9.2 Toward a Humanities of the Digital?

Reading Search Engines as a Concordance

Johanna Sprondel

Introduction

In their seminal paper ‘The Verbal Concordance to the Scriptures’ from 1974, R.H. and M.A. Rouse characterize concordances to the scriptures to be ‘not only one of the earliest but probably the most important [technical aid], because its system of reference, its method of compilation and its successful application of complete alphabetization were used by generations of later tool-makers’. To what extent this holds true for more recent inventions, such search engines, and more specifically web search engines, is the question I shall address in this paper: Can we consider Google & Co. as concordance?

I will begin by examining the term ‘concordance’ (1), the use of which has increasingly broadened over the last centuries. Deriving from the Latin concordare (to agree, be united, harmonize), the implications of ‘concordance’ moved from Samuel Johnson’s definition ‘a book which shows in how many texts of Scripture any word occurs’ toward a more extensive meaning. Today, the term is commonly used in disciplines such as law, politics, cosmology, genetics, mathematics, psychology, medicine, linguistics, and grammar. I shall concentrate on the implications of the term following from Johnson’s above-quoted definition and what we would commonly refer to as a concordance in philology, theology and literary studies. What unites all these usages throughout the disciplines, however, is a sense of harmonization, reflecting the term’s Latin roots with respect to content, interests, ideas, data or methods. In a second step, I will focus on search engines (2). Although the history of computerized search engines is a recent one, I will show how the pattern of search, the harmonization of information, and the idea behind search engines is on the one hand congruent with that of concordances, while on the other hand it significantly differs from the latter (3) and will conclude with some brief remarks on how, nevertheless, comparisons like this might give us insights
on the understanding of digital developments and should encourage a critical comparison of practices (4).

Concordance and the idea of harmony

Johnson’s definition may be seen as an abbreviation of what we find in Alexander Cruden’s preface to the first complete and perhaps best-known concordance to the Holy Scriptures in English, published in 1737, where he defines the term as

an Index to the Bible, wherein all the words used through the inspired writings are ranged alphabetically, and the various places where they occur are referred to, to assist us in finding out passages, and comparing the several significations of the same word.

This definition characterizes the function as well as the capacity of a concordance, also hinting at the incentive that made many authors write concordances. While we can see the finding of a passage by using a concordance as a direct function of it, it is the indirect function – the comparison and interpretation of different applications of a word – that makes concordances an indispensable instrument for textual criticism in numerous fields of research and, as Rouse and Rouse remarked, a source for numerous systematic and methodological approaches. To understand the core meaning of the term, it seems advisable to approach it from an historical perspective. The term concordance was first introduced into the English language in 1387 when John of Trevisa translated Ranulf Higden’s Polychronicon (1327 or 1343).

These lines refer to what is assumed to be the first known concordance: most likely around 1244, a word index to the Vulgate, named Concordantiae Sacrorum Bibliorum was completed that began under the instructions of Hugh of Saint-Cher (Frere Hewe). After providing a critical edition of the Bible (Sacra Biblia recognita et emendata, around 1236) and planning to write a commentary on the same, Hugh realized that an index of the words occurring and the places of their occurrence was required to compare different passages and to see whether the
meaning of the words altered in these. He produced a sheer word index, an *index verborum*: a concordance of the simplest kind.

To understand the implications of the term *concordance* though, it is noteworthy that John of Trevisa not only translates, but *coins* the term concordance in a way that points to ‘biblical criticism’. In order to understand why, we need to look at what John of Trevisa translates as ‘concordance’: Ranulf Higden writes ‘et magnas super bibliam concordancias compilavit’, what we might translate as ‘and he has brought together many agreements all over the Bible’. A single *concordancia*, a medieval formation of the participle *concordare* (to agree), is to be seen as a bundle of two or more passages where the same word appears in Scripture, making the plural a collection of these bundles in a book. And in this broad sense the term had in fact been used before, e.g., for the lists Langton prepared in his glosses and that he named *concordantia*. But differing from what the plural ‘*magnas* [...] *concordancias*’ might suggest, John of Trevisa explicitly speaks of a single agreement, turning Hugh’s *Concordantiae Sacrorum Bibliorum* into a certain concordance.

