Routledge
Taylor & Francis Group

Historical Methods: A Journal of Quantitative and Interdisciplinary History

ISSN: 0161-5440 (Print) 1940-1906 (Online) Journal homepage: http://www.tandfonline.com/loi/vhim20

Phrasing history: Selecting sources in digital repositories

Hieke Huistra & Bram Mellink

To cite this article: Hieke Huistra & Bram Mellink (2016) Phrasing history: Selecting sources in digital repositories, Historical Methods: A Journal of Quantitative and Interdisciplinary History, 49:4, 220-229, DOI: 10.1080/01615440.2016.1205964

To link to this article: http://dx.doi.org/10.1080/01615440.2016.1205964

HISTORICAL METHODS
2016, VOL. 49, NO. 4, 220–229
http://dx.doi.org/10.1080/01615440.2016.1205964

OPEN ACCESS

Phrasing history: Selecting sources in digital repositories

Hieke Huistra[a] and Bram Mellink[b]

[a]Descartes Centre for the History and Philosophy of the Sciences and the Humanities, Freudenthal Institute, Utrecht University; [b]Department of History, University of Amsterdam

ABSTRACT
In recent years, mass digitization has opened up voluminous text corpora to human interpretation. Full-text search lets historians now find new sources that can change their understanding of thoroughly studied historical episodes. At the same time, it forces scholars to access historical sources in a new way: through specific words. This article analyses the consequences of this new way of accessing sources and investigates which search technologies are best suited for historical source selection in digital repositories. It argues that to seize the opportunities that digitization offers, historians must refine their search technologies so that they are based on words but are less dependent on exact phraseology.

KEYWORDS
Digital humanities; digitization; full-text search; keyword search; source selection

We have all been there: sitting behind your computer screen, browsing the national newspaper database, searching for sources on the public reception of, say, the assassination of John F. Kennedy. First, you enter the president's name and limit the time period to 1963, but it feels feeble to look only for sources containing "John F. Kennedy" in full, and it returns many articles on irrelevant aspects of Kennedy's life. You then extend your search with synonyms ("president," "Kennedy") and words more specifically related to the subject ("assassination," "shooting," "Dallas"), but still the results are either too many (when you search for any of these words) or too few (when you search for all of them). At the end, you randomly select a few newspaper articles, some of which prove useful. You add them to your footnotes as if you consulted them in the archive, perhaps happy with their added value, perhaps with a twinge of guilt about your haphazard selection. You then close your browser and return to the paper sources (business as usual).

But for how long? The digital world is closing in. Libraries around the world are digitizing newspaper and journal collections, archives such as the Newton Papers and London's Old Bailey Court Proceedings have become available online, books are being scanned and made full-text searchable.[1] Historians use these digitized sources a lot (Chassanoff 2013; Milligan 2013), sometimes because access to the paper versions is restricted to protect them, but more often because accessing sources digitally is both easier and cheaper than travelling to an archive and ploughing through boxes manually. Researchers working on contemporary history often do not even have a choice: An increasing share of their material has become digital-only (Hampshire and Johnson 2009; Patel 2011).

The new digital sources can be explored in many ways. Scholars can, for example, count words to see which ones are used most or which ones often appear close together (e.g., Guldi 2012; Nicholson 2012). They can visualize the results of their counting with word clouds or Ngram graphs (e.g., Huijnen et al. 2014; Michel et al. 2011). They can visualize networks of peoples and places that emerge from the sources (e.g., Burrows 2012; Shell forthcoming). These are all examples of the advanced techniques we identify with the "digital turn." Far more important for most historians is the "digitized turn" (Putnam 2016, 378–79): the use of full-text search to retrieve manageable sets of sources from large repositories to study them manually.

Yet, although full-text search is the most used digital technology, it is also the one least talked about, and it remains remarkably undertheorized (Putnam 2016, 379–80). Most historians seem to assume that digitization has not really changed anything, so much so that most of us do not even feel the need to mention the use of digital sources in our footnotes (Hitchcock 2013, 12). This is problematic because technologies, no matter how mundane, are never neutral and co-shape the outcome of the research they are used in (Drucker 2011; Leca-Tsiomis

CONTACT Hieke Huistra h.m.huistra@uu.nl Descartes Centre for the History and Philosophy of the Sciences and the Humanities, Freudenthal Institute, Utrecht University, P.O. Box 85.170, 3508 AD Utrecht, The Netherlands.

2013, 468–69; Liu 2013, 416–17; Solberg 2012, 55; Turkel 2011). This implies that some search technologies for accessing digitized sources are better suited for doing historical research than others. We must figure out which technologies these are because only when we know which technologies we prefer, can we ensure that these are indeed the ones available to us.

Insofar as full-text searching is being debated, it is usually contrasted with browsing (Brake 2012, 222–224; Hitchcock 2008, 2013, 13–16; Nicholson 2013; Putnam 2016). This comparison helps to show how digitization fundamentally changes our research practices, but it also risks suggesting that "searching" is a uniform practice. This is not true: There is more than one way to use a search box. Hence, historians must choose not only between browsing and searching, but also between different search technologies.

This article aims to assist with the latter choice. To do so, it offers a refinement of the emerging debate on full-text searching in digitized sources: Instead of contrasting searching with browsing, it compares different search technologies with each other. Through this comparison, we show that to seize the opportunities digitized sources offer, we must refine our search technologies so that they are based on words but are less dependent on exact phraseology. Our argument is based on an analysis of recent digital history projects and on our personal experiences as historians in a digital history project.

