Curating a historical source corpus of 20th century patient organization periodicals

Gijs Aangenendt[1,*], Maria Skeppstedt[2] and Ylva Söderfeldt[1]

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

Abstract

Acting out Disease: How Patient Organizations Shaped Modern Medicine (ActDisease) explores the history of patient organizations in 20th century Europe. By combining traditional historiographic methods with text mining techniques, the project aims to shed light on how patient organizations co-constructed concepts of and management of disease. Part of the project is to digitize print sources and build a digital corpus for historical text mining. The corpus consists of periodical publications from selected British, French, German and Swedish patient organizations, a type of material that poses a number of challenges in scan quality, layout, and lack of consistency. This paper discusses the technical process of building the ActDisease corpus from digitizing patient organization periodicals to OCR post-processing. It touches upon the methodological questions and challenges of curating a corpus of fragmented and heterogeneous historical source material tailored to a specific project.

Keywords

Corpus curation, historical text digitization, OCR processing, patient organizations

1. Introduction

A well-known challenge in writing the history of medicine is that the sources available regarding past experiences of illness, management of disease, and ideas about health overwhelmingly derive from other people than those who were actually sick. Medical literature, health administration files, and patient records all share fundamental limitations as historical sources to how most people felt, thought, and acted around health and illness. Since the 1980s, historians of medicine have debated how to overcome this obstacle, whether it is possible to capture such a thing as "the patient's view" or if a historical patient-as-subject can be said to exist at all [1, 2, 3]. Considering this crucial problem in the field, it is surprising that a large body of printed media produced by and for people living with disease has so far been mostly neglected by historians: Beginning in the late 19th century, clubs and organizations for people suffering from particular illnesses began forming in the US and Europe and by the 1940s, there was a considerable number of such patient organizations with remarkable resources and influence. Contrary to the belief held in most research on patient advocacy, the patient organization as a historical phenomenon hence far predates the 1960s and -70s "new social movements" [4].

huminfra Conference 2024, Gothenburg, 10–11 January 2024.
*Corresponding author.
gijs.aangenendt@idehist.uu.se (G. Aangenendt); maria.skeppstedt@abm.uu.se (M. Skeppstedt); ylva.soderfeldt@idehist.uu.se (Y. Söderfeldt)

Since patient organizations from early on issued printed publications, excellent source material is available for studying their aims, activities, and ideas. Not only do these sources reflect life with a particular disease – for example diabetes – from a viewpoint closer to the actual everyday realities of patients, they also profoundly shaped the experience of living with that illness. In our project, ActDisease, we aim to study how patient organizations co-constructed concepts of and management of disease in the 20th century.

The aim of the project is not to write the history of any particular disease, nor of a specific healthcare system, but of the patient organization as a historical actor. Traditional historiographic methods based on close reading alone therefore do not suffice. Our hypothesis is that illness characteristics are expressed in the aggregated vocabularies of the patient organization periodicals, and hence that their variance across lay and professional contexts, between countries and specific illnesses, as well as their transformations over time can be tracked through text mining techniques.

Other than professional medical literature, however, these sources have hitherto not been available in digital form. Rather than confining ourselves to already digitized material, we are digitizing the periodicals of a selected number of patient organizations. Digitizing the sources within the project enables a higher level of transparency in the methodology. Any corpus curation, including ours, involves a myriad of decisions and steps that shape the material in ways that have consequences for the analysis. By curating our own corpus in collaboration with in-house computer scientists, the historians in the project do not need to rely on selections and technical decisions made by other entities like libraries, publishers, or tech companies, which otherwise threaten to affect the analysis in unknown ways [5]. Nevertheless, our process also involves potential sources of selection bias and unknown distortion, as we will discuss below.

2. Project team

The ActDisease project assembles an interdisciplinary research team which in its current form consists of the PI, a research engineer and a research assistant. In 2024, two history postdoctoral researchers, focusing on Great Britain and France respectively, are joining the project as well as one Digital Humanities postdoctoral researcher focusing on methodological development in the field of historical text mining.

The project is conducted in close collaboration with the Centre for Medical Humanities (CHM) and the Centre for Digital Humanities and Social Sciences (CDHU) at Uppsala University. The main support provided by CDHU is 1) access to a pool of qualified research engineers, with expertise in relevant areas such as Optical Character Recognition (OCR), Natural Language Processing (NLP), webscraping, and front- and backend development, and 2) access to computational resources for data processing, data storage, and data backup. Additionally, the collaboration provides the opportunity to create synergies with other ongoing research projects that CDHU is involved in. For example, the research infrastructure project Communicating Medicine: Digitalisation of Swedish Medical Periodicals, 1781–2011 (SweMPer). Both projects concern medical periodicals and have overlapping technical and infrastructural needs which CDHU can provide. Considering the similarities, the aim is to integrate the ActDisease and SweMPer materials into a shared database. Lastly, the collaboration allows us to make tools and models for text processing and exploration developed within the project available to a wider audience as Huminfra resources.