Being merely a list of the words in the Bible with references to their location, when Hugh completed his work in 1244, John of Trevisa’s interpretation of the term would have seemed somewhat inept. But when John completed the translation of Higden’s work in 1397, the *Concordantiae Sacrorum Bibliorum* had indeed become a concordance, mainly owing to the work of three English monks who had not only revised and added quotes of the passages in question to the references of the work of Hugh but had basically produced a new concordance the *Concordantiae Anglicanae* – that was nevertheless commonly received as a ‘second edition’ of Hugh’s work. Furthermore: John of Trevisa, knowingly or not, boarded a ‘hermeneutic-semantic’ train that, for example, changed the title of Gratian’s *Decretum* from *Concordia discordantium canonum* into *Concordantia discordantium canonum*. And it is idea that Cruden identifies as the gist of a concordance that moved more and more into the focus: ‘interpreting Scripture, namely [...] comparing one Scripture with another’.

Starting with Hugh of Saint-Cher’s work, concordances to the Bible, both verbal and topical but also marginal, became increasingly common and were produced in almost every language considered ‘franca’. Along with this came a broader acknowledgement and awareness of the functions and possibilities that a concordance offered: a ‘key to the knowledge of the scripture’ and a ‘higher truth’, an idea that is reflected throughout the history of concordances and perhaps in the most allegorical way by Isaac Nathan ben Kalonymus, who in the fifteenth century made a Hebrew Concordance of the Bible that he called not ‘Concordance’ but * Меир Нетиб*, which is Hebrew for ‘Light to the Path’. In the introduction (*Petihat Меир Нетиб*) Nathan explains that a Hebrew concordance was long
due to allow the Jewish to benefit from a system that made the central text of a whole culture (the Holy Scripture) and its information easier accessible and thus easier interpretable. He also explains that he chose to do a Hebrew concordance in order to be able to argue with his Christian adversaries more easily.\textsuperscript{16}

Moreover, a reflection upon the method, its choice, and a justification of which words were chosen became part of most introductions. A critical and guiding summary is given by Kenneth McKenzie in his considerations upon concordances on the works of Petrarch and Dante:

\begin{quote}
\textit{[I]}t must be borne in mind that the editor is not presenting the results of an investigation, but a work which will be referred to, whether incidentally or systematically, for certain information; hence he should keep his own personal judgment as much as possible in the background, and should arrange the material in the most natural and practical way. In a word, the arrangement most convenient for practical use is the most scholarly. The editor should then introduce no system of classification except on a universally accepted basis, such as the order of the alphabet, or, as the case may be, the actual order of occurrence in the text.\textsuperscript{17}
\end{quote}

I shall not further depict the development of concordances in detail here, since the focus of this paper lies on the concept, method and idea behind concordances rather than on their historical development. It goes without saying that Johannes Gutenberg’s innovation around the year 1439 not only radically changed the ‘book market’ but also allowed concordances to be more widely spread, letting them be present in many churches – often \textit{ad usum communem} – most monasteries and surely fostering exegesis. And in the course of political and social development of that time and especially during the Enlightenment, when the \textit{litterae humaniores} became increasingly popular, it was suggested that not only Scripture should enjoy the privilege of being furnished with a concordance. Over two centuries the emphasis shifted more and more from clerical writings to other types of literature, and the conviction grew that not only the Bible and religious texts were books in which every word counted.

But while this widening of the ‘canon’ of works supplied with concordances is mostly of interest in terms of a change in the history of ideas, one might say that the next milestone in terms of the practices of composing a concordance is to be seen in the advent of computing and especially latent semantic indexing, developed by the French mathematician Jean-Paul Benzécri. Making a concordance became significantly easier, and works as well as authors that until then had not been recorded were provided with a concordance,\textsuperscript{18} but, of course, the ‘classics’ were among those who first were digitized and consequently furnished with con-
cordances in the late 1960s. Scholars were quick to agree that digital concordances offered a tremendous gain, and their multiple applicability surfaced with the turn in the so-called ‘Dead Sea Scrolls controversy’: Scholars were granted only limited access to the scrolls discovered in 1947. In 1991, Michael Abegg used a complete concordance of all Scrolls from the 1950s that lists every usage of every word appearing in the scrolls, including the words that flank it on either side and the name of the scroll where it appeared. By an invert use of this concordance, Abegg reconstructed the parts that, until then, had remained unpublished, programming his computer to identify overlaps of ‘word strings’, and to assemble these into larger groupings. A work that would have taken Hugh and his monks years to complete, once the passages were entered into the computer, took Abegg’s Mac SE ‘less than 15 minutes’, as he explained.

