1 Introduction

Historians are trained to critically interpret the past. To do this they are instructed in a variety of archival and writing skills, as well as critical thinking, and are taught a research workflow in which searching for primary sources, verifying their authenticity, and undertaking close reading to understand, analyze, and interpret them, are all fundamental to writing critical and comprehensive reflections of past events.

Although, for historians, search is just one aspect of their repertoire, it is a vital skill and one in which they become very efficient. Historians have always relied on this competency: searching for people (e.g. librarians and archivists) or within archive catalogs to direct them to relevant material, as well as searching or browsing through primary sources in order to find useful information. Without search, there are no sources or text passages to be read and interpreted. Of course, the term “research” itself is a derivative of the word “search.”¹ In other words, search is where all historical investigations start – hence why search – as opposed to other skills of the historian such as analyzing and interpreting – is central to this chapter.

Since the advent of digital history and digital humanities in the late twentieth century, the historian’s “traditional” workflow has often been juxtaposed with these “new” forms of research practices and tools. Digital history has become associated with terms such as big data, algorithms, programming languages, text mining, topic modeling, network analysis, etc. Even now, some humanities scholars can be wary of incorporating certain approaches promoted within digital humanities. They have come to associate digital history with

“black boxes,” uncritical research outcomes, and computational approaches that replace or downplay the defining research skills of historians. However, this does not have to be the case.

At the same time, within digital history the concept of search has become an even more crucial feature of the historian’s research repertoire and workflow, especially when big data is involved. Far from supplanting historians’ original research practices, digital history can provide additional or extended forms of search to aid historians in the exploration and analysis of source material. Digital tools and approaches do not have to be something utterly foreign: text mining, topic modeling, and other derivates are in essence “different species of search.”

In the following paragraphs I want to challenge the reader to think about how essential search is, what the benefits and drawbacks of different digital search tactics are, and what the future of digital search might involve. I draw from my own educational background as a “traditionally schooled” historian, experimenting with digital sources and tools that enable me to use digital search in order to analyze the history of psychiatry from a transnational point of view. To contextualize this, I first explain my research project, as well as the meaning of search itself. I then focus on the different stages of the historian’s research process – including searching for sources in order to find material with which to answer research questions, searching for tools in order to manage the exploration of large data sets of collected sources, and searching for relevant content within our sources that will facilitate reading, analysis, and interpretation – and the importance of search in each of these.

2 Search: A means to an end

During my time as an MA history student I developed an interest in the history of psychiatry and transnational history. For my PhD project I wanted to pursue these domains on a bigger scale than I had done before – to leave behind any form of nationally contained histories and instead investigate how psychiatric

2 Statement by Lincoln Mullen, Associate Professor at George Masson University, online meeting June 25, 2020.
knowledge had circulated throughout Europe during the nineteenth and early twentieth centuries.\(^5\) To answer my research questions I studied the main psychiatric journal of each of five different countries – Belgium, France, the Netherlands, Germany, and the United Kingdom – between 1843 and 1925 (an 82-year period).\(^6\) Together, these sources amounted to a substantial corpus of over 460 volumes and more than 300,000 pages to investigate.\(^7\)

Scale is undoubtedly one of the main challenges with transnational research. Digital history, and more specifically digital search, seemed at first sight to offer easy solutions to this problem.\(^8\) The transnational and digital turns are becoming more and more intertwined due to source digitization which facilitates virtual cross-border research, as well as the growing possibilities of the search box which make (transnational) research possible at a pace and range that was not feasible before.\(^9\) As Putnam aptly states, “Digital search has become the unacknowledged handmaiden of transnational history.”\(^10\)

---

\(^5\) Some key questions in my research are: How did the transnational sphere influence thinking about and the reception of psychiatric concepts and practices? What kinds of negotiations took place in psychiatric circles before certain information or ideas were deemed important, true, false, or even useless? Why did or didn’t knowledge transfers succeed?.

\(^6\) The reason I chose these countries was mostly a pragmatic one due to my knowledge of the languages spoken in these countries, but also already being familiar to some extent with the history of psychiatry in these specific countries. The reason for choosing this timeframe was based on two parameters. Firstly, 1843 is the earliest date that an issue of one of the journals under study appeared. Then 1925 was chosen as an end date because the journals were harder to acquire in a digital format after this date. The journals under study were: Bulletin de la Société de Médecine Mentale de Belgique (Belgium), Annales de la Société Medico-Psychologiques (France), Psychiatrische Bladen (the Netherlands), Allgemeine Zeitschrift für Psychiatrie und psychisch-gerichtliche Medicin (Germany), and Journal of Mental Science (the United Kingdom) – although their titles have changed over time.

