ActDisease interface so it uses "Clean URLs", where the URL address matches the filename in the text corpus. The user can thereby start their exploration by identifying potentially interesting texts in the graph, based on the timeline visualization, and then click on these texts to reach the standard ActDisease database interface. In the database interface, users can then inspect the scanned image and the extracted text, but also browse adjacent pages of the selected text. This is a development compared to previous studies [2, 13], where the user was directed to a separate webpage displaying only the scanned image, without any links to adjacent pages in the magazine.

3. Case Study

To test and evaluate the developments of Topic Timelines, we generated a visualization for the corpus of periodicals published by the Swedish Diabetes Association and connected it with the graphical user interface for the ActDisease database. Finally, the historians in the project inspected the timelines and the automatically generated ordering of the topics.

[3]The materials scanned by Gothenburg University Library can be found here: https://gupea.ub.gu.se/handle/2077/64597
[4]An example of this format can be found here: https://github.com/CDHUppsala/topic-timelines/tree/main/diabetes_topics

3.1. Method and configuration parameters for the topic extraction

The method used for creating topics was sentence transformers [15], which is a technique for encoding the meaning of a text unit into a vector. The distance between these vectors corresponds to the similarity between texts, and it is therefore possible to extract topics by performing a distance-based clustering of the vectors. Clusters of vectors then correspond to groups of texts belonging to the same topic.

We encoded the text units (i.e., in most cases the pages)[5], into vectors using the *paraphrase-multilingual-MiniLM-L12-v2*[6] SentenceTransformer [15]. We then applied the functionality for agglomerative clustering, which is available through scikit-learn [16], on these vectors. Agglomerative clustering is a clustering method that starts with each vector belonging to its own separate cluster, and then iteratively merges the two clusters that are closest to each other into larger clusters, until a user-specified distance threshold is met.

As stressed by Da [17, p 625], the output from automatic topic extraction is heavily dependent on the methods used and the parameters settings selected. Therefore, the topics extracted and visualized should be looked upon as a set of potentially interesting topics in the corpus, rather than as *the* topics in the corpus[7]. When experimenting with different configurations settings, our aim was consequently to render timelines containing topics with the potential to be interesting to the user. Therefore, to generate topics that were focused and specific enough, we varied the threshold for when to stop the merge of clusters. In addition, to avoid clusters containing too few elements, we implemented the functionality of allowing small clusters to be merged with the closest neighbouring cluster, regardless of distance. Finally, to avoid topics being created based on similarity between uninteresting function words, rather than between content words, we applied stop word filtering before encoding the texts into vectors.

We iteratively refined the clusters, resulting in the following configuration: a cosine distance threshold of 0.47, a minimum cluster size of 50 texts, and the standard Swedish stop word list from NLTK [19], extended with 80 corpus-specific stop words. This resulted in 68 topics being extracted and visualized.

Also for creating the high-level clusters, agglomerative clustering was used, with a cosine distance threshold value of 0.35. This resulted in 12 high-level groups, and there were 7 outlier topics, not fitting in any of the groups.

3.2. The resulting visualization

The resulting timeline graph is shown in Figure 3, and a description of the graph components is shown in Figure 4. Each one of the 68 topics is represented by a horizontal lane, and each text in the corpus is represented by a vertical line[8]. The position on the timeline for the vertical text-line is determined by the date when the text was published. In the case of the periodical issues, several texts are published on the same date. As a solution, the text line corresponding to the first page is placed on the position corresponding to the publication date, and the lines for the following pages are moved slightly to the right. The automatic division into high-level groups of topics is represented by color coding.

When a text belongs to a topic, this is indicated by a circle in the graph at the point where the text line and topic lane intersect. The more typical a text for a topic, the larger the circle representing the text-topic connection[9]. The circles are semi-transparent, which has the effect that the prevalence of the same topic in many adjacent texts is represented by a pattern of overlapping, semi-transparent circles.

[5]The corpus investigated is divided into pages. For most cases, a text unit therefore means a page. However, long pages had to be split into smaller subtext units, to accommodate for the maximum token limit of the sentence transformer used (512 tokens). For these pages, blank lines – which often signal paragraph breaks – were used for splitting the texts. The 1 112 pages in the corpus that contained more than 300 words were split into subpages, resulting in a total of 9 775 texts to cluster.

[6]https://huggingface.co/DataikuNLP/paraphrase-multilingual-MiniLM-L12-v2

[7]As a more objective measure of the corpus content, we would argue that frequency-based methods instead form a better choice [18].