As a preliminary conclusion, we may sum up the key characteristics of a concordance in the following way: A concordance – and as the ideal type we shall define a complete concordance – is, first of all, a tool that enables us (a) to search for the occurrences of words or subjects in a text. It thereby allows us (b) easier access to works. To support this, a concordance should (c) be based upon a system of classification on a universally accepted basis. We may use the results of our search (d) to identify certain passages and even writings. Also they grant us (the user) the possibility (e) to compare, upon this basis, dispositions of words and then (f) to interpret the latter or even a whole work. Thus a concordance should (g) be objective and neutral. As for the subject matter of concordances, we may furthermore note that concordances mostly exist for (h) major works or major authors of a culture.

Searching the Web: Index, corpus and engine

The history of web search engines began in 1990, when the World Wide Web still consisted of a manageable amount of websites. A simple directory would then serve all purposes and the content of pages could be searched manually. With an ever-growing number of websites, documents, pictures and other data sets available online, search engines were needed that not only displayed all pages available, but also their specific content, making the latter a concern of the search. Furthermore, with the growth of the overall amount of data, searches that would fit the request of the user became more important, while also the time needed for an individual search became an issue (which is, for example, why a Boolean search that would list every single occurrence of a word turned out to be no longer manageable). Also sign-n-grams are no longer a basis in web-based search engines,
which they were earlier and still are, for instance, in archives where manuscripts are scanned and converted to digital documents by OCR. We therefore have to face the fact that the results of a query may be detailed, but are never a complete record of what could be found – and what we find varies widely, also due to a highly unstable corpus.

Web-based search engines such as Google, Yahoo!, Bing and others used today operate in three steps – web crawling, indexing, and searching – and only in the last of these steps is the user involved directly. These operations, to put it simply, rely on two basic concepts: an (inverted) index and an algorithm. The World Wide Web is crawled regularly, that is, it is systematically browsed, and whatever the crawler (a.k.a. the spider) finds is stored in an index. While the inverted index is a record of all words existing in all documents available with a notation of their location, the algorithm takes care of these being interpreted in a relevant and (for the user) satisfactory way once a search is placed. This user query then results in a list of search results, the so-called hits, that are sorted following numerous heuristic factors, described in the case of Google by the term PageRank; one might mention stochastic factors like the proximity of words searched and the number of their occurrence in a document or website, the number of links that refer to documents or websites available, the number of clicks a website or document has received so far (documented in a ‘query log’), but also an estimation of the relevance of the site as a whole and other criteria, to most of which we remain ignorant. The aim of a search engine is to make results available as fast as possible and, at the same time, to match the query as close as possible, thereby following the ‘principle of least astonishment’ (POLA/PLA). The POLA is based on the idea that users already have a certain knowledge of what they are searching (otherwise they would not search for it) and therefore a search result should meet their expectations, which is one of the reasons why data of the users’ queries is saved and introduced as a factor to future queries.

The fulfillment of these aims demands several preconditions and concessions. One is the availability of an index server that keeps at hand all the information needed, so that the search engine does not have to search the World Wide Web itself for the information queried again and again but can fall back on a stored index. Considering that the World Wide Web is a constantly growing, neither already nor yet complete and thus unstable corpus, the problem of information stored on the server expiring or not being available is obvious. Another problem consists in limitations. Beyond the fact that a search result never covers every incidence of the word queried for, the words searched for are not necessarily identical to the actual keywords entered in the query: firstly, ‘stop-words’ (in, on, of, the, single numbers or letters and more) are ignored, i.e., not indexed; secondly, search engines rely on a so-called ‘similar words match’, based on the linguistic
method of stemming. Stemming is the (computer-)linguistic term for reducing (inflected) words to their stems or root forms and searching for these as well as for all its possible inflections and even derivative morphemes; e.g., the stem of ‘children’ would be ‘child’, possible results therefore might include ‘children’, ‘childish’,’childlike’ but might also include ‘childless’ or ’childe’ and are stored on a ‘lookup table’ on the server. This not only implies a search for words beyond their grammatical setting but also leads to a broadening of the results that treats the different inflections and derivative morphemes synonymously. In contrast to a search engine working with a closed set of data that refers to a limited indexed corpus and therefore limits the field of results to the possible words, a stemming algorithm that is introduced to a corpus like the World Wide Web tends to be either over- or understemming words, owing to the fact that the algorithm does not necessarily retain stem words in the semantic relation to inflected word in its index.