In the first section, we investigate how digital technologies influence historical research, focusing specifically on search technologies and how they shape source selection. A more detailed analysis of keyword search follows in section two, which discusses problems in choosing the right search words. Informed by examples from the field of digital humanities, we then (in the third section) discuss alternative forms of full-text search, based on words but less dependent on exact phraseology. To ensure that digital repositories enable such alternative search technologies, we must get involved in digitization projects early on, when we can still influence how digitization is carried out. Hence, to conclude the article, we suggest a digital agenda for scholars working with historical sources, aimed at improved accessibility of digitized sources and increased transparency in digital research.

Assessing digital techniques

In the last decade, the number of digitally available sources has exploded. Archival material such as the Newton papers, Darwin's correspondence and manuscripts, and letters written by and to Alfred Wallace have been published online; Google has digitized millions of books, while non-commercial parties have created much smaller but more structured and transparent online book repositories; and national libraries rapidly digitize vast amounts of newspapers. As of early 2016, for example, the Dutch National Library had digitized approximately 8 million newspaper pages, while the British Newspaper Archive (set up in partnership with the British Library) then contained over 13 million pages.[2]

The new digital sources differ from the paper originals. For one, they smell different. Most of us might shrug this off, but not the medical historian who wants to sniff letters looking for traces of vinegar (once used to prevent the spread of cholera, now an invaluable source for charting past epidemics) (Brown and Duguid 2000, 173–74). The digitized correspondences we have available are useless for this kind of research. This might seem unavoidable, but digital historian William Turkel (2011, 288–90) has shown we can preserve smell in digitized sources, if we want to. The loss of smell is not an inevitable consequence of digitization, but the result of human choice. Digitization transforms our sources, but we can influence which source characteristics are preserved in the transformation because we decide which technologies we want to use specifically.

Virtually no scholars lament the usual loss of smell in digitization, but many are concerned about a loss of context (Bingham 2010, 230; Broersma 2012, 40–42). Context is essential in historical research. Take newspapers: When researchers browse them manually, they study more than relevant article text. The tedious process of searching for relevant articles inevitably draws their attention to competing topics and thus enables the researcher to intuitively assess the importance of the given topic at that time. Historians have long (and often unconsciously) relied on this "*sitfleisch*-based test of statistical significance" (Putnam 2016, 392), but digitization has made this test optional. Similarly, source characteristics changing over time are less visible in digitized sources. Newspapers in the 1950s are a different type of medium than they were in the 1900s, or the 1850s, something that is immediately apparent when leafing through them by hand, but easily missed when searching them electronically.

Digitization does not mean we must give up this contextual study of text, though: Like smell, context can be saved. Unlike smell, in current repositories it often is, although some parts of it more than others. The context of the page is usually well-preserved. The Dutch National Library's newspaper database, for example, provides thumbnails of the pages in the result list produced by a search.[3] The thumbnails occur next to the results, and the article containing the search words is highlighted. If you select a specific result, you get an image of the page, where you can zoom in and out to read the article and study its surroundings. While you do this, you get visual feedback of your position on the page as a whole on a thumbnail adjacent to this image. Reading the full article often requires some mouse clicks and scrolls to zoom in and out, but this might be a good thing: Literary scholar Laurel Brake (2012, 223) has argued that the required effort actually makes scholars *more* aware of the context of the page. Brake is less optimistic about the context of the issue and the newspaper as a whole; in the repositories she studied, figuring out how to "leaf through" successive pages and issues is discouragingly difficult.

Indeed, as she pointed out, virtually all digital periodical repositories are focused on searching, not browsing (see also Towheed 2010). The search field dominates the access to newspapers. Again, this is not an inevitable consequence of digitization, but the result of human choice: the decision to design the interface in such a way that the search field is on top (clearly in view), while browsing opportunities (if present) are initially hidden. Scholars may very well challenge this decision.

This does not mean we should reject searching altogether, for the search field has its merits as well. It allows scholars to extract articles on specific topics from voluminous text corpora in a fraction of the time it would take to do so manually. This offers major opportunities: It lets us find new sources that can change our understanding of thoroughly studied historical episodes (Furnée 2015; Goodrich 2013; Kaalund 2014; Lee 2014). Our aim here is to investigate how to seize these opportunities because searching fruitfully is not as self-evident as it may seem. Searching can be done in many different ways. Currently, available search technologies vary between digital repositories: Many offer mainly variations of basic free-text searching; some allow researchers to carry out more advanced complex queries. Again, as with dominance of searching over browsing and the loss of smell, which search technologies are on offer co-shapes the outcome of our research and is the result of a human decision.[4] Therefore, we must think about which search technologies suit our needs best, and then we must get involved in digitization projects to make sure these technologies are indeed available in the digital repositories we use.

To think about which search technologies would be most helpful in selecting sources, we first must gain an understanding of full-text search and its consequences. Full-text search fundamentally alters our access to sources. Digital historian Bob Nicholson (2013, 66–67) has conceptualized this as a shift from top-down to bottom-up access. Traditionally, Nicholson explains, researchers access their sources top-down. When studying a certain event (Nicholson uses the Titanic disaster as an example), they start by deciding on a specific newspaper title in which they want to look up the event. They then select a specific issue at a specific date (in the Titanic case, shortly after April 15, 1912), looking for specific headers which indicate the topic, thus arriving at the article in question. In a full-text searchable newspaper database, the process is turned around. Researchers start with entering keywords ("Titanic"), which provide direct access to the text of articles. Information such as headers, issues, and newspaper titles are still available, but do not play a crucial role in the selection of articles.