\(^7\) Referring to my sources as a “substantial corpus” is of course relative in terms of the amount of data that is used in other historical research, such as the newspaper Impresso project, which contains over 5,445,822 pages (see: Matteo Romanello and Maud Ehrmann, “What’s in Our Corpus?,” impresso blog, January 23, 2020, accessed August 2, 2021, https://impresso-project.ch/news/2020/01/23/state-corpus-january2020.html) or the amount of data that is often used in computational sciences.

\(^8\) This does not mean that transnational research did not exist before the introduction of the search box and digital repositories, but it was much more time-consuming to write and expensive to investigate, often resulting in specifically selected examples rather than trying to study a particular subject in its entirety. Lara Putnam, “The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast,” The American Historical Review 121, no. 2 (2016): 382–3 and 394.


“Search” is not an end goal; it is always a means to an end regardless of whether we are talking about search in its analog or its digital form. The former consists in most cases of skim-reading page after page until a certain title, passage, phrase, or word catches our eye – often almost as if by accident – in our search for relevant material.\footnote{Regarding the difference between manual browsing and digital searching, see Bob Nicholson, “The Digital Turn,” Media History 19, no. 1 (2013): 59–73; Hieke Huiistra and Bram Mellink, “Phrasing History: Selecting Sources in Digital Repositories,” Historical Methods: A Journal of Quantitative and Interdisciplinary History 49, no. 4 (2016): 220–9; and Adrian Bingham, “The Digitization of Newspaper Archives: Opportunities and Challenges for Historians,” Twentieth Century British History 21, no. 2 (2010): 225–31.} Depending on the type of source this can also be facilitated by searching through physical tables of contents or indexes. In essence this is what can be called a top-down approach to searching, whereas digital search applies a bottom-up approach dominated by the search box.\footnote{Nicholson, “The Digital Turn,” 66–7.}

But if digital search is a bottom-up approach, doesn’t that mean that it is something different from what historians are taught? Yes and no. Instead of phrases or words catching our eye as we read for hours, they now come to us almost instantaneously via digital search.\footnote{This is not to say that this is without its problems. Many scholars have warned about the loss of context in these cases and the idea that what is in fact scarce now looks prominent or abundant.} The reason our eyes pick up on certain passages within a source when browsing manually is because we consciously or unconsciously build a list of words in our minds around the topic we are studying. For example, to explore the use of mind-altering substances in psychiatry we would pay attention to words like alcohol, morphine, addiction, or wine. When we search digitally, we still use our same background knowledge and word lists regarding this topic, only now we enter them into a digital interface. Suddenly, digital search seems less alien.

On the other hand, there are certain aspects of digital search that we need to be careful about, although through critical reflection and transparency potential issues can be mitigated. When it comes to historical research, one of its challenges and even dangers is its seeming simplicity. We all know the search box and use it daily, either in our personal lives or for our professional activities. Most of the time we do not think about how we use it or how it works, and do not take into account the variety of ways in which digital search can trigger different or skewed results – especially arising from the many forms that search and the search box can take.\footnote{Tim Hitchcock, for example, has warned scholars about this on multiple occasions. See, for instance, “Lecture Tim Hitchcock – Beyond Close and Distant Reading: Recording and Interview,” June 18, 2019, accessed April 13, 2021, https://www.c2dh.uni.lu/data/lecture-tim-}
3 Searching for the perfect digitized source

When we apply digital search, the first phase in which we do this is while searching for relevant digital sources. This often means searching with the Google search engine to, in my case, find different journals to investigate, or do a keyword search within the digital repositories of archives and libraries.

But keyword search is more than just typing words in a search bar. Depending on the platform, a variety of options can be offered. These can include “basic” options such as introducing a date range, or placing limits on titles, genres, source types, or places of publication, as well as using Boolean operators (such as “AND,” “OR,” and “NOT”) between keywords or using multiword expressions. But there are many other forms of search: fuzzy search, proximity search, the use of wildcards and query auto-complete options. These types of search can be further optimized and improved via the use of correction after the optical character recognition (OCR) process (post-OCR correction), named entity recognition, entity linking, sentiment analysis, or topic modeling. Many of these more advanced features are less integrated into the interfaces of online repositories and, if they are present, are often hidden.

Because we are talking about searching for and identifying digital sources for our research it is also important to reflect on the digital sources themselves as, aside from the algorithms applied in a search environment, their quality has a tremendous impact on search functionality – as well as on the displayed results we will later have at our disposal. This is as true for the initial task of locating digital sources as for the application of search tactics within sources.

The quality and accuracy of a digital source is determined by the factors that help transform the analog source into a digital version. In this regard, Owens made the accurate observation that “all digitized objects are surrogates for the originals.” This transformation process can be captured in three stages: scanning

hitchcock-beyond-close-and-distant-reading-recording-and-interview (“Recording of the conference” – see especially from 13 minutes 13 seconds onward).