[8]We refer to https://github.com/CDHUppsala/topic-timelines and previous publications for a more detailed description of the visualization technique.

[9]We here measured how typical the text was, by its distance to the centroid of the cluster. I.e., the closer to the centroid of the cluster, the more typical is the text.

When clicking on a circle, the user is directed to the ActDisease database interface, showing the original text in which the topic occurs (as shown in Figure 5).

3.3. Reading the timelines

The most distant view of the graph is provided by the color-coded grouping into high-level topics. This view tells us that the corpus contains topics related to (a) nutrition, (b) food and sweeteners, (c) insulin and measuring of blood sugar, (d) insulin pumps and medical effects of diabetes, (e) insulin, activities for members, (f) children and youth, (g) members and organizational matters, (h) disability and economical matters, (i) the organization's chairperson and (perhaps) different types of letters, (j) travel, (k) economical matters and taking insulin, and (l) (perhaps) personal patient stories.

By zooming in one level into the visualization, we can move from the high-level clusters to study the temporal characteristics of the individual topics. For instance, topic 25 (about insulin) occurs frequently during the entire time period studied, topic 22 (about economical matters) occurs regularly once per year during a large part of the time series, the topics on patient stories become frequent during the 1980s, and finally topic 40 (about driving licenses), is very concentrated to the year 1965. This indicates that certain themes emerge at "flashpoints" – in the case of topic 40, campaigning around medical certification for driving – while others represent the consistency of organizational administration (financial reporting) or more slow-progressing cultural changes within the formatting and editorial choices of the magazine (the foregrounding of personal narratives and experiences). Advertisements likewise evolve differently, e.g., while topics 14 (sweeteners) and 53 (self-testing kits for urine) cover a long chronological range, the marketing of newer products which only entered the market during the 1980s (topics 46 and 61) is rendered more pronounced by the visualization.

Finally, the most detailed level is accessed by clicking on the circles to reach the original texts, as they are presented in the ActDisease database interface. For instance, a closer study of the texts related to the "advertisement" topics, allows us to see that the high concentration of these topics in the timeline visualization in the 1980s is not only a testament to the substantial advertising campaigns led by manufacturers (as shown in the example in Figure 5) but is also a consequence of, e.g., technical devices appearing in other parts of the periodical, including reviews and personal narratives of use.

3.4. Discussion

The principal benefit of the integration between the database and the timeline is that it simplifies not only navigation between the data points but also contextualization of the results. This allows the automatic topic extraction to exist in dialogue with more empirical methodologies, as we preserve the ability to move backwards and forwards in the magazine from a relevant data point as well as the physical experience of browsing the magazine as an artefact. This movement between the different "levels" of the corpora – as statistical data and modelled topics on the one hand, and textual, visual and formatted documents (albeit mediated through a digitalized database) on the other – offers the potential for fruitful and more intuitive uses by historians whose primary training is not in quantitative methods. It also allows users to merge more traditional and digital methodologies by collapsing the digital/material divide between the sources.

It can further be noted that the automated generation of topic groups suggests an interesting tension in the function of these visualizations. Here, the (auto)generated high-level clusters suggest broader themes which are of interest to historians: the mediation of consumption and self-care practices by people with diabetes (clusters a to d, h and k), the administrative life of the patient organizations themselves (clusters e, f, g, h and i), and the rise of individual narrative practices in the later half of the twentieth century (potentially clusters i and l). Yet cutting across the higher-level clusters are topics which also offer entry points into the different genres of text existing in the magazine corpora. Topics 14, 46, 53 and 61 are advertisements for a range of specialized products for people with diabetes, including sugar-free sweeteners (14), urine glucose-testing kits (53), and digital blood-glucose monitors (46 and 61). In previous iterations of these timelines, historians have sometimes tended to classify as much by genre as by theme – to make these advertisements more visible. As this case demonstrates, there are no "correct"

or “incorrect” classifications and clusters, and it is helpful to consider one’s own clustering in the context of other possible classification logics.