With the Titanic, the top-down approach works reasonably well because the topic is strongly linked to a specific date. For many topics, it is unclear which specific newspaper issues are relevant until we have found them; think about the long-term reception of the assassination of Kennedy, or the changing public perception on the safety of travel by boat. For such "non-dated" topics, bottom-up access to newspaper articles is a lifesaver because it circumvents the need to decide which newspaper issues are specifically relevant before you have found them.

Accessing newspapers bottom-up, through the search field, means we must type in search words. Our source selection now depends on the specific search words we choose. What does this mean in practice? The ability to search a voluminous corpus with keywords enables a scholar to obtain instant information about people, organizations, and topics, even if they have not yet been covered in existing literature (Solberg 2012, 58). When a scholar attempts to use keywords to delimit a certain topic, say homosexuality in the nineteenth century, issues quickly arise. How does one account for the various ways in which this topic can be described, without generating far too many results? How does one account for possible blind spots in the search query, and how does one judge the representativeness of the generated results? In other words, which search technologies should we use, and hence, which search technologies do we want to be incorporated in our digital repositories?

The challenges of digital source selection

Medical historian Robert Proctor (2011, 9), confronted with 14 million digitized tobacco industry documents when writing his history of the cigarette, described full-text searching as "a powerful magnet": It allows scholars "to pull out rhetorical needles from large and formidable document haystacks." This way of working with digital sources has also been discussed by the American historian Charles Upchurch (2012), who carried out a 10-year digital research project on the representation of sex between men in British newspapers between 1820 and 1870. When the project started, in 1999, he had one digital resource: a CD-ROM with the nineteenth-century *Palmer's Index to the Times*. At the end of the project, in 2009, he had 45 full-text searchable British newspapers at his disposal from the online repository "British Library newspapers 1800–1900."

Upchurch sought to take advantage of the newly available full-text digitized newspapers to extend his earlier research based on *Palmer's Index*, but he concluded that *Palmer's Index* had proven itself a more reliable guide to his research topic than keyword search (2012, 91). The main reason was that *Palmer's Index* was an index of subjects composed by humans and therefore delivered direct access to the topic Upchurch was interested in. This circumvented the problem that topics can be described in many different words and that words can have multiple meanings. An example illustrates this. For his research with *Palmer's Index*, Upchurch compiled a list of the labels that the nineteenth-century indexers used for articles discussing sex between men. To do this, he worked backward: He looked up articles on cases he had extracted from other primary sources to see under which headings they had been indexed. Examples of the headings he found were "indecent assault," "indecent conduct," "gross indecency," "unnatural offence," and "abominable crime." Each new heading led him to additional articles, many of which proved relevant for his research. When the newspapers became available digitally, he performed full-text searches using the headings as keywords, hoping to find even more relevant articles.

Although he did indeed find more articles, many of them turned out to be not relevant at all. "Abominable crime," for example, was used more often in the context of murders, non-sexual attacks, and cruelties overseas than in relation to sex between men; "gross indecency" was mainly used for heterosexual sex or to judge events varying from suicides to art exhibitions. In *Palmer's Index*, these headings had indicated articles on sex between men in the vast majority of the cases (Upchurch 2012, 96). What is at play here is the difference between topics and words. The headings in *Palmer's Index* are *topics*, but when they are used as search terms, they return to being simply *words*. In *Palmer's Index*, the indexers have interpreted the words in the context of the article, which removes many false positives, as Upchurch's results show. It also solves another problem: false negatives, articles that are *not* found, because they discuss the topic in other words than we would expect.

Neither of these two problems are unique to full-text search. They occur in all digital techniques that access digitized sources through words. Thus, virtually every digital method for working with textual sources is confronted with these problems, not just retrieval methods such as source selection through keyword search, but also more quantitative analytical methods (e.g., King, Kübler, and Hooper 2015; Lijffijt et al. 2016; Verhoef 2015, 56–57). As soon as one uses words to access sources, one is confronted with the equivocality of natural language: The meaning of words varies between contexts. However, the consequences of the resulting problems differ significantly between digital methods. For approaches in which computers "read" the texts, such as quantitative analysis of large data sets, false positives are a major concern. Too many false positives skew the results of the analysis, and since the individual texts are not read manually, false positives easily go unnoticed. For approaches in which humans read the texts (after they have been selected with help of a computer, for example through full-text search), false negatives are more of a problem than false positives. The latter are always noticed and can then be filtered out either manually or by adapting the search strategy.

Upchurch only ran into the problems caused by equivocality once he started doing full-text searches, because he then lacked the interpretational work of the indexers who had filtered out false positives and false negatives. Unfortunately, for most newspapers, full-text search is the only choice: *Palmer's Index to the Times* is a unique resource that covers only one newspaper. This means we must find another solution to the problem. Upchurch examined the results of his keyword searches and discarded the false positives manually. This does not solve the problem of false negatives, since that would require doing exactly what we are trying to avoid: going through the full corpus by hand. Moreover, accounting for false positives manually is no longer feasible if the number of results increases, which easily happens when working with large newspaper corpora, as we experienced in our own research project.