16 Ehrmann, Bunout, and Düring, “Historical Newspaper User Interfaces,” 12.

17 Ehrmann, Bunout, and Düring, “Historical Newspaper User Interfaces,” 14.

18 Ehrmann, Bunout, and Düring, “Historical Newspaper User Interfaces,” 12.

the source; optimizing the source to enable more and better search functionalities; and the online consultation or downloading of sources.

Firstly, scans can be made by high definition cameras, (semi)automatic book scanners or overhead scanners (with or without the use of a V-shaped book cradle). All make the digital version of a source somewhat disparate from its original and can lead to visual and analytical discrepancies between the original and its digital copy, as well as between digital versions. This can have a lasting impact on the different search capacities that can be integrated. The severity of this impact depends on the accuracy and completeness of a source (missing, skewed, or badly scanned pages), its readability by humans and machines, and its aesthetics and visual representation (e.g. the difference between black-and-white, grayscale, or multitone scans, or the (dis)use of thumb-removal software).

In a second stage, the sources are optimized to maximize the search functionalities. The three most important processes we find here are: creating single pages out of double scanned pages, which improves the OCR accuracy; applying OCR; and applying post-OCR corrections through software. Much can be said about OCR software and protocols, but what is important to note is that (re)search with digitized sources relies tremendously on the recognition of letters and words within a corpus. When sources are not properly optimized this can lead to discrepancies between the material that can be found within digital repositories and the search hits within a source. The impact that poor scanning can have on the readability of the source by a machine (due to inaccurate OCR), but also that the source becomes almost unreadable for the researcher too, making even analog search within this digital source less efficient.

Lastly, these digital sources are stored on personal hard drives or the servers of archives and libraries and, in the case of the latter two, made accessible via an online repository. Depending on how carefully the previous steps were carried out, digital sources can be found more or less easily. There are still some online platforms that do not apply OCR when scanning their source

---

material, making search possible only via the metadata (e.g. title, year) that are provided by the institution that stores them.

Due to the mass digitization of sources, different copies of a single source can be found on the internet. This becomes especially visible in repositories such as HathiTrust and Archive.org and can potentially lead to different research results, depending on the accuracy and quality of each of these copies, which in turn determine the degree of search that is possible. It is not always clear which copies are better and should be preferred over other digitized copies. There are often no clear ways to notify providers about the discrepancies that are sometimes found within a digitized source either – nor to ask them to rectify this.

This whole transformation process lays bare two important shortcomings of digitized corpora: the historian’s dependence on the input and diligence of others in their search for sources, and the efficiency of search and search results. We often rely on third-party data providers such as libraries or online archives – and the companies (e.g. Google) they work with – in order to deliver and provide complete and well-scanned historical material. Where it was previously just the historian and stacks of physical sources under their control, there is now an intermediator standing between the historian and the sources in the form of those who scan and provide the material, as well as the machines used to make those scans. Of course, some mediation also takes place between the historian and the archivist, as the latter often makes a selection of which documents are preserved and which are not. Likewise, intermediation has in some cases also become less extreme. This is, for example, noticeable in the online access of archival catalogues.

4 Searching for tools: A process of trial and error

“Searching for tools” does not mean the same as searching for sources in repositories, developing search tactics, or exploring search results. Nevertheless, searching for suitable tools is important when we want to apply digital search. Many of the search functionalities mentioned earlier are also found in standalone (re)search tools. The range of possibilities, algorithms, and online and standalone tools that offer all or some of these functionalities seems almost endless.21 However, the internal mechanisms and modi operandi of these tools

21 For a broad introductory overview see, for example: Shawn Graham, Ian Milligan, and Scott Weingart, Exploring Big Historical Data: The Historian’s Macroscope (London: Imperial College Press, 2016).
are not always explained, or they are difficult to use and understand for inexperienced users who are unfamiliar with this multitude of search functionalities and thus not able to use them properly.

This takes us to the dreaded “black box” of the digital humanities, which can become problematic during the use of digital search if we are not transparent about – and careful and consistent with – the research practices applied. This amalgam of tools is not necessarily a one-size-fits-all solution for each and every research project, although that can be a common misconception. Below I outline some of the tools I experimented with to find a form of digital search that fitted my project and research questions, and my research workflow as a historian – a tool that was also understandable to me. Corpus linguistics, text mining, and more specifically keyword search and topic modeling, were the search practices and techniques I used to digitally search for relevant content in order to be able to analyze the circulation of psychiatric knowledge later on in my research. The search tools I explored were Voyant Tools, MALLET, AntConc, and histograph.

Voyant Tools is a “web-based reading and analysis environment for digital texts” that is “designed to facilitate reading and interpretive practices for digital humanities students and scholars as well as for the general public.” It is an example of text mining via “simple” keyword search – but, although it has an important goal in mind, I didn’t find the tool suitable for my own research.