Table 1
Overview of the ActDisease corpus for Swedish and German patient organization periodicals

Patient organization (Disease)	Periodical	Period (Number of pages)
Sweden:		
De lungsjukas Riksförbund (Lung disease, later also heart disease)	Status	1938-1991 (16 790)
Riksförbundet för sockersjuka (Diabetes)	Diabetes	1949-1990 (8 891)
Riksförbundet för mot astma och allergi (Allergies and asthma)	Allergia	1957-1990 (4 054)
Germany:		
Deutscher Diabetiker Bund (Diabetes)	Diabetiker Journal	1951-1990 (19 324)
Deutsche Multipel Sklerose Gesellschaft (Multiple sclerosis)	MS Gesellschaft	1954-1990 (5 646)
Deutscher Allergiker und Asthmatiker Bund (Allergies and asthma)	Der Allergiker	1959-1985 (2 397)
	Jahresberichte	1901-1972 (8 529)

3. Constructing a corpus

The criteria for selecting the patient organizations aimed at covering as much of the 20th century as possible and to include the main European languages (English, German, French) plus Swedish. We selected for the study the two or three oldest patient organizations from each country that had issued a periodical publication (newsletter, magazine, annual report). The earliest publication started in 1899 and the most recent in 1959. We digitized every selected series from its first preserved issue up until and including the year 1990.

Compared to other datasets typically used for historical text mining, like collections of books, parliamentary print, or newspapers, the material of the patient organizations is relatively small. For Sweden and Germany, the ActDisease corpus consists of periodical publications from three Swedish and three German patient organizations (Table 1). In the near future, the corpus will be extended with periodical publications from British and French patient organizations which are currently being digitized. In total, our corpus will comprise about 150.000 pages but the size of the individual series varies considerably. Some volumes and issues are missing, creating gaps in the series. The format and layout varies greatly between the series and also over time. Furthermore, the patient organization journals contain a very diverse range of texts and images: besides regular articles also advertisements, crossword puzzles, cartoons, charts, lists and much else. Finding ways to meaningfully compare texts of such diverse volume and type, as well as in different languages, is one of the objectives of the project.

The scanned periodicals were delivered as PDF files. A utility script developed by CDHU using ImageMagick was used to extract the image data from the PDF files into individual image files [6]. These image files formed the basis for OCR processing. To be able to integrate the ActDisease and SweMPer materials into a shared database in the future, a joint consistent filename scheme was adopted, including information such as periodical name, year, volume, and issue.

4. Performing OCR

The OCR processing stage posed several methodological questions and challenges originating from the scan quality, evolving design/layout of the periodicals, and the intended uses of the OCR output. As the decisions taken during this stage influenced how the periodicals could be used in the future, these questions and challenges had to be addressed collectively by the research team. For the output, we selected formats that would support traditional historiographic methods, computer-based analysis, as well as the creation of a database: TXT, searchable PDF, and XML.