We work as historians in the digital history project "Translantis" at Utrecht University and the University of Amsterdam.[5] This project, which started in 2013, aims to map the rise of the United States as a "reference culture" in Dutch public discourse during the twentieth century using the voluminous newspaper database of the Dutch National Library, which contains roughly 94 million newspaper articles (Van Eijnatten, Pieters, and Verheul 2013). With such amounts of data, delimiting a topic is difficult. Take debates on the welfare state, one of the project's subthemes. The search query "social security" returns about 21,000 articles.[6] On the one hand, the staggering number of articles prohibits filtering them manually, while we know many of the 21,000 might be irrelevant. On the other hand, the 21,000 returned articles are just the tip of the iceberg, since welfare politics are being discussed in many different words, including "planning," "welfare state," "Keynesian economics," "national health service," and "New Deal." Searching for one of these words seems painstakingly arbitrary. How does one delimit a topic on justifiable grounds?

These source selection issues are already difficult of their own accord. In the 1940s, Roberto Busa (the Jesuit priest often seen as the first "digital humanist") studied the concept of "presence" in Thomas Aquinas's oeuvre. He quickly realized that the direct Latin translations (*praesens* and *praesentia*) were much less relevant as a marker for his topic than the preposition *in* (Busa 1980, 83). Busa concluded that the conceptual system that an author has in mind cannot be grasped by focusing on isolated words: The hidden structures of texts are of key value for a proper understanding, and Thomas Aquinas's conceptual understanding of "presence" could not be singled out with keywords alone.

The limits of keyword-based delimitation become even more apparent when research questions exceed the boundaries of a single text. If a scholar would use the works of Thomas Aquinas to gain insights into the historical development of scholasticism in the High Middle Ages, Thomas Aquinas's *Summa theologica* would be only one of the many sources that shed light on the overarching topic, covered by multiple authors in places during two or three centuries. Consequently, the scholar must account for differences in word usage among persons, between places, and across time. Thus, the variation in word usage and intended meanings inevitably multiplies, making a delimitation through keywords even more problematic than when one text or oeuvre is being studied in (relative) isolation.

The resulting gap between phraseology and meaning largely explains why it is so difficult for the majority of historical researchers to choose proper keywords while accessing digitized sources. Their topics of interest are described through words but cannot be pinpointed through simple keywords alone. Yet, once sources have been digitized, entering isolated keywords often becomes a prerequisite for access. When a scholar aims to delimit a certain topic (e.g., homosexuality or welfare politics) by means of keywords, he or she cannot ignore that fact that the topic in question can be described by many different (key)words, while each of these words may carry multiple meanings.

To overcome these challenges, we must think about other search technologies than the standard keyword search currently offered by most digital repositories. To do so, we must figure out how to use keywords to our advantage so that we can use multiple words to describe a topic, while being able to filter out irrelevant results. Practical examples are more useful here than abstract reflections. Therefore, we will now discuss two experiments with digital source selection.

Sophisticated selection of digital sources

The examples in the previous section seem to provide a double bind. When throwing out voluminous "word nets," full-text search generates far too many results. Scholars then become overwhelmed with data: Reading all returned sources is impossible, but they do not know which of their results are false positives and should be discarded. On the other hand, if they would limit the number of search words, their source selection would be arbitrary and would return far too few results to be convincing (the problem of false negatives).

Scholars could avoid this situation if they could use a maximum number of synonyms to describe a topic, while keeping the number of returned search results limited and relevant. This can be done through developing so-called combined search queries, in which two or more topics are described by a wide variety of synonyms and searched for in relation to each other. The research project "War in Parliament" provides an example. The project investigated political accusations of national socialist sympathies (topic A) addressed at members of a controversial right-wing political party in the Netherlands, the Boerenpartij (actor B) (Piersma et al. 2014). Dutch historians have always claimed that the Boerenpartij (Farmer's Party) was regularly accused of Nazi sympathies during the 1960s, but they have never examined this claim methodically (Donselaar 1991, 125–33; Nooij 1969, 215–17). "War in Parliament" aimed to do this digitally, by connecting general references to the Nazis during the period 1963–81 with general references to the Boerenpartij in the digitized parliamentary proceedings of the Netherlands. They confirmed their findings manually.

This kind of source selection can be understood as building digital sets. The first set contains all documents in the parliamentary proceedings with references to national socialist sympathies in the broadest sense of the word, such as "fascism," "political delinquent," "anti-Semitism," "war," and "collaboration." This set is created with the search query below. The query contains wildcards (*) to account for inflections and spelling variations. For example, "fascis*" can refer to both "fascism" and "fascist." The OR operator indicates that the computer will return all texts in which either of the listed words on either side of it appears. The date range is set to include all parliamentary proceedings from 1963 to 1981:

> fascis* OR NSB OR "politiek delinquent" OR "politieke delinquent" OR "politieke delinquenten" OR collaborat* OR "nationaal socialisme" OR "nationaal socialistisch" OR "nationaal socialistische" OR antisemitis* OR oorlo* OR Hitler OR Mussert OR Roskam OR Boerenleider OR Jeugdstor* OR NSK* OR Waffen-S* OR Landstand (Piersma et al. 2014, block 15)[7]

This query generated 8,000 texts containing (possible) references to national socialist sympathies. A second query contained the name Boerenpartij and the names of all its members in Parliament and Senate between 1963 and 1981:

> Boerenpartij OR Adams OR Koekoek OR Voogd OR Brake OR Harmsen OR Harselaar OR Bossche OR Koning OR Kronenburg OR Leffertstra OR Nuijens OR Verlaan (Piersma et al. 2014, block 16)

This generated about 12,000 hits. After both sets were created, the scholars combined the two queries, searching for texts with references to both national socialist sympathies (topic A) *and* (members of) the Boerenpartij (actor B). This resulted in 179 hits, which they studied manually.