First of all, the web application often reacted extremely slowly or crashed due to the large amount of data I had. Secondly, as with many other search and explorative tools, Voyant only shows plain text versions of the data, while it could also be valuable to see the original scans next to these. Thirdly, although Voyant Tools offers 28 different ways to explore and search through a corpus, this breadth of options is overwhelming for a beginner. Furthermore, not all sub-tools allow the user to switch between the visualization and the text file. Lastly, many of the visualization options – let’s call them “visual search” options – are nice to look at but, for detailed search tactics and source analysis, they will scarcely provide the researcher with the information they are looking for (Fig. 1). This is a problem very common with visualizations in the humanities, as “beautiful” graphs are often bad representations of data or easily open to misinterpretation.

---

24 Although running Voyant Tools on your own computer is possible, I was not aware of this feature before I abandoned this tool.
Fig. 1: Five different ways of corpus exploration in Voyant Tools: (1) a table view of terms that appear in the corpus; (2) a way to explore high frequency terms and their collocates; (3) a line graph with the distribution of a word’s occurrence across a corpus; (4) a “visualization of the terms in a document that includes a weighted centroid of terms and an arc that follows the terms in document order”; and (5) a visualization of the frequency and distribution of terms in a corpus. Source: https://voyant-tools.org/. © Stéfan Sinclair and Geoffrey Rockwell (cc 2021). Data processing and visualization by Eva Andersen.
Aside from the use of keyword search, where a certain amount of background knowledge is required from the researcher, another tool that allows digital search is topic modeling. In a very simplistic manner, a computer algorithm tells the researcher which topics or subjects are present in a certain corpus. Topic modeling works as follows: a document (e.g. a book, journal issue, or article) consists of a collection of words and via a statistical process the computer classifies these words into sets of words that occur frequently together, forming different topics in the process.

One of the topic modeling tools that I briefly explored was the Topic Modeling Tool (TMT), which provides a graphical user interface (GUI) for MALLET.25 MALLET, which is used via the command line, is a topic modeling algorithm that is very frequently used within the humanities.26 I only conducted a few experiments with the GUI for MALLET because the tool represented a black box for me. The exact mechanics behind the algorithms and the different settings and options that could be selected and implemented were not clear to me. In cases like this especially, continuing to use this kind of search tool would have led me to make errors in my analysis and conclusions later on.

Aside from these explorations I also began to work with AntConc and historigraph. Both tools were used extensively during my (re)search process. In Section 5 I highlight and explain the different digital search tactics I applied in order to show their drawbacks and benefits.

5 Applying search tactics to locate what to read

The main reason I made use of various different forms of search was to overcome the obstacle of the overabundance of source material that I had acquired. This overabundance was often problematic to my starting to perform a valuable and thorough analysis of my sources. Without computational support it was not only difficult to search for and provide answers to specific research questions but also to, for example, locate interesting and useful subjects that could serve as case studies.

So how to search this goldmine of psychiatric journals? How could I harness digital search to get control over the sheer volume of my sources? These questions were a constant concern. Without a digital search tactic, I could not start to carry out this essential step of my research workflow as a historian. Without it I would not be able to close-read crucial parts of these journals for my analysis of psychiatric knowledge circulation across Europe.

Historians Damerow and Wintergrün have demonstrated the importance of having full control over a corpus even within “a digital framework” because “historical research relies on trust in its sources.” This trust in sources is a precarious balancing act for historians. How can we find and trace knowledge circulation in these substantial corpora? Where and how should we start distant reading, and later on close reading? How can we find relevant information for close reading? How do we zoom in and out of the material? How accurate is the output of the search tools? These were some of the challenges I faced in seeking one or multiple search tactics. As will become clear in this section, keyword search would form a major part of my research tactics but would take on different forms.

5.1 AntConc

AntConc is an off-the-shelf application developed in 2014 by Laurence Anthony. Its goal is to make textual analysis and explorative research of text files easier and more manageable. The tool can create concordance tables, n-gram clusters, and collocations, among other outputs. To make my use of this tool more explicit, and to contextualize it within the scope of the transnational history of psychiatry, I framed my search exploration with AntConc around the non-restraint system that came into vogue during the nineteenth century. During this time, debates were held about the (un)suitability of using mechanical restraints such as straitjackets and iron cuffs on patients. Laying bare the

29 Some were of the opinion that this formed a necessary part of therapy, as well as a practical element necessary to keep control over a large number of patients. Others were of the opinion that this had no therapeutic value at all and that patients did not have to be confined in this manner, but instead should be able to walk around freely and allowed to enjoy the outside air, games, or working in the kitchen or gardens of the asylum.
non-restraint debate across time and space by manually combing through all the journals to search for relevant articles would have been an immensely time-consuming task. However, by using AntConc as a search tool, in combination with close reading, this became more feasible.