Finally, based on what topics were extracted, it seems that the topic timeline method is unusually suited to certain kinds of textual genres – and especially to the advertisements – because they are relatively stable texts in temporal and historical terms. Brand names, slogans, key selling points, and product designations often remain unchanged across long time periods, even if the advertisements themselves could be renewed or updated regularly to maintain reader interest. This is visible from the preponderance of proprietary names such as “sionon” (topic 14), “reflolux” and “boehringer [mannheim]” (topic 46), or “clinitest” and “ames” (topic 53) in the topic timeline. It is possible that this relative textual stability produces a slight bias within the generation of topic timelines. In addition, the format of the magazines, and the marketing economics of the manufacturing companies, meant that many of the advertisements for the testing devices were full-page by the 1980s (for an example, see Figure 5). Since advertisements are relatively short, and the texts were normally split for clustering on page level, this might result in a higher degree of coherency for the full-page advertisements than for the other texts, which might also have introduced a bias for advertisements.

4. Conclusions and future work

Overall, we would suggest that interactive navigation of the magazine database by automatic topic extraction and timeline visualizations offers potential for historians beyond usual keyword searches. Such methods can allow historians to pass between the “levels” of distant versus close textual reading, while also preserving the ability to interact with the visual and aesthetic dimensions of the magazines. Nonetheless, it is preferable to retain the flexibility with regard to the high-level clustering of topics: automated versions of this can prove useful in stimulating fresh perspectives, but they may have limitations when it comes to the presentation of results.

As a next step, we will develop the sentence transformers clustering described here into a Python module generally applicable on temporal corpora. We then aim to make the text clustering and the Topic Timelines visualization available as a resource within the Huminfra research infrastructure.

We also plan to employ the genre classifier [20] developed for the corpora digitalized within the ActDisease project to more closely investigate the relationship between genre and topic. For instance, by creating topics only for texts belonging to a certain text genre, or by investigating the prevalence of different genres within automatically extracted topics. This work also includes an exploration of more semantically informed methods for dividing the pages into subtexts.

Additionally, we are investigating further integration of the database into the visualization. This would enable the creation of an interface that could allow the user to browse the database not just chronologically by material in the same issue/volume of the magazine, but also according to related data points in the topic modeling and timeline. This would open up new possibilities for reading the database against the chronological and thematic grain, while preserving the possibility to still explore the magazines in a more traditional way.

Acknowledgements

The work described here was conducted within the ActDisease project. ActDisease is funded by the European Union (ERC ActDisease, ERC-2021-STG 101040999). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

The development of Topic Timelines has been partly supported by Huminfra (the Swedish Research Council, grant 2021-00176) and InfraVis (the Swedish Research Council, grant 2021-00181).

References

[1] G. Aangenendt, M. Skeppstedt, Y. Söderfeldt, Curating a historical source corpus of 20th century patient organization periodicals, in: Proceedings of the Huminfra Conference (HiC 2024), 2024, pp. 76–82. URL: https://ecp.ep.liu.se/index.php/hic/article/view/895. doi:10.3384/ecp205011.

[2] Y. Söderfeldt, A. Burchell, J. Reed, M. Skeppstedt, Topic timelines for enabling close and distant reading of discursive shifts. a pilot case using periodicals of european diabetes organizations, Journal of Open Humanities Data (2025). doi:10.5334/johd.286.

[3] M. Grootendorst, Dynamic topic modeling, visualization, https://maartengr.github.io/BERTopic/getting_started/topicsovertime/topicsovertime.html, 2023.

[4] D. Blei, J. Lafferty, Dynamic topic models, in: ACM International Conference Proceeding Series; Vol. 148: Proceedings of the 23rd international conference on Machine learning; 25-29 June 2006, ACM, 2006, pp. 113–120.

[5] S. Sheehan, S. Luz, M. Masoodian, TeMoTopic: Temporal mosaic visualisation of topic distribution, keywords, and context, in: H. Toivonen, M. Boggia (Eds.), Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, Association for Computational Linguistics, Online, 2021, pp. 56–61. URL: https://aclanthology.org/2021.hackashop-1.8.

[6] S. Malik, A. Smith, T. Hawes, P. Papadatos, J. Li, C. Dunne, B. Shneiderman, Topicflow: visualizing topic alignment of twitter data over time, in: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 720–726. URL: https://doi.org/10.1145/2492517.2492639. doi:10.1145/2492517.2492639.

[7] C. Meaney, M. Escobar, T. A. Stukel, P. C. Austin, L. Jaakkimainen, Comparison of methods for estimating temporal topic models from primary care clinical text data: Retrospective closed cohort study, JMIR medical informatics (2022).

[8] S. Gad, W. Javed, S. Ghani, N. Elmqvist, T. Ewing, K. N. Hampton, N. Ramakrishnan, Themedelta: Dynamic segmentations over temporal topic models, IEEE Transactions on Visualization and Computer Graphics 21 (2015) 672–685. doi:10.1109/TVCG.2014.2388208.