This method, henceforth referred to as "related searching," makes the scholar less dependent on exact word usage than simple keyword search. Searching only for the 19 general references to national socialism generates too many hits for human interpretation. The same goes for all references to the Boerenpartij and its members. Relating both categories refines the source selection and results in a few highly relevant articles. Using a wide variety of synonyms increases the representativeness of the results. A search query limited to "Boerenpartij" would be arbitrary. Once a scholar uses synonyms or includes all parliamentary members of this political party, the findings gain credibility. Naturally, the quality of the keyword choice remains crucial, and no selection method alone can uncover all relevant sources, but this applies to traditional source selection as well. Related searching involves conscious choices, which can be debated, but which make it more transparent and credible than a search with one or two arbitrary keywords.

The strengths of related searching are closely related to its central weakness: It hinges on the relation between two word sets. It works well for the Boerenpartij case study because the research question investigates an assumed relation between A (ascribed national socialist sympathies) and B (the Boerenpartij). Historians studying, for instance, the changing political discourse of national socialists, however, cannot acquire their sources with related searching. Because many historical topics cannot be delimited by relating them to other topics, related searching does not provide solutions for all historical questions. Therefore, we now turn to a method to retrieve sources without confining the scholar to fixed relations between sets of keywords.

"Weighted searching" weights keywords to search for relevant sources: The more specific and relevant a keyword, the higher its value. This method has been experimented with in the research project "ePol. Verbundprojekt Postdemokratie und Neoliberalismus," set up by a group of German political scientists. This project investigates post-war neoliberal political discourse in four (West-)German newspapers (Lemke 2014; Niekler, Wiedemann, and Heyer 2014, 9). As neoliberal discourse can be signaled through many different keywords, a source selection method that only allows for synonyms ("free market" OR liberal OR right) would provide too many results, while a source selection method aimed at word combinations ("free market" AND liberal) would provide too few. The ePol project aimed at a broad discourse, so the methods described above do not suffice.[8]

The ePol group went about searching their newspapers in a rather different way. They attributed relative numeric values to all words deemed relevant for the neoliberal discourse that they were studying. They based this weighted word list on a starting corpus of 36 texts generally believed to have shaped the neoliberal discourse, such as Friedrich Hayek's classical study, *The Road to Serfdom* (1944), and Milton Friedman's famous 1951 lecture, "Neoliberalism and its Prospects." They then analysed this corpus using computer-based linguistic methods and manual research to draft a list of 500 words and phrases indicating neoliberalism, like "free market", "privatization," or "personal responsibility", and to score these words according to how likely they were to indicate neoliberalism. The word "privatization," for instance, got a higher value than "freedom." To identify arguments, the scholars created a separate list of 127 words or word combinations regularly used in argumentative texts, such as "secondly," "contrary to," and "because of." By combining the neoliberal dictionary with relative word values and the argumentative dictionary (an application of related searching), they ranked newspaper articles according to the probability they contained neoliberal arguments (Dumm and Lemke 2013).

Including relative word values is a simple addition to existing forms of keyword search, but it has considerable advantages. Most importantly, weighted searching frees scholars from the obligation to provide words that should either appear or not appear in the corpus. Attaching relative values to words results in a search method that can easily be customized and refined. When a certain combination of words produces results that are inconsistent with other sources or that have clear gaps, and one suspects that selection process must be refined, it is possible to do so both by adjusting the word list and through adjusting the word values.

Of course, such refining and customizing should be disclosed and explained, in order to avoid a historiographical equivalent to data massaging: altering word lists and word values randomly, without justification, until they yield a publishable outcome. In the ePol project, random empirical manipulation was avoided by first studying the smaller starting corpus to establish the word lists and relative word values, and then using the results of this preliminary study of canonical neoliberal texts to approach the much larger newspaper corpus (which, unlike the 36 texts, could not be studied by hand). Like all source selection strategies, this approach involves choices that can be disputed. For instance, how do we know the 36 selected texts are indeed representative of the neoliberal discourse?

These are legitimate questions, but they are not unique to digital research. Sources, whether analogue or digital, are easy to manipulate, and scholars should always explain to their audiences why they believe their source selection to be coherent and representative. Instead of expecting the computer to solve problems with representativeness for us, or assuming that a quantitative approach results in higher degrees of objectivity, we should continue to take matters into our own hands and account for our selection criteria. Simultaneously, we should urge digital repositories to disclose the algorithms used by their search technologies.

The ePol project does not provide a one-size-fits-all solution for selecting sources. Its reliance on a specific set of theoretical and highly political texts as a starting set fits in relatively well with certain historical approaches such as intellectual history and the history of ideas, but seems less suitable for writing a history from below, for example. Even within the history of ideas, one might object that taking a single text corpus as a point of departure for historical research is ahistoric: It assumes continuity in terminology and might therefore miss shifts in language.