5.1.1 Concordance plots and the vitality of keyword lists

The main feature of AntConc that I used to search through my corpus was the concordance plot. This component shows concordance results plotted in what Laurence Anthony calls a “barcode” format (Fig. 2). This allows you to see the position of one or multiple search terms in different documents in an abstract representation. Each line in the barcode visualization (distant reading) is clickable and brings forward the uploaded text file for close reading, highlighting the selected keyword(s). In the case of my research about non-restraint I always worked with multiple keyword lists, as this search tactic made tracing relevant spots within the corpus (keyword clusters) easier and more consistent. Below I explain why this was the case and why it can be a useful search strategy.

The building of keyword lists proved essential with this form of text mining. As a point of departure, I used the terms “non-restraint” and “mechanical restraint,” two phrases that typified the core of the debate. However, this did not capture all places where the debate was mentioned. It is important that the researcher already has some understanding of the subject at hand in order to make decisions about the terms that will be included (e.g. knowing which terms were customary). In a second stage I added other keywords that represented these restraint systems, such as “padded room” and “straitjacket.”

Due to the use of corpora in multiple languages, I compiled a list of terms associated with restraint and non-restraint for each language. This was accomplished by, on the one hand, translating already-known terms to other languages, but also by alternating between distant and close reading: examination of specific sections within a source revealed variations of word use within each language. This is a strategy that has also been highlighted by Berridge et al.32

31 See AntConc Help file at Anthony, “AntConc Homepage.”
Furthermore, I took the different spelling variations of keywords, some due to OCR mistakes, into account where possible (e.g. no restraint, no-restraint, non-restraint, non\^restraint).

Compiling these keyword lists needs to be done thoroughly, as the creation of too limited or too generic or broad a list can create its own problems. Not considering one or multiple keywords can have an impact on the output results of search queries, potentially misleading the researcher. This became tangible while analyzing the German psychiatric journal. I had started out with a limited set of keywords for this particular language, due to my limited knowledge of German. But by translating some of the terms found in the other corpora while combining this with close and distant reading, relevant sections within the corpus were highlighted more distinctly and gave a more concrete image of relevant starting points for further corpus exploration (Fig. 2).

While sparse keyword lists can miss relevant spots in a corpus, the use of words that are too generic can clutter the results and create a mass of data that is not easily processed, as other research has also shown.\(^{33}\) To give an example: the French word “\textit{cellule}” could either refer to an isolation cell or human cells. The word “\textit{restraint}” could refer to non-restraint, mechanical restraint, or emotional/behavioral restraint, hence why I opted to use specific words that would not be ambiguous in their use (e.g. isolation cell and \textit{restraint absolu}). I used the same search tactic for zooming in, gathering and extracting information about the editors and editorial decisions, or references to international conferences, from the journals.

A key drawback of AntConc was that, although relevant sections became easier to spot, reading these sections was less straightforward due to its use of text files only: a representation of the source that does not correspond to the original from an aesthetic or visual point of view. Unstructured text files are not always efficiently readable for the human eye. In order to make close reading possible I was obliged to switch between AntConc’s visualization, the text files, and the original PDF documents of the sources – the latter to stay as close to

Fig. 2: Frequency distribution of words in AntConc related to non-restraint in the German *Zeitschrift für Psychiatrie*. On the left, my corpus of terms was not yet complete (e.g. it omits words related to *Zwangsmittel* and *Beschränkung*), as opposed to the image on the right with a more extended list of keywords. Notice how the number of hits and relevant places within the corpus changes. See, for example, plot 1: on the left, 12 hits; and the same plot on the right, 65 hits. 2022. © Eva Andersen.
reading the original source as possible. This was a time-consuming workflow which would be improved by using the histograph web app.

The search tactics with keyword lists used in AntConc require substantial background knowledge of a subject in order to thoroughly study it via the barcode visualization. This therefore omits many other topics that stay hidden from the researcher. Via other search tactics (e.g. topic modeling), this can be overcome to a certain extent.

5.2 Histograph

Although the use of AntConc solved some problems, the sheer volume and diversity of the corpus still posed challenges: How can a historian find and trace relevant information to analyze the evolution of specific ideas throughout large corpora? The use of off-the-shelf applications could only go so far. This stimulated a collaboration between me and my colleagues at the C\textsuperscript{2}DH.\textsuperscript{34} This cooperation, which included major brainstorming sessions about data quality and inconsistently digitized corpora, as well as the nature of the research project, made sure that the search workflow could stay as close as possible to my own research process.

As a result of this collaboration I did not have to adapt to the constraints of a specific tool, as is often the case, but rather vice versa. This approach simultaneously provided me with a better understanding of the technical processes operating in the background, avoiding the black box effect. The result was a tool for corpus exploration modeled on an earlier version of histograph. Initially developed to provide “graph-based exploration and crowd-based indexation for multimedia collections,” through which related documents could be discovered via filtering entities, date ranges, and document types, histograph also reveals relationships between people and keeps track of relevant documents.\textsuperscript{35} My

\textsuperscript{34} For a more detailed excursion into the necessity of collaboration between historians and computer scientists, as well as for all technical details about the processes and algorithms used, see: Eva Andersen et al., “How to Read the 52.000 Pages of the British Journal of Psychiatry? A Collaborative Approach to Source Exploration,” Journal of Data Mining and Digital Humanities (Histoinformatics), 2020.

colleagues adapted the first version of histograph to fit my particular research and sources – e.g. adding topic modeling and visualizations to maximize the search functionalities.