[9] S. Havre, B. Hetzler, L. Nowell, Themeriver: visualizing theme changes over time, in: IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings, 2000, pp. 115–123. doi:10.1109/INFVIS.2000.885098.

[10] N. Günnemann, M. Derntl, R. Klamma, M. Jarke, An interactive system for visual analytics of dynamic topic models, Datenbank-Spektrum 13 (2013) 213–223. URL: https://doi.org/10.1007/s13222-013-0134-x. doi:10.1007/s13222-013-0134-x.

[11] N. Günnemann, D-vita: A visual interactive text analysis system using dynamic topic mining, in: Datenbanksysteme für Business, Technologie und Web, 2013. URL: https://api.semanticscholar.org/CorpusID:15848321.

[12] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, X. Tong, Textflow: Towards better understanding of evolving topics in text, IEEE Trans. Vis. Comput. Graph. 17 (2011) 2412–2421. URL: https://doi.org/10.1109/TVCG.2011.239. doi:10.1109/TVCG.2011.239.

[13] M. Skeppstedt, G. Aangenendt, V. Danilova, Y. Söderfeldt, Topics in periodicals from the swedish diabetes association 1949 – 1990: Extending the topic modelling tool topics2themes with a timeline visualisation, in: Selected papers from the CLARIN Annual Conference 2023, Linköping University Electronic Press, 2024. doi:https://doi.org/10.3384/ecp210015.

[14] M. Skeppstedt, K. Kucher, M. Stede, A. Kerren, Topics2Themes: Computer-assisted argument extraction by visual analysis of important topics, in: Proceedings of the LREC Workshop on Visualization as Added Value in the Development, Use and Evaluation of Language Resources, 2018, pp. 9–16.

[15] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing,

Association for Computational Linguistics, 2019. URL: http://arxiv.org/abs/1908.10084.
[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[17] N. Da, The computational case against computational literary studies, Critical Inquiry 45 (2019) 601–639. doi:10.1086/702594.
[18] M. Skeppstedt, M. Ahltorp, K. Kucher, M. Lindström, From word clouds to Word Rain: Revisiting the classic word cloud to visualize climate change texts, Information Visualization (2024). doi:10.1177/14738716241236188.
[19] S. Bird, NLTK: The natural language toolkit, in: Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, 2002.
[20] V. Danilova, Y. Söderfeldt, Classifying textual genre in historical magazines (1875-1990), in: Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 160–171. doi:10.18653/v1/2025.latechclfl-1.15.

A. Appendix

As an appendix, we have included the pdf version of the Topic Timelines visualization, as well as screenshots of the ActDisease text database interface.