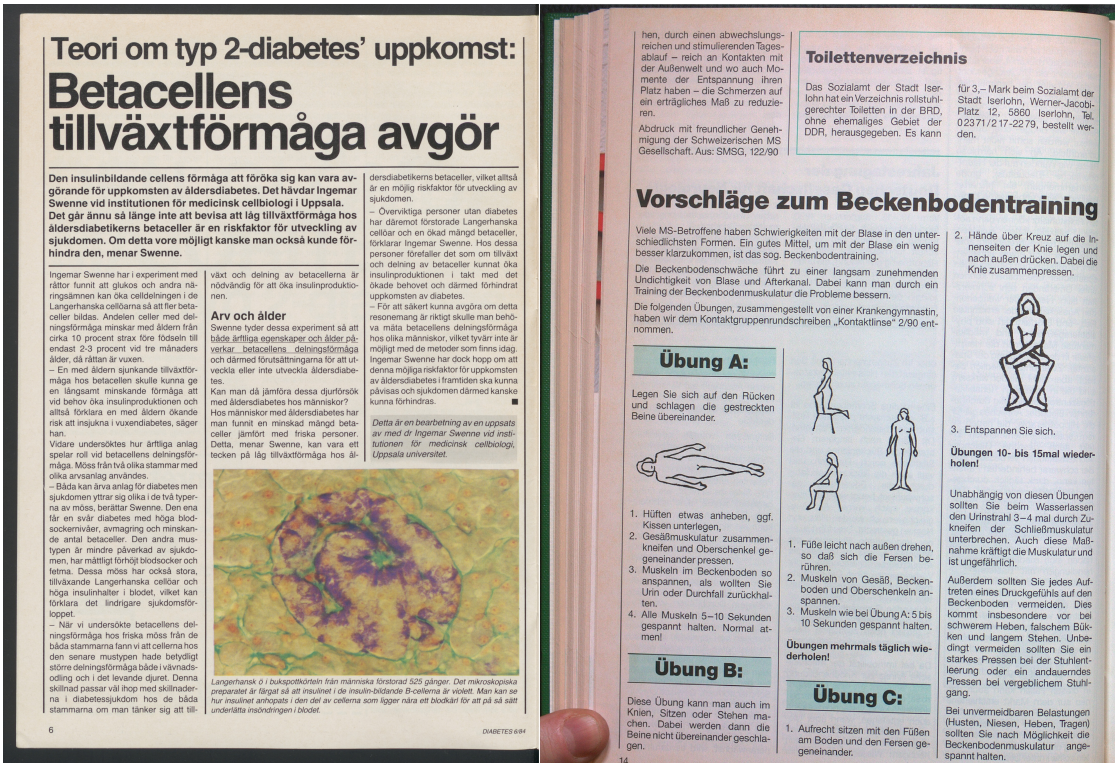

Figure 1: Example pages of the Swedish periodical Diabetes 6:1984 p. 6 and the German periodical Mitteilungsblatt, 149:1990 p. 14.

The periodicals of the Swedish patient organizations were professionally scanned by the University Library of Gothenburg University. The periodicals of the German patient organizations, on the other hand, were scanned with the intention to only be used for close reading. This resulted in a more inconsistent scan quality compared to the Swedish material and presented a challenge for achieving sufficient OCR results. Figure 1 illustrates the difference in scan quality between the Swedish (left) and the German periodicals (right) in terms of lighting, orientation of text lines and paragraphs, and warping of the page. Furthermore, the example pages showcase the complexity of the layout of the periodicals. The layout of the Swedish and German periodicals became gradually more complex over time, with multiple columns, images, illustrations, in-text advertisements, and creative cursive fonts appearing more frequently. Combined with other layout features such as colored backgrounds, fracture script, and text inside images, this added another

layer of complexity to the OCR process.

The quality of the OCR output has a direct impact on the implementation of NLP text mining methods, such as concordance analysis, named entity recognition, and topic modeling. To facilitate subsequent NLP text mining tasks and achieve reliable results, we strived for an OCR quality of 90%. Meaning that of all the words on a page, at least 90% would be correctly recognized [7, 8]. Advanced NLP methods require a good OCR quality not only in terms of correctly recognized words, but also in terms of accurate layout analysis and reading order detection [9]. Due to the complex and inconsistent layout of the periodicals, ensuring the correct reading order for each page was labor intensive and in many cases required judgment calls from the historians. Not only was it impossible to define a set rule for reading order (e.g. left to right, top to bottom), there were also many pages where a "correct" reading order was difficult to determine even for a human reader.

Similarly, making a consistent rule for what would count as a text unit (e.g. an article) and extracting individual texts from the material, similar to the repositories of many scientific journals, proved to be challenging. Texts would regularly be broken up over several nonconsecutive pages or issues, and the distinction between what constitutes, for instance, a text versus a subsection of a text was ambiguous. For that reason, we decided that the page would constitute our base text unit. This decision has the consequence of obscuring the association between words that appear in a multi-page article, and creating an association between words that appear in different texts on the same page. However, this remoteness/closeness between words also represents a material fact that is consistent with the situation when the material was being read by its intended audience. Segmenting the material into individual texts, we concluded, would have created a greater discrepancy between the digital output and the paper original.

ABBYY FineReader Server 14 was used for the OCR processing. For each periodical a separate workflow was designed. Depending on the scan quality, preprocessing steps were included in the periodical's workflow, such as straightening the text lines or deskewing of the page. In order to ensure a good quality, a confidence character threshold was set. If the low confidence character rate of a page exceeded 5%, the image was manually reviewed in the software's verification station. In practice, this meant that between 5 to 10% of the pages were sent to the verification station for review. Here text and image areas were added, removed, and/or redrawn, OCR errors corrected, and the reading order checked. The confidence character rate of a page was not always accurate. The software could be confident about interpreting a character correctly while it in reality did not, or the other way around, be insecure about a character while it was accurately recognized. Overall, the performance of the OCR software on the Swedish and German material was good on a character and word level. However, the layout and reading order detection was sometimes incorrect. Given the complex page layout combined with the inconsistent scan quality of the German periodicals, this was not surprising. Combining multiple preprocessing methods in one workflow seemed to negatively impact the accuracy of recognized reading order during our tests. Therefore, we opted to only perform basic preprocessing if it significantly improved the quality of the OCR.