5.2.1 Establishing an optimal way to search content

As seen earlier, historical corpora are often unstructured and irregular due to poor OCR quality and textual errors, as well as the lack of a regular volume structure which, for example, makes the detection of individual articles within a corpus extremely difficult. Some preprocessing steps were required to make our exploration tool useful. Firstly, this included choosing a logical boundary unit – in this case, at the page level (one page = one document). Choosing this boundary also meant that the structure of the documents would be the same across the multiple corpora. A second preprocessing step involved removing stop words and maximizing the use of content-bearing pages.

Instead of relying on concordance plots, collocations or n-grams, we used topic modeling to enable more control over the corpora and the search functionality. This enabled me to discover which topics were covered in the journal, where, and to what extent – which was important for being able to select relevant parts of the corpora for close reading.

We used non-negative matrix factorization (NMF)\(^{36}\) instead of the latent Dirichlet allocation (LDA)\(^{37}\) that is more often used within humanities research. This approach was chosen because instead of specifying three parameters, as is the case with LDA, only two needed to be specified (the number of topics and the number of words in a list). Furthermore, when applying LDA the words per topic will change every time the program runs over a set of documents. With NMF this is not the case and thus provides better topic stability from a historical point of view, making my (re)search more consistent.

---

To study psychiatric phenomena across a wide timespan my colleagues generated two kinds of topic: “window topics” – the standard calculation of X topics (where X = 10 to 20 in my case) for each year of a corpus (Fig. 3) – and “dynamic topics” – less standardized, but allowing better spotting of the development of psychiatric topics and their (vocabulary) variations through time. A good example is the development of general paralysis (GP) as a psychiatric disease in the nineteenth century and its connection with syphilis, as well as with the technical developments that took place within medicine. GP was a mental disease in which patients slowly lost all mental and motor functions – including total loss of speech, writing abilities, and movement – often paired with hallucinations and dementia. No cure existed throughout the nineteenth century and for some time into the early twentieth century, until penicillin was discovered and mass produced from the 1940s. During the time period under consideration, a discussion surfaced regarding syphilis as a possible cause of GP.

<table>
<thead>
<tr>
<th>Rank</th>
<th>1876_01</th>
<th>1876_02</th>
<th>1876_03</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>asylum</td>
<td>case</td>
<td>morphia</td>
</tr>
<tr>
<td>2</td>
<td>insane</td>
<td>brain</td>
<td>dose</td>
</tr>
<tr>
<td>3</td>
<td>patient</td>
<td>left</td>
<td>mania</td>
</tr>
<tr>
<td>4</td>
<td>medical</td>
<td>cell</td>
<td>injection</td>
</tr>
<tr>
<td>5</td>
<td>report</td>
<td>side</td>
<td>sleep</td>
</tr>
<tr>
<td>6</td>
<td>hospital</td>
<td>nerve</td>
<td>excitement</td>
</tr>
<tr>
<td>7</td>
<td>case</td>
<td>vessel</td>
<td>sedative</td>
</tr>
<tr>
<td>8</td>
<td>lunatic</td>
<td>paralysis</td>
<td>vomit</td>
</tr>
<tr>
<td>9</td>
<td>attendant</td>
<td>frontal</td>
<td>patient</td>
</tr>
<tr>
<td>10</td>
<td>superintendent</td>
<td>centre</td>
<td>hypodermic</td>
</tr>
<tr>
<td>11</td>
<td>association</td>
<td>general</td>
<td>case</td>
</tr>
<tr>
<td>12</td>
<td>number</td>
<td>part</td>
<td>hour</td>
</tr>
<tr>
<td>13</td>
<td>treatment</td>
<td>muscle</td>
<td>night</td>
</tr>
<tr>
<td>14</td>
<td>appoint</td>
<td>motor</td>
<td>acute</td>
</tr>
<tr>
<td>15</td>
<td>class</td>
<td>disease</td>
<td>effect</td>
</tr>
<tr>
<td>16</td>
<td>officer</td>
<td>convolution</td>
<td>subcutaneously</td>
</tr>
<tr>
<td>17</td>
<td>county</td>
<td>blood</td>
<td>action</td>
</tr>
<tr>
<td>18</td>
<td>committee</td>
<td>layer</td>
<td>drug</td>
</tr>
<tr>
<td>19</td>
<td>bethlem</td>
<td>eye</td>
<td>hypnotic</td>
</tr>
<tr>
<td>20</td>
<td>great</td>
<td>corpus</td>
<td>administer</td>
</tr>
</tbody>
</table>

Fig. 3: Window topic word assignment. This figure displays three topics that are present in the *British Journal of Psychiatry*, 1876. Via the words associated with each topic, a tag could easily be assigned for each subject: the first topic is about asylum management, the second is broadly related to neurology, and the third is about drugs. 2019. © Eva Andersen.