Finally, it should be mentioned that search engines do depend on the user. Not only does a search engine save the queries of users and the choices they made from the results offered and learns from this. Also the correct use of operators by users leads to more detailed and exact results and therefore alters future results, helping the search engine to interpret (i) the user’s query and (2) the index in the light desired by the user.

We may thus note the following key features for web-based search engines: Web-based search engines pursue the goal to (a) search the World Wide Web for a keyword or a set of keywords. They thereby (b) put the material available into an order. The search is (c) based on an index that is produced by the crawlers that are part of the system of the search engine itself. The way the (d) index is searched and results are given relies on an algorithm that is not fully known to the user. Therefore (e) the user is not searching the index himself, recurring to his own means and criteria of interpretation, but (f) the search engine is an agent. Keywords are furthermore (g) limited and (h) altered due to stemming and other processes. Search engines tackle (i) a constantly changing corpus but are (j) able to ‘learn’, i.e., to implement information gathered from earlier search processes.

**Same but different:**

**Retrieving knowledge, scraping information and sorting data**

Against this background, and the two sets of key features in particular, several parallels can be drawn between concordances and search engines. First of all, we may state that the World Wide Web is, although it is not a major single work of
our culture, the most present medium of our times. If we understand the World Wide Web as a (possibly unstable) corpus, it would therefore deserve, we may argue, a concordance, ‘a book which shows in how many texts of scripture any word occurs’, to return to Johnson’s definition once again. As a matter of fact, the search engines that compile data since the 1990s (and in the way we know today since approximately the year 2000) surely allow us an easier access to works or information by enabling us to search for the occurrences of words or subjects, for the material available on the World Wide Web. As frequently proven, web-based search engines enable us to identify passages and even writings by entering keywords.32 So they in many ways serve as a form of concordance already, allowing the user to use the results for further purposes, such as interpretation. Thus if we take into account what, on a surface level, search engines do, it appears that they are a concordance of the World Wide Web. If, furthermore, we consider that from the material available on the World Wide Web, most existent concordances and many others could be compiled by using search engines, it seems that we may even consider web-based search engines the concordance par excellence. Or, at least, a highly potential concordance.

Having said this, it is just as vital to point out the profound differences between concordances and search engines, some of which become clear if one considers the operating methods of both, others if one looks at the idea underlying their concept. Many of these differences might be seen marginal in everyday usage; however, when we ponder the question whether a search engine may be referred to as a concordance, these aspects require scrutiny.