5. OCR post-correction

To achieve an estimate of the OCR quality independent of ABBYY's own estimation — as well as to correct some of the errors still remaining in the corpus — we implemented a word-list based approach for OCR post-correction based on previous work by Thompson et al. [10]. We based the implementation on a spellchecker [11], which we extended with additional functionalities such as compound splitting. With Diabetes as a first test case, we configured the post-correction to only suggest corrections that 1) had an edit distance of one from the original, unknown word, and 2) were more frequent in the OCR text output than the original word. As reference data for the spellchecker and for the OCR quality estimation, we used word-lists gathered from corpora and from medical terminology, which we manually extended with correct words that were flagged as unknown by the spellchecker. Finally, we manually verified the spelling corrections suggested by the spellchecker before replacing the original word as it had been interpreted by the OCR software. Similar to the estimations made by the ABBYY OCR software, our word-list based quality measurements indicate a consistent and low error rate, with only 382 of 8 991 pages exceeding an error rate of 5% unknown words[1].

6. Conclusion

Tailoring source corpora to the scope and questions of a specific project opens up a wider field for using text mining in historical research. It also allows a greater transparency in the research method since the decisions made in the digitization and post-processing stage are consequential for the analysis. In our corpus of 20th century patient organization periodicals, we achieved a high OCR quality on the character and word level which could further be improved with spell checking, but layout parsing presented more challenges. The software's difficulty in determining the correct reading order derives in part from an inherent ambiguity in the layout. Our material is unusually diverse, complex, and fragmented from a text mining standpoint, but typical for historical source materials. Using digital methods more broadly for historical research will require improved methods for handling structurally inconsistent and fragmented materials.

Acknowledgements

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.

[1] Our implementation of Thompson et al.'s [10] algorithm for OCR post-correction and quality estimation, as well as instructions for how to practically apply it, will be made available through CDHU as a Huminfra resource. It is available at CDHU's GitHub: https://github.com/CDHUppsala/Customised-OCR-Correction.

References

[1] R. Porter, The Patient's View: Doing Medical History from below, Theory and Society 14 (1985) 175–198. URL: http://www.jstor.org/stable/657089.

[2] L. Jordanova, The Social Construction of Medical Knowledge, Social History of Medicine 8 (1995) 361–381. URL: https://academic.oup.com/shm/article-lookup/doi/10.1093/shm/8.3.361. doi:10.1093/shm/8.3.361.

[3] F. Condrau, The Patient's View Meets the Clinical Gaze, Social History of Medicine 20 (2007) 525–540. URL: https://doi.org/10.1093/shm/hkm076. doi:10.1093/shm/hkm076.

[4] Y. Söderfeldt, The Truth Within: Making Medical Knowledge in the Hay Fever Association of Heligoland, 1899–1909, Isis 112 (2021) 531–547. URL: https://www.journals.uchicago.edu/doi/10.1086/715653. doi:10.1086/715653.

[5] M. Fridlund, Digital history 1.5: A middle way between normal and paradigmatic digital historical research, in: M. Fridlund, M. Oiva, P. Paju (Eds.), Digital Histories: Emergent Approaches within the New Digital History, Helsinki University Press, Helsinki, 2020, pp. 69–87. doi:10.1145/90417.90738.

[6] CDHU, cdhu ocrscripts, https://github.com/CDHUppsala/cdhu-ocrscripts, 2023.

[7] D. van Strien, K. Beelen, M. Ardanuy, K. Hosseini, B. McGillivray, G. Colavizza, Assessing the Impact of OCR Quality on Downstream NLP Tasks:, in: Proceedings of the 12th International Conference on Agents and Artificial Intelligence, SCITEPRESS - Science and Technology Publications, Valletta, Malta, 2020, pp. 484–496. URL: http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0009169004840496. doi:10.5220/0009169004840496.

[8] M. J. Hill, S. Hengchen, Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study, Digital Scholarship in the Humanities 34 (2019) 825–843. URL: https://academic.oup.com/dsh/article/34/4/825/5476122. doi:10.1093/llc/fqz024.

[9] C. Neudecker, K. Baierer, M. Gerber, C. Clausner, A. Antonacopoulos, S. Pletschacher, A survey of OCR evaluation tools and metrics, in: The 6th International Workshop on Historical Document Imaging and Processing, ACM, Lausanne Switzerland, 2021, pp. 13–18. URL: https://dl.acm.org/doi/10.1145/3476887.3476888. doi:10.1145/3476887.3476888.

[10] P. Thompson, J. McNaught, S. Ananiadou, Customised OCR correction for historical medical text, in: 2015 Digital Heritage, IEEE, Granada, Spain, 2015, pp. 35–42. URL: http://ieeexplore.ieee.org/document/7413829/. doi:10.1109/DigitalHeritage.2015.7413829.

[11] T. Barrus, pyspellchecker, https://pyspellchecker.readthedocs.io, 2018.