38 These dynamic topics are no longer based on the original page content but on the window topics that were created earlier. Derek Greene, “Derekgreene/Dynamic-Nmf,” September 16, 2020, accessed June 23, 2021, https://github.com/derekgreene/dynamic-nmf.
When reviewing the keywords in the dynamic topic, a couple of interesting points can be observed. First of all, the British corpus I used begins in the 1850s, while the dynamic topic indicates that the word “syphilis” in connection with GP only appears from the 1880s onwards. This tells us at a glance when the connection between GP and syphilis became more central. Secondly, the vocabulary used became more technical around the turn of the century. This is for example noticeable in the use of the words “spinal” and “wassermann.” Both these terms refer to August Paul von Wassermann and his Wassermann test, which was developed to discover the presence of syphilis via the extraction of blood and/or spinal fluid.

### 5.2.2 Multiple ways of searching and exploring via an interface

While I found this raw topic modeling output understandable and usable, it also required my repeatedly switching between the given output and the digitized sources in order to read the content of the psychiatric journals. As with AntConc, this slowed the exploration process down considerably. This was improved by importing the topic modeling pipeline into histograph, thus creating a direct link between the topic modeling output, the digitized sources, and the textual transcription of the sources.

Making use of histograph enabled me to integrate multiple ways of searching – a necessary search strategy to explore the corpus to its fullest. As Coles et al. have said, “[...] distant reading visualizations cannot replace close reading, but they can direct the reader to sections that may deserve further investigation.”³⁹ One of the many advantages that these search tactics brought to the fore was that this kind of tool can be a valuable addition to the use of more conventional methods (such as finding information only via tables of contents or indexes). Below I highlight some of the different search mechanisms that I used.

A first way to explore the corpus was via the visualization of the generated topics (Fig. 4). Based on the tone and size of the dots shown in the visualization I could observe how often a certain topic appeared over time. This was especially useful in searching for and selecting subjects (such as general paralysis) that could function as case studies for my PhD dissertation. In addition, there is the possibility to zoom in and out of this dot-visualization in order to view the pages related to a specifically selected year.

---

**Fig. 4:** Dot-visualization of the different topics and their occurrence over the *British Journal of Psychiatry* corpus via color-coding and circle size in histogram. Topic labels were assigned based on my knowledge of the history of psychiatry. 2019. © University of Luxembourg, http://histograph.eu/, developed by Lars Wieneke, Marten Düring, and Daniele Guido.
Using these search approaches within histograph helped me extract relevant content, and especially to discover otherwise hidden content and “weak signals” around very distinct topics, which would not have been made visible by a manual search approach. Sometimes there were considerable discrepancies between what could be found within a table of contents and the relevant locations suggested by the system. Furthermore, histograph allowed me to find the proverbial “needle in a haystack” while I investigated a particular theme.

To come back to the example of GP: within historical research there has been quite some emphasis on physicians, alienists, and syphilographers developing a cure for this disease. However, their interest in GP entailed more than a race for a cure. Their research would take many directions and become quite diverse. Medical practitioners, for example, conducted research on the sense of smell in GP patients, the presence of peptone in urine and changes in body temperature – all of which were examined as possible indicators of the disease. Although these instances are rather infrequent, they do help paint a broader picture of physicians’ and alienists’ interest in GP. Without the use of digital search, and more specifically the use of the above-described search tactics in histograph, these instances would have been almost impossible to trace.

A second way of searching within histograph was configured by clicking on a specific topic, for which all related pages were then displayed. Via additional keyword search these results could then be fine-tuned to my specific interests. Aside from a connection between GP and syphilis, many other potential causal links to GP were generated. To get a better grasp on these I was able to, for example, study GP and its relation to alcoholism. In this case, histograph displayed only those pages on which the words “general paralysis” and “alcoholism” appear together within the selected GP topic (Fig. 5). This facilitated the search for only relevant content within the psychiatric journals, reducing the number of pages that I needed to close read.