I shall start with the limitations applied to the actual content of the query itself and the problems that arise from it. As pointed out earlier, search engines, by means of stemming or of not indexing stop-words, modify the search to improve its results. Unless put into inverted commas, the keyword is altered through stemming or even ignored when being a stop-word. Now, considering that one of the denotations of a concordance – and in fact the incentive for the very first one – was that the different usage of words throughout a work could be compared, we recognize this as a central disagreement between the two concepts. The same holds for the tackling of stop-words. Articles, particles, conjunctions and other small but often-used words would surely downsize the relevance of the results of a search-engine search. But historically, as we have seen before, and also semantically, they are essential.33 Moreover, it would be erroneous to think that a search engine allows us to search strings that are not presorted in a ‘lexematic’ or even ‘semantic’ way. Yet again: if a concordance is an instrument for interpretation and answering questions on the text or the content underlying it rather then about it, the precision as much as the complete and equal representation of a corpus must be considered essential.
As for the resorting to an index listing all words and the pages on which they occur, we touch on another problem that leads us beyond the materiality of the search engine or the concordance itself, approaching the actual idea that is behind a concordance. We might state that the index that is retrieved by a crawler and resort to by the search engine when a user conducts a query can be understood as a concordance. But we have to be aware of the fact that this turns the search engine itself into a kind of intermediary between the index/data and the user (who enters the query). The task of this intermediary agent is then to interpret the query as well as the matches in the index, which at times – as with any intermediary agent – can be a challenge. In fact, it can be so in multiple ways: firstly, the bigger and more open to changes a corpus, the broader are its results and the more it is likely, in comparison to a search in a concordance that is applied to a single corpus or a limited field, to come up with numerous results that might not fit the query. While the ‘user’ of a traditional concordance might, for example, just search for the word ‘honor’ and its occurrences in *Troilus and Cressida* using a concordance of Shakespeare’s complete works, the user of a search engine is depending on the search engine to interpret his query in that intention and to show her twenty-two appearances of the word – instead of an interpretation on the question of honor in the play, which is all that ‘my’ Google offers me for the first thirty-eight results. In fact, this makes obvious, again, the problem of displaying information that can be found in the text or the content underlying it as opposed to information about it. At the same time, an engine that corresponds to a large, ever-changing corpus is supplied with an algorithm that evaluates the user’s picks, and learns. So can the user: every search engine can be optimized by the user’s effort to learn how to use operators correctly and therefore get more detailed and exact search results. So overall search results can be ‘improved’ – which is, of course, not the case in a concordance, but there would be no need for it anyway, assuming it was sufficiently assembled and the canon established. And it would furthermore not be in the interest of a concordance to represent previous searches and filter the results in favor of these.

Secondly, the interpreting intermediary agent can be bribed, which neither the concordance nor its author nor its user are likely to be. The objective and universally transparent and understandable criteria mentioned before dictate a concordance to be sorted in an alphabetic order. The order of the listing of the passages would then be defined by the position of the occurrence in the book, etc. In web-based search engines, this is not the case. However, the intermediary agent may try to interpret the query and the index on a neutral basis, so the index is constantly under attack by external parties who try to improve their own presence in the search result rankings, due to the fact that the higher an entry is listed the more often it is clicked by a user. Search engine optimization (SEO) is a flourishing area of business. Therefore, in this respect as well, search engines
do not seem to be objective scholarly concordances but highly hierarchical ones. This, in a way, makes their use much less deliberate than we might wish it to be. Although this is only one aspect, it seems to be a substantial one.

From what has been said it also becomes clear that a web-based search engine not only does not distinguish between primary sources and works that are already an interpretation, but Google’s latest developments, the so-called ‘knowledge graph’, goes even one step further by actually conveying the user that this is ‘the truth’: Offered as an extra search result, displayed in an emphasized way this search result seems immune to interpretation – an idea that stands in opposition to the concept of a concordance.

**The potential of reading search engines as a concordance**

As of June 30, 2012, the number of World Wide Web users was calculated to be 2,405,518,376, which at that point equaled 34.3% of the overall population of the world. 67.7% of the online population were physically living in Asia, Europe, or North America, in the latter with almost 80% coverage in the population. Given the fact that the indexed web contains at least 14.48 billion pages today and is continuously growing, it may appear that we are looking at an infinite, unsorted, unmanageable, and, even more important, not understandable amount of information. Thus it is hardly surprising that a search engine, namely Google, is by far the most visited website in global rankings, receiving over 2,000,000 queries per minute, suggesting the impression that most of the world’s population is hardly able to imagine using the World Wide Web without using search engines – maybe simply because they are overwhelmed by the amount of data and information available.