Such objections result from this specific application of weighted searching; they are not inherent to the approach. Not only should it be possible to use different text corpora for different decades, but it is also conceivable that weighted searching will be used for different purposes, for instance, to trace networks of historical actors by using dictionaries consisting of names instead of phrases and concepts. Furthermore, we can refine weighted searching by designating special word clusters. When you are, for instance, investigating the introduction of scientific management (a late-nineteenth-century method for increasing labour efficiency) in Europe, words like "labour," "time," and "process" may be too general as indicators when used separately, but can be specific signifiers of a discourse about scientific management when appearing close to each other.

"War in Parliament" and ePol show how simple measures like relating sets of words and adding word values can significantly refine source selection. Computer technology does not provide one-size-fits-all solutions, nor does it solve longstanding issues regarding representativeness and objectivity once and for all. Small methodological inventions do allow scholars to select sources through keywords without having to rely on exact phraseology. Such sophisticated approaches require time investment; they do not provide a source corpus with one mouse click. They do enable researchers to approach text corpora in new ways, making newspapers, magazines, journals, and other large text collections more manageable and accessible than ever before. Researchers cannot decide to start using these approaches unilaterally; they must work with the creators of digital repositories. They should not just think about useful search technologies, but also act upon their conclusions. This brings us to our last point: three suggestions for topics that should be added to the digital historical agenda.

Three topics for a digital historical agenda

All historical research relies on proper source selection. Mass digitization opens up new types of sources and offers promising new opportunities for source selection. To seize these opportunities, full-text search approaches based on single keywords or basic word combinations do not suffice: Single words can have multiple meanings, and a single topic can be indicated by many different words. Scholars working with historical sources must master more advanced technologies, like related and weighted searching; otherwise, they will drown in the digital world, just as they would in an archive without inventories. This also means that digital repositories must allow for the use of more advanced search techniques. To reach this goal, historians must not take all digital matters completely into their own hands, but they must cooperate more closely with archivists and librarians in charge of digitization projects.

Bigger is not automatically better in digitization: A newspaper corpus of 94 million articles sounds impressive, but no matter how big the data, scholars still must account for their selection of sources. At the moment, digitized texts are mostly accessible like catalogues: Scholars can only enter (simple combinations of) keywords to obtain material, as if searching for a specific book or article. Source selection, however, would profit from different forms of access. Luckily, relatively small measures can improve the usability of digital corpora significantly. Archives and national libraries can contribute to this by providing more advanced search techniques such as related and weighted searching, context-aware accessibility such as browsing, or improved data indexation (Walma 2015). Scholars, on the other hand, should learn to refine their process of source selection through full-text search. If scholars can use (extensive lists of) synonyms, delimit sources through relating sets of synonyms, and manually attach numeric values to words, their control over the corpus will increase significantly. Taking control over the process of digital source selection will improve not only the usability of the results of digital searches, but also transparency, as it forces scholars to formulate explicit selection criteria to underpin the representativeness of their sources.

Second, scholars working with historical sources must account for their digital source selection. In most cases, a general description of the search methods and software together with a brief description of their consequences should suffice. For annotation, requirements vary according to the methods that have been used. In most cases, however, footnotes in which a corpus is mentioned (e.g., parliamentary proceedings) should be enriched with the search tool (if more than one tool can be used to access the corpus), the search query (e.g., "fascism OR Hitler OR Waffen-SS"), the time period (e.g., 1963–1981), and if applicable other metadata filters, the number of hits and the date the search was carried out.[9] This provides orientation points similar to archive names and inventory numbers, and increases the transparency of selection criteria and search results.

Finally, scholars working with historical sources should step into the world of digital humanities and talk with digital scholars from other (humanities) disciplines more often than they currently do. Many of them still seem hesitant to enter digital humanities debates, but they have a lot to gain from participating in them (Gibbs and Cohen 2011, 69–70; Kemman, Kleppe, and Scagliola 2014; Laubichler, Maienschein, and Renn 2013, 122; Weller 2013, 1–3). For although the exact effects of digitization differ between disciplines, virtually all disciplines are affected by the replacement of paper text with digital text, whether through the digitization of their research materials, the digitization of their secondary literature, or simply because most recent sources are born-digital and filed electronically, if they are kept at all.[10] Working with scholars from other disciplines to figure out how to best use digital texts will help us to maximize their accessibility and preservation, and benefit from digitization without giving up on sound criteria for source selection. The digital world offers scholars both challenges and benefits; to overcome the first and collect the second, we need reflection and sense of direction. With this article, we have aimed to provide a bit of both.

Acknowledgements

The authors thank the following people for their helpful comments during the writing process: James Kennedy, Clé Lesger, Peter van Dam, Bert Theunissen, three anonymous reviewers, the members of the research projects *Translantis* and *Asymmetrical Encounters* at Utrecht University, and the participants in the history of science staff meetings at VU University Amsterdam. Special thanks to Hermione Giffard, who combined her constructive criticism with editorial advice.

Funding

This work was supported by the Netherlands Organization for Scientific Research (NWO) under grant 317-52-010.

Notes

[1] For the Newton Papers, see http://www.lib.cam.ac.uk/deptserv/manuscripts/newton.html; for the Old Baily Proceedings, see http://www.oldbaileyonline.org/.