A third method of searching involved using “keyword mentions.” One or multiple keywords were specified by me and were displayed via a bar chart at the same time, also making it possible to directly access the pages with these particular words. This proved to be a useful search tactic to study for example the presence of specific psychiatrists. One of them was the internationally renowned Belgian alienist Jules Morel. By implementing keywords with different variations of the spelling of his name (jules morel, jul. morel and j. morel) a

Fig. 5: A histograph results page showing, within the topic “General paralysis,” the additional keywords “general paralysis” and “alcoholism” (top left), and their corresponding pages (center). 2019. © University of Luxembourg, http://histograph.eu/, developed by Lars Wienke, Marten Düring, and Daniele Guido.
straightforward overview of his name’s occurrence within the British psychiatric journal was generated, which created a basis for more in-depth exploration. Via the use of these different layers and its accompanying search tactics, I had the opportunity to be more precise, as well as flexible about what I wanted to investigate. During my research process I have used these layers for different purposes, ranging from the discovery of relevant study subjects via topic modeling, to the discovery and fine-tuning of my already-selected case studies via “keyword mentions.” These examples are of course also traceable in a similar fashion through tools such as AntConc or Voyant Tools which, to a certain extent, use similar mechanisms. However, with histograph, due to the incorporation of different search strategies as well as its more “natural” visualization of the sources, the research process was improved and accelerated.

6 Conclusion: Digital search as an extension of the historian’s workflow

One of the first tasks in the historian’s research workflow is “search.” This is where all historical research begins and it is one of many skills that historians are proficient in. However, within the scope of my PhD project researching the dissemination of psychiatric knowledge across the nineteenth and early twentieth centuries, it became apparent that using only an analog search approach, either for finding my sources (psychiatric journals) or gathering all relevant text fragments within my sources, would not be sufficient as a tactic. This is where digital search became a central aspect of my research workflow.

An implied question that runs throughout this chapter is in how far the historian’s skill in analog or traditional search is (dis)similar to that of digital search, and whether the latter undermines the former. This cannot be answered with a simple yes or no answer. Firstly, in essence, analog search and digital search are not that different: their common factor being keywords. Furthermore, analog search often remains present – whether consciously or unconsciously – within the boundaries of digital search. With digital search we are directed to potentially relevant sections within a source. But as historians we will always investigate these specific pages in more detail. It is within this process of close reading specific sections that we (un)consciously apply traditional search. If we, for example, were being directed to a section about non-restraint via a digital keywords search, our eye might be caught (just as in analog search) by certain other words or phrases that may be relevant and which could help to fine-tune our search tactics.
However, this does not mean that we do not need to be aware of some aspects that make digital search disparate from analog search. This awareness starts with the realization that the ways in which one can digitally search (including keyword search, fuzzy search, topic modeling, time range selection, etc.) are far more extensive than when we talk about analog search. In addition, not every project or research question benefits from the same search tactics—and refining our search approach is a process that often involves trial and error.

Secondly, digitization, including the multiple functionalities of the search box, “[. . .] opens shortcuts that enable ignorance as well as knowledge.”\(^4\) We need to be aware of the pitfalls that can await the historian when applying digital search. The impact on search possibilities and strategies can be quite tremendous and starts with the digitization of sources. From the scanning machines and scan settings used, through the choice of OCR software and its accuracy, to the source’s document format—all have an impact on which search tools the historian can ultimately use.

Thirdly, after locating our digital sources and deciding on our search tools, there comes the problem of developing one or multiple search tactics. Because the information that the historian is looking for is often complex it is better to make use of multiple search tactics. Diverse search functionalities can help us reassess our existing knowledge of particular topoi within history more easily, and can also lead us to discover new or forgotten subjects of interest. Of course, each search function comes with its own opportunities and drawbacks. However, I think that as long as we try to fully understand these functions and be transparent about how we use them—not forgetting to combine distant and close reading—using multiple approaches to digital search can contribute to many realms of historical research, including my own fields of transnational and psychiatric history.

The question now is whether the many technologies available to assist us in searching and gathering information can also enable us to absorb information faster (e.g. speeding up information processing).\(^4\) The effort needed to interpret and close read texts on past events takes time. While the action of interpreting has not sped up as rapidly as technological innovations—we are just human after all—I do think that using digital search tools can speed up certain parts of the search process, as well as the further exploration and analysis of

\(^{41}\) Putnam, “The Transnational and the Text-Searchable,” 379.

sources. This is especially true when either looking for specific information in a large volume of data (that elusive needle in a haystack) or wanting to investigate large amounts of data over multiple years and corpora.

With this in mind, I also want to briefly highlight a couple of interesting avenues that still need to be explored in relation to overcoming some of the limitations that digital search currently experiences. Further research efforts need to be invested in topic modeling across languages, making dynamic and cross-lingual explorations possible. When it comes to transnational knowledge exchange, cross-lingual exploration might be one of the most significant approaches that could help researchers discover patterns of exchange over a wide geographical region. Another area that merits further exploration relates to aspects such as the expansion of keyword lists, the use of word embeddings for historical corpora and the use of word co-occurrences – since a researcher never can be aware of all historical variations of certain terms or all their misspellings.

One of the key reasons that we need to continue developing digital search techniques and interfaces is precisely because search is such an essential element within a historian’s research practices. Digital search stands closer to, and is more of a continuity of, analog historical scholarship than many often think. Historians do not need to give up on their ways of practicing history via close reading – rather, the option of digital search can function as an extension of already-existing practices.

References