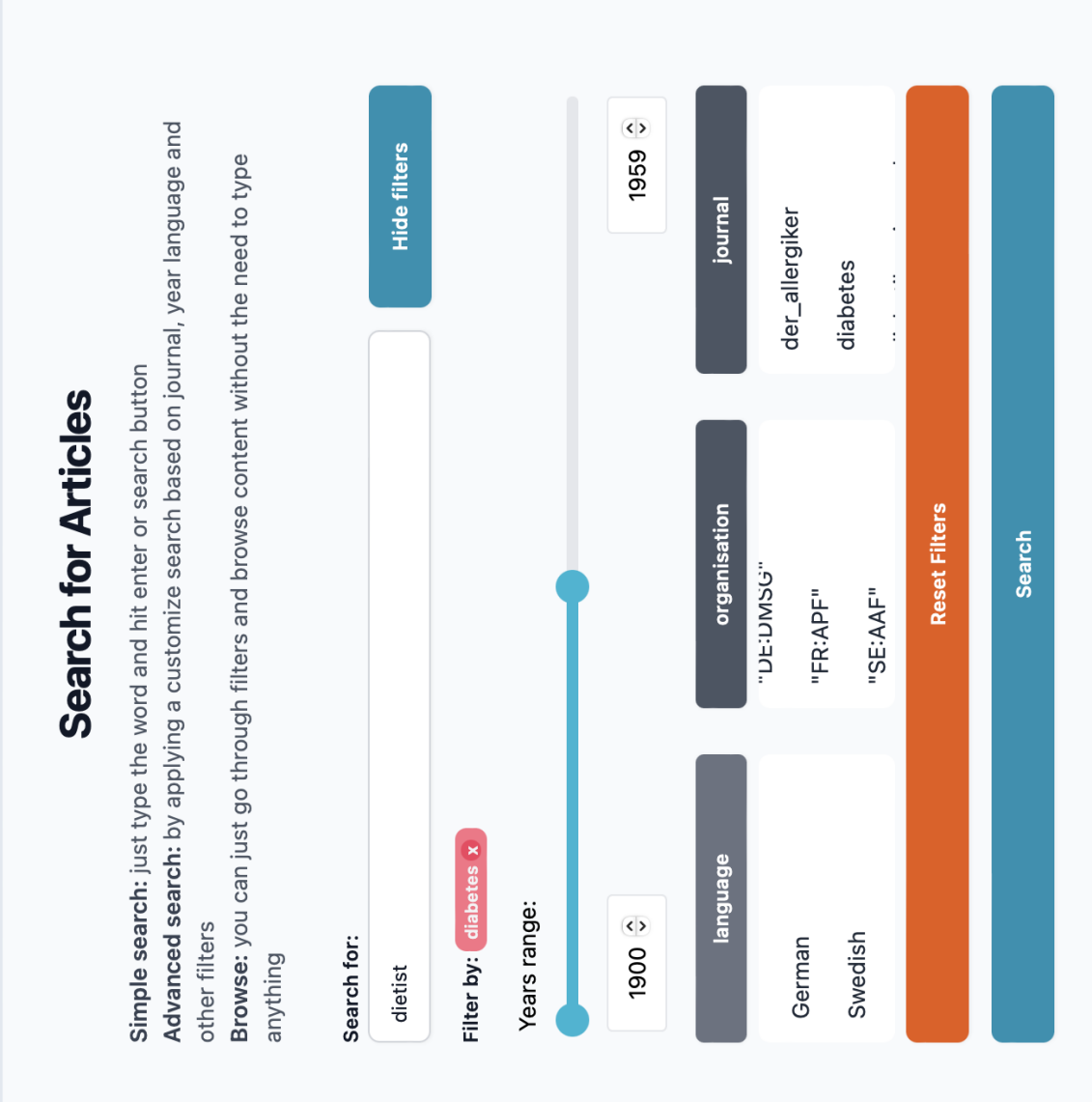

Figure 1: The graphical user interface for searching and filtering the ActDisease database.