[2] http://www.britishnewspaperarchive.co.uk/ (accessed February 15, 2016; http://www.delpher.nl/nl/platform/pages?title=collecties (accessed February 15, 2016).

[3] The Dutch National Library's newspaper database "Delpher" can be found at www.delpher.nl.

[4] Technology choices in digitization projects affect more aspects than the ones mentioned here. Other issues include changing serendipity, the false sense of completeness easily evoked by digital repositories, and the quantitative objectivity suggested by some visualizations of digitized material.

[5] See http://www.translantis.nl.

[6] Kranten.delpher.nl, query: [welvaartsstaat], period: [1 Jan 1890 – 31 Dec 1995], metadata filters: [none]; search date: [21 June 2016].

[7] Translated into English, the query reads: fascis* OR NSB [the Dutch national-socialist movement] OR "political delinquent" OR "political delinquent" [adjectives can have multiple suffixes in Dutch; this is something that could also have been solved with wildcards] OR "political delinquents" OR collaborat* OR "national socialism" OR "national-socialist" OR "national-socialistic" OR antisemiti* OR war* OR Hitler OR Mussert [the leader of the Dutch national-socialist movement] OR Roskam [Dutch national-socialist] OR Boerenleider [title used to indicate Roskam] OR Jeugdstor* [Dutch equivalent of the Hitlerjugend] OR NSK* [paramilitary national-socialist organization] OR Waffen-S* OR Landstand [national-socialist farmers movement].

[8] Dutch historian Pim Huijnen currently aims to apply ePol's methods to the digitized newspaper database of the Dutch National Library; a short introduction to his project can be found at http://pimhuijnen.com/2015/02/11/kb-onderzoek-de-taal-van-het-taylorisme/.

[9] For instance, "Op de koffie bij Koekoek," De telegraaf, 3 July 1968, 5. Selected by kranten.delpher.nl, query: [Boerenpartij AND Koekoek AND fascistisch], period: [1963–1981], metadata filters: [none], hits: [12], search date: [22 June 2015].

[10] Just one example of how other disciplines are affected is that recent research by three computational biologists suggests that the type of articles cited in various scientific disciplines might have changed due to the rise of electronic journals. They found that articles with long, jargon-laded abstracts are cited more often than those with short, clear

abstracts, probably because the former are easier to find through keyword searches. Thus, the search methods currently offered by journal databases stimulate scholars to write less readable abstracts, not a particularly desirable situation, it seems (Weinberger, Evans, and Allesina 2015).

References

Bingham, A. 2010. The digitization of newspaper archives: Opportunities and challenges for historians. *Twentieth Century British History* 21:225–31.

Brake, L. 2012. Half full and half empty. *Journal of Victorian Culture* 17:222–29.

Broersma, M. J. 2012. Nooit meer bladeren? Digitale krantenarchieven als bron. *Tijdschrift voor mediageschiedenis* 14:29–55.

Brown, J. S., and P. Duguid 2000. *The social life of information*. Boston: Harvard Business School Press.

Burrows, S. 2012. How Swiss was the Société Typographique de Neuchâtel? A digital case study of French book trade networks. *Journal of Digital Humanities* 1(3):55–65. http://journalofdigitalhumanities.org/1–3/how-swiss-was-the-stn-by-simon-burrows-and-mark-curran/.

Busa, R. 1980. The annals of humanities computing: The index Thomisticus. *Computers and the Humanities* 14:83–90.

Chassanoff, A. 2013. Historians and the use of primary source materials in the digital age. *The American Archivist* 76:458–80.

Donselaar van, J. 1991. *Fout na de oorlog: Fascistische en racistische organisaties in Nederland, 1950–1990*. Amsterdam: Bert Bakker.

Drucker, J. 2011. Humanities approaches to graphical display. *Digital Humanities Quarterly* 5 (1). http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html.

Dumm, S., and M. Lemke 2013. Argumentmarker: Definition, Generierung und Anwendung im Rahmen eines semiautomatischen dokument-retrieval-verfahrens. Discussion Paper Number 3, Schriftenreihe des Verbundprojekts Postdemokratie und Neoliberalismus. http://www.epol-projekt.de/wp-content/uploads/2014/10/Discussion-Paper-epol-3_dumm_lemke_CC.pdf.

Furnée, J. H. 2015. Winkelen als bevrijding? Vrouwen en stedelijke ruimte in Amsterdam, 1863–1913. *BMGN—Low Countries Historical Review* 130:92–122.

Gibbs, F. W., and D. J. Cohen 2011. A conversation with data: Prospecting Victorian words and ideas. *Victorian Studies* 54:69–77.

Goodrich, A. 2013. Understanding a language of "aristocracy," 1700–1850. *Historical Journal* 56:369–98.

Guldi, J. 2012. The history of walking and the digital turn: Stride and lounge in London, 1808–1851. *The Journal of Modern History* 84:116–44.

Hampshire, E., and V. Johnson 2009. The digital world and the future of historical research. *Twentieth Century British History* 20:396–414.

Hitchcock, T. 2008. Digital searching and the re-formulation of historical knowledge. In *The virtual representation of the past*, ed. M. Greengrass and L. Hughes, 81–90. Farnham: Ashgate.