Such an interpretation in the face of new technologies has been dramatically declared by authors such as Neil Postman. The dissemination of information, according to Postman, is nowadays (and this still holds today) no longer the problem; instead, he argues, the problem today is ‘how to transform information into knowledge, and how to transform knowledge into wisdom’, knowledge being defined as organized information – information that is embedded in some context; information that has a purpose, that leads one to seek further information in order to understand something about the world. Without organized information, we may know something of the world, but very little about it. When one has knowledge, one knows how to make sense of information, knows how to relate information to one’s life, and, especially, knows when information is irrelevant.
Besides brooding fears among scholars of vanishing paper work, libraries and corpora,\textsuperscript{16} the biggest problem that I see, however, seems to be connected to the ‘users’ and a losing of the ability to make sense of what is left when there is just data – or even information – and concerns understanding and making sense of information and turning it into knowledge. Applications as the ‘knowledge graph’ are actually supporting the approach that a search engine as an agent can supply information that seemingly lies beyond interpretation. That users accept this as factual knowledge might be a consequence of an increasing lack of procedural knowledge on how to retrieve, and critically review, the information provided; for example, by applying long-learned practices of research and interpretation. But it may just as well be a lack of awareness that the results we are offered by web-based search engines such as Google are neither objective nor based on a fully transparent methodological approach. Moreover, there may be a lack of awareness of the fact that the search engine is already offering an interpretation of the listings found in an index: The search results represent a statistical harmonization rather than a harmonization reflecting the meaning or the content as a concordance would. Again, a simple example might come from Shakespeare: In his play 	extit{Troilus and Cressida} the word ‘lord’ comes up a hundred times, the word ‘honor’ twenty-two times, ‘love’ appears fifty times, but ‘hierarchy’ is not used once. As a result, it seems, the index in a concordance would possibly categorize the play as being mainly about a lord (or lords), about the experience of love, somewhat less about honor, and not at all involving hierarchy – an inaccurate assessment to be sure.

Without doubt, the World Wide Web is an outcome of the technologies that are as well a representation as an expression of our current culture. Of course, it is one of the basic ideas of cultural studies and the social sciences that the change of culture and society are connected to the change of technologies (Innis, Benjamin). Media are no longer only instruments to spread or to bear information but ‘apparatuses’ (Benjamin), ‘codes’ (Flusser), ‘dispositivs’ (Foucault), ‘extensions’ (McLuhan), or 	extit{Aufschreibesysteme} (Kittler), to mention but a few approaches, and thus are themselves sources of cultural, scientific and social practice whose influence we can not ignore but approach in a way that valuates them as such. Criticism geared toward the digital humanities and the practices of the ‘digital age’ is mostly concerned with its methods, its phenomenological and psychological impact, and the focus that is put on pure data rather than what one might call content; finally, many scholars question its reliability and sustainability (most prominently by Fish, Fitzpatrick and Ramsay). So far, however, neither the critics nor the champions of the field have exhaustively addressed the possibility that the effects of the digital technologies on the humanities and scholarship more generally their new ‘learned practices’ and ‘habits’, can be considered as no-
table features able to reflect the changes and impacts themselves. Precisely this, to compare the practices of the ‘traditional’ humanities to those of the ‘digital’ humanities in order to better understand the shifts and changes that the transition from one to the other produces, is an approach I think valuable if we want to grasp not only developments in a ‘digital age’ but also a ‘making of the humanities’ as we experience it today. Rather than merely celebrating or criticizing the assumed potential of a great renewal that can counter the often presumed ‘crisis of the humanities’, the task might be to produce a genealogy of methods and their media. To me it seems promising to not only assume an ‘exceptional promise for the renewal of humanistic scholarship [...] enabled by networked, digital environments’ but to approach new technologies using older ideas and concepts now hidden behind them in order to see both of them in a different light: Models and practices such as, for instance, concordance, but also, for example, source criticism have lost nothing of their significance. And they correspond to practices and techniques on the Web and in digital technologies (such as search engines, the Text Encoding Initiative (TEI), topic models, stylometrics, to mention but a few); at the same time, however, they can serve to make problems and derivatives in methodologies and interpretations of both, the digital humanities and the humanities, transparent. Turning around the convoluted project of the digital humanities and ask for a new ‘humanities of the digital’, comparing practices of the humanities with the altered forms that result from the mingling with digital practices and the influence of multifaceted technological innovations seems promising, also to analyze concepts of knowledge and understanding in the ‘digital age’.

My claim would be to approach new technologies using the idea behind, for example, a concordance as a vehicle. We may surely consider the results as bundles that have been nicely brought together, each with a title and indication of the source, the webpage – ‘Google, qui magnas super World Wide Webem concordancias compilavit’, we may say – but only to continue, putting the results into a relation to our query and interpreting and scrutinizing these bundles as carefully as we would the indices of Frere Hewe, Cruden and others; and from there, turn to the passages displayed and again start to search for meaning. Because that, the Erkenntnis des Erkannten (in the words of August Boeckh), will always be on us, the users.