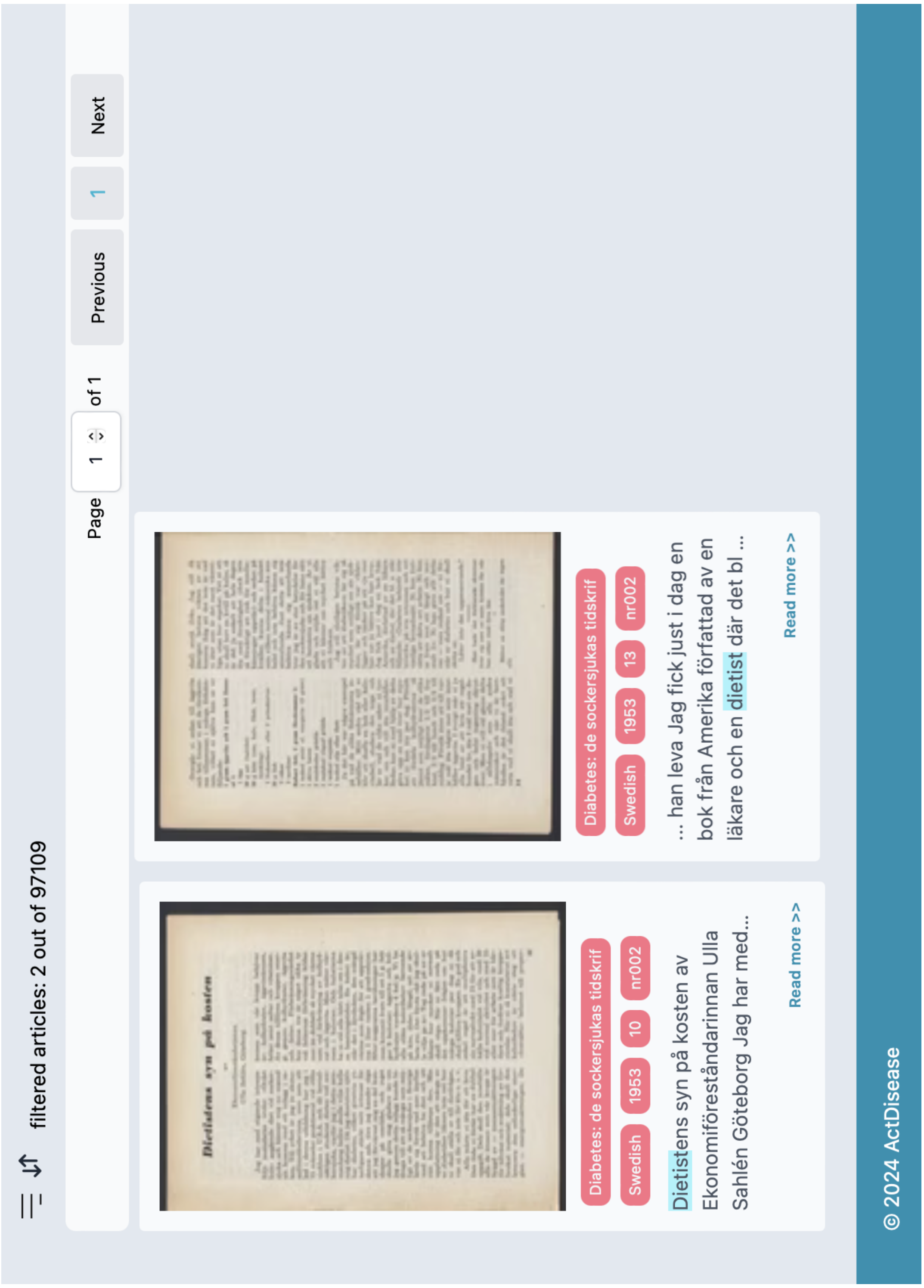

Figure 2: The ActDisease database interface showing the results of searching and filtering.

Figure 3: Timeline for the Swedish Diabetes Association member publications.

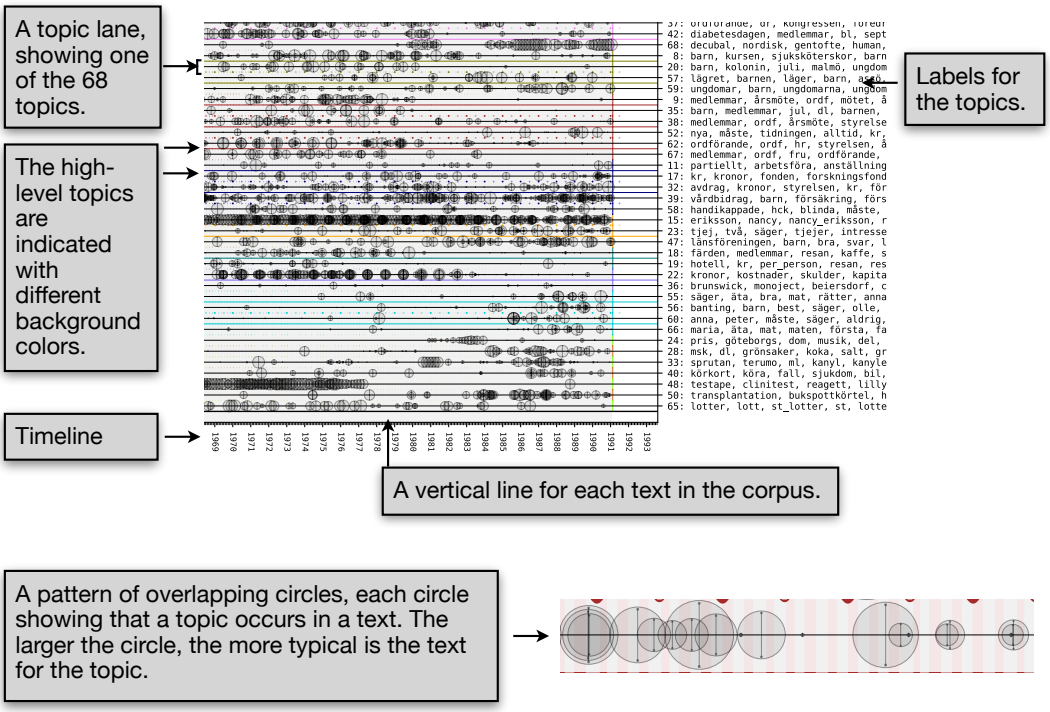

Figure 4: Components of a Topic Timelines graph.

Figure 5: Example of an advertisement, as shown in the user interface for the ActDisease database.