Hitchcock, T. 2013. Confronting the digital: Or how academic history writing lost the plot. *Cultural and Social History* 10:9–23.

Huijnen, P., F. Laan, M. de Rijke, and T. Pieters. 2014. A digital humanities approach to the history of science. In *Social Informatics: SocInfo 2013 International Workshops, QMC and Histoinformatics*. Kyoto, Japan, November 25, 2013, revised selected papers, ed. A. Nadamoto, A. Jatowt, A. Weirzbicki, and J. L. Leidner, 71–85. Berlin: Springer.

Kaalund, N. K. L. 2014. Oxford serialized: Revisiting the Huxley-Wilberforce debate through the periodical press. *History of Science* 52:429–53.

Kemman, M., M. Kleppe, and S. Scagliola 2014. Just Google it: Digital research practices of humanities scholars. *Studies in the Digital Humanities*. http://www.hrionline.ac.uk/openbook/chapter/dhc2012-kemman.

King, L., S. Kübler, and W. Hooper 2015. Word-level language identification in the chymistry of Isaac Newton. *Digital Scholarship in the Humanities* 30:532–40.

Laubichler, M. D., J. Maienschein, and J. Renn 2013. Computational perspectives in the history of science. *Isis* 104:119–30.

Leca-Tsiomis, M. 2013. The use and abuse of the digital humanities in the history of ideas: How to study the *Encyclopédie*. *History of European Ideas* 39:467–76.

Lee, J. 2014. King Demos and his laureate. *Media History* 20:51–66.

Lemke, M. 2014. Postdemocracy and neoliberalism: Digital strategies for the identification of economic references in large text data collections. Paper presented at *Reference Cultures and Imagined Empires in Western History: Global Perspectives, 1815–2000*, Utrecht, June 13.

Lijffijt, J., T. Nevalainen, T. Säily, P. Papapetrou, K. Puolamäki, and H. Mannila. 2016. Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities* 31:374–97.

Liu, A. 2013. The meaning of the digital humanities. *Publications of the Modern Language Association* 128:409–23.

Michel, J.-B., Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331:176–82.

Milligan, I. 2013. Illusionary order: Online databases, optical character recognition, and Canadian history, 1997–2010. *Canadian Historical Review* 94:540–69.

Nicholson, B. 2012. Counting culture; or, how to read Victorian newspapers from a distance. *Journal of Victorian Culture* 17:238–46.

Nicholson, B. 2013. The digital turn: Exploring the methodological perspectives of digital newspaper archives. *Media History* 19:59–73.

Niekler, A., G. Wiedemann, and G. Heyer 2014. Leipzig Corpus Miner—A text mining infrastructure for qualitative data analysis. Paper presented at *Terminology and Knowledge Engineering*, Berlin, June 20. http://hal.archivesouvertes.fr/hal-01005878.

Nooij, A. 1969. *De Boerenpartij: Desoriëntatie en radikalisme onder de boeren*. Meppel: J. A. Boom.

Patel, K. K. 2011. Zeitgeschichte im digitalen Zeitalter: Neue und alte Herausforderungen. *Vierteljahrshefte für Zeitgeschichte das zentrale Forum der Zeitgeschichtsforschung* 59:331–51.

Piersma, H. I., I. Tames, L. Buitinck, J. van Doornik, and M. Marx. 2014. War in parliament: What a digital approach can add to the study of parliamentary history. *Digital*

Humanities Quarterly 8 (1). http://www.digitalhumanities.org/dhq/vol/8/1/000176/000176.html.

Proctor, R. 2011. *Golden holocaust: Origins of the cigarette catastrophe and the case for abolition*. Berkeley: University of California Press.

Putnam, L. 2016. The transnational and the text-searchable: Digitized sources and the shadows they cast. *American Historical Review* 121:377–402.

Shell, J. Forthcoming. Mapping the geography of Karl Marx's capital. *Digital Scholarship in the Humanities*. doi:10.1093/llc/fqw002 (accessed September 13, 2016).

Solberg, J. 2012. Googling the archive: Digital tools and the practice of history. *Advances in the History of Rhetoric* 15:53–76.

Towheed, S. 2010. Reading in the digital archive. *Journal of Victorian Culture* 15:139–43.

Turkel, W. J. 2011. Intervention: Hacking history, from analogue to digital and back again. *Rethinking History* 15: 287–96.

Upchurch, C. 2012. Full-text databases and historical research: Cautionary results from a ten-year study. *Journal of Social History* 46:89–105.

Van Eijnatten, J., T. Pieters, and J. Verheul 2013. Big data for global history: The transformative promise of digital humanities. *BMGN—Low Countries Historical Review* 128:55–77.

Verhoef, J. 2015. The cultural-historical value of and problems with digitized advertisements: Historical newspapers and the portable radio, 1950–1969. *Tijdschrift voor tijdschriftstudies* 19:51–60.

Walma, L. 2015. Filtering the "news": Uncovering morphine's multiple meanings on Delpher's Dutch newspapers and the need to distinguish more article types. *Tijdschrift voor tijdschriftstudies* 19:61–78.

Weinberger, C. J., J. A. Evans, and S. Allesina 2015. Ten simple (empirical) rules for writing science. *PLoS Computational Biology* 11:e1004205.

Weller, T. 2013. Introduction: History in the digital age. In *History in the digital age*, ed. T. Weller, 1–21. London: Routledge.