Notes

2 Samuel Johnson, ‘Concordance’, A Dictionary of the English Language (Heidelberg, 1828), vol. 1, 209. This is understood as the first definition of the term in a dictionary, given in 1755. Note that a concordance is not a dictionary, but a dictionary is a type of concor-
dance. Johnson’s referring to the ‘(Holy) Scripture’ is owed to the circumstance that the first word concordances were made for the Bible, which was the only book considered true throughout and thus worthy. Other writings that may be called early concordances (though mostly in the sense of a ‘harmony’ of numerous writings mainly of interest to textual criticism) are the Concordantiae Morales (attributed to Antony of Padua, possibly around 1225), the Concordia discordantium canonum (Gratian, most likely around 1140) and Origen’s Hexapla, an extensive comparative study of various translations of the Old Testament, compiled in the 3rd century. There were also concordances for other central religious writings, such as the Qu’ran, the Torah or ancient Chinese texts. This paper, however, focuses on the European tradition.

3 This broadening of an already broad applicability of ‘concordance’ started around the time of the raise of the ‘Geisteswissenschaften’ in the late 18th century. Unfortunately, to my knowledge, as of today there is no comprehensive study of the idea of concordances through the ages and disciplines.

4 Earlier, John Marbeck had completed a concordance, though this must be seen as a index verborum: John Marbeck, A Concordance, That Is to Saie, a Work Wherein by the Ordre of the Letters of the A.B.C. Ye Maie Redely Finde any Words Conteigned in the Whole Bible (London, 1550). (Alternate spellings are Marbecke or Merbecke.)


6 For the history of concordances I am indebted to several introductions to concordances, particularly to Bindseil, Bruder, Cruden, Mandelkern, Mengenot, Rouse and Rouse, who provide an overview of this wide but surprisingly understudied topic.

7 Oxford English Dictionary: ‘Concordance.’ To my knowledge, the term was not used in this sense in any other language before that; French uses it in diplomatic vocabulary, starting in 1190, and in this meaning the term later also occurs in English. Cf.: Denis Diderot, ‘Concord,’ Encyclopédie (Paris, 1751), vol. 1.

8 Later released under the title Postillae in universa Biblia iuxta quadruplicem sensum literalem, allegoricum, moralem, anagogicum.

9 Heinrich Ernst Bindseil, Concordantiarum Homericarum specimen, cum Prolegomenis in quibus concordantiae biblicae recensetur (Halle, 1867), vif (Prolegomena).

10 John of Trevisa here translates a passage in which Higden reflects upon the reign of Pope Innocence IV, mentioning a certain ‘frere Hewe’ who was made a Cardinal Priest by Innocence IV. However, it seems that this coining of the term ‘concordance’ in the context of what we may today understand as ‘biblical criticism’ can be considered a by-product that only in retrospect receives a deeper meaning. Hugh of Saint-Cher’s role in clerical politics under the reign of Innocence IV was surely considered his greatest achievement throughout the Middle Ages, which is why Ranulf Higden mentions him in the Polychronicon. The sentence ‘þat expownede al þe bible, and made a greet concordaunce uppon þe bible’ in my opinion is mainly a predication of Hugh of Saint-Cher. In a similar way this information is treated in the Memoriale potestatum Regniusium, ‘[...] dominus Ugo Cardinalis, [...] qui doctor eximius doctrina sana et praelucida totam Bibliam postillavit; Concordantiarum Bibliae primus auctor fuit.’ Cf. Du Cange et al., Glossarium mediae et infimae latinitatis (Niort, 1883-1887).


12 In his synoptic harmony from around 1140, Gratian brought works of potentially contradictory contents into accordance, thus the title Concordia discordantium canonum.
Cruden, *Concordance to the Holy Scriptures*, preface.

There is good reason why verbal indexes and concordances, for example, on the works of Homer (1656), were given names such as *Clavis Homerica*.

The cover page of the first edition read 'Yaïr Natib' (It/He Will Light the Path, after Job 41.24, the description of Leviathan [A. V. 32]).

This is a common motivation. John of Regusa, for instance, president of the Council of Basel from 1431 to 1449, when getting into a quarrel over the particles nisi, ex and per, realized that these were not part of any concordance so far and tasked John of Segovia with the *Concordantiae biblicae vocum indeclinabilium* (Basel, 1476), cf. Bindseil, *Concordantiarum Homericarum specimen*, xii, which also shows that detailed information on texts was becoming more and more essential.

Kenneth McKenzie, *Means and End in Making a Concordance, with Special Reference to Dante and Petrarch*, *Annual Reports of the Dante Society* 25 (1906), 19-46, esp. 29.


Sometimes these attempts were rather overdoing, as has to be stated. Spevak's concordance on Shakespeare that was compiled by an IBM 7094, for instance, takes into account every single word of Shakespeare's works. Not only does this concordance fail to compile different spellings of the same word in one entry, but in fact it also shows all entries for so-called stop-words such as 'a' and 'the', urging Frank Kermode to close his 1969 review in *The New York Times* close with the line: '[W]e shall learn, *inter alia*, what proportion of the poet's total vocabulary is claimed by such words as *love*, *black*, *power*, and *will*. But his favorite word will, I predict, turn out to be *the* (Frank Kermode, 'The IBM Shakespeare', *The New York Review of Books*, January 30, 1969).


Note that 'search engine' is an umbrella term covering numerous types: web-based search engines, selection-based search, meta search engines, desktop search tools, search engines within applications and with limited corpora, and web portals or vertical market websites offering a search facility for their online databases. This paper focuses on web-based search engines, i.e., specific software codes, designed to retrieve information from the World Wide Web, regardless of their origin. Examples of these are Yahoo!, Soso, Bing, or Google. Again, individual web search engines work differently, mostly with different algorithms. I have chosen Google as a case study for two reasons: (1) Google is by far the most frequently visited website in global rankings, and therefore also the most frequently used web-based search engine. Cf. URL: http://www.alexa.com/topsites; and (2) despite the fact that details of the algorithm itself are being kept secret, the basic idea of how Google works is accessible and well documented.

For a brief characterization of the differences between search engines in light of their history, see Kuyoro Shade O et al., 'Trends in Web-Based-Search Engine', *International Journal of Emerging Trends in Computing and Information Sciences* 3.3 (June 2012), 942-948.

Not to be confused with Google's service 'Ngram Viewer' that is not only limited to the Google Books corpus but also tackles words as grams.

The terminology is somewhat confusing, since this is what one would refer to as an index or even concordance from the aforementioned. 'Inverted' refers to the following: in com-
puter science an 'index' lists the documents and from there accesses the words/signs in it. An inverted index lists the words/signs and the documents available from there.


26 The idea of search being lead by the idea of knowing what should be found is a common place in the history of ideas, see, e.g., Martin Heidegger, *Phenomenology of Religion*, §9, h (‘What does it mean to search’).

27 URL: http://www.link-assistant.com/seo-stop-words.html provides an exhaustive, though seemingly rather exaggerated list.


29 To be clear on this: Google does not list synonyms.

30 Overstemming means that words of the same stem but a possibly different field of reference are brought together; understemming denotes the case where words have a different stem but a similar semantic concept.


32 At least those that are available and searchable online. One may see the various plagiarism affairs of the past years as an example, e.g., http://de.guttenplag.wikia.com/wiki/Tools,_um_Plagiate_aufzusp%C3%BCren.

33 For a possible reading of the value of semantic detachment, cf. Boris Groys, *Google: Words beyond Grammar* (Ostfildern, 2011). Groys argues that Google is the incarnation of a philosophical machine *because* it detaches words from their grammatical limitations, thus crossing boundaries beyond deconstructivism.


36 URL: http://www.worldwidewebsize.com/.


39 Ibid., 93.

40 Cf. Anthony Grafton, *Codex in Crisis* (New York, 2008), who clearly represents a more optimistic position than many other scholars in the humanities and whose fear of the loss of what we may see as a certain reader and reading attitude I fully share.

41 Anne Burdick et al., *Digital_Humanities* (Cambridge, 2012), 7.

42 For a comprehensive study on methods of information management in ancient and medieval Europe, see Ann M. Blair, *Too Much to Know: Managing Scholarly Information before the Modern Age* (New Haven, 2011).
