



Tiresias: A Novel Approach for Mining Book Indices

Moshe Blidstein <mblidstei_at_univ_dot_haifa_dot_ac_dot_il>, University of Haifa  <https://orcid.org/0000-0001-7488-6324>
Daphne Raban <draban_at_univ_dot_haifa_dot_ac_dot_il>, University of Haifa  <https://orcid.org/0000-0003-1791-0310>

Abstract

Tiresias (<https://tiresias.haifa.ac.il/>) is a free database designed and constructed as an efficient tool to access and understand ancient texts for research purposes. It uniquely associates end-of-book subject indices with the index locorum. Based on indices of about 460 scholarly books in fields such as history, religion, biblical studies and more, Tiresias is multilingual, contains about 140,000 subject tags and about 16.4 million references to ancient primary sources, including non-digitized sources. It is especially helpful for comparative research between areas, cultures or languages, or for scholars working on long-term trends, who are therefore less familiar with sources which are not in their area of specialization. The paper describes three usage scenarios employing this newly available resource as an integral part of historical research and study. Tiresias provides inspiration for the construction of other databases making use of multiple book indices in various areas of the humanities and social sciences.

1. Introduction: the challenges

Locating primary source (henceforth: source) material for humanities' research is labor-intensive. Three methods frequently used to perform this task are: 1. Reading large amounts of source texts and manually collating the relevant passages; 2. Reading prior research (henceforth: books) on the subject, collating references from these books, and locating source texts according to these references; 3. Searching by keyword in full-text corpora (usually in the source language, such as Perseus, Thesaurus Linguae Graecae, papyri.info, or Ma'agarim) [Green 2000] [Dalton and Charnigo 2004] [Sinn and Soares 2014].

Each of these methods has certain disadvantages. Many researchers' continued recourse to methods 1 and 2 demonstrates the current limits of full-text keyword search [Green 2000]. Further criticisms of keyword searches describe forcing historians to retrieve texts in a specific way, producing biases, blind spots and false positives [Hitchcock 2013]. "[Historians'] 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" [Huistra and Mellink 2016].

In this paper, we describe Tiresias, a novel database created by juxtaposing back-of-book indices in scholarly books. Tiresias' aims are to provide simple and efficient retrieval of ancient sources by subject, and to be a tool useful for beginners, advanced students and researchers in the disciplines of ancient history, epigraphy, papyrology, theology, and kindred areas, as a complement to full text keyword searches. First, we shall describe how the database was created and the possibilities it affords for the study of ancient history. We then review some of the available databases and their limitations and describe how Tiresias fills the gap of providing a digital search tool while maintaining the intellectual value of book indices.

2. Tiresias

Tiresias (<https://tiresias.haifa.ac.il/>), named after the mythological blind prophet, is a detailed, searchable database of

1

2

3

4

subject tags to ancient texts and artifacts, currently consisting of almost 140,000 subject tags for about 16 million references to sources. The search form allows for searches of several keywords, filtering by various fields (e.g., author, work, subject, keyword), and immediate full-text retrieval for efficient perusal of the relevant sources.

The database was constructed through the following innovative method. Many research books in the fields of ancient history, classics, or biblical studies, are published with two indices: one for subjects (called a subject index, index of terms, or *index rerum*) and one for ancient source references (also known as an *index locorum*). Through these indices, each page of the indexed book was identified as relating to certain subjects as well as certain sources, indicating with a high probability that these source references can be tagged as related to these subjects. This probability was improved by a validation method which we describe and assess in this article. Using this method, we produced a *subject tag* for each source reference, i.e., a short description of a subject highly relevant to the source reference. For example, the source reference “Genesis 1.3” could be tagged with the subjects “light” and “creation.” With the help of a computer program, the tags were combined to create a general database of subject tags for ancient sources. The subject-source database is thus based on existing expert-made back-of-book indices, unified and assisted by digital means.

3. Corpora and methods

The creation of Tiresias can be divided into the following stages:

3.1 Selection and retrieval of books

The database focuses on ancient Mediterranean religion. This appeared to be a relatively well-defined area with many academic books, of which many have both subject and source indices. The books used are mostly by leading publishers in the field (Oxford University Press, Cambridge University Press, Brill and De Gruyter), all of which have a strong presence in digital publishing, from the following areas: classics, biblical studies, Jewish studies, patristics, and ancient history. The Ancient Near East is not included for technical reasons – too few indices meeting the relevant criteria.

There were two essential criteria for inclusion of books: a. the existence of both an *index locorum* and a subject index. b. cheap, convenient, and legal access to a high-quality scan of the indices. It was clearly preferable to use books for which PDFs with embedded text are readily available, in order to obviate the stage of OCR of scanned text. This meant that the books used were usually published after 2000, as indicated in Figure 1. About 600 books were identified as meeting all these criteria and their indices were downloaded via subscriptions of the Younes & Soraya Nazarian Library at the University of Haifa. Additional book indices are continuously entered with expanded availability and continued publication of new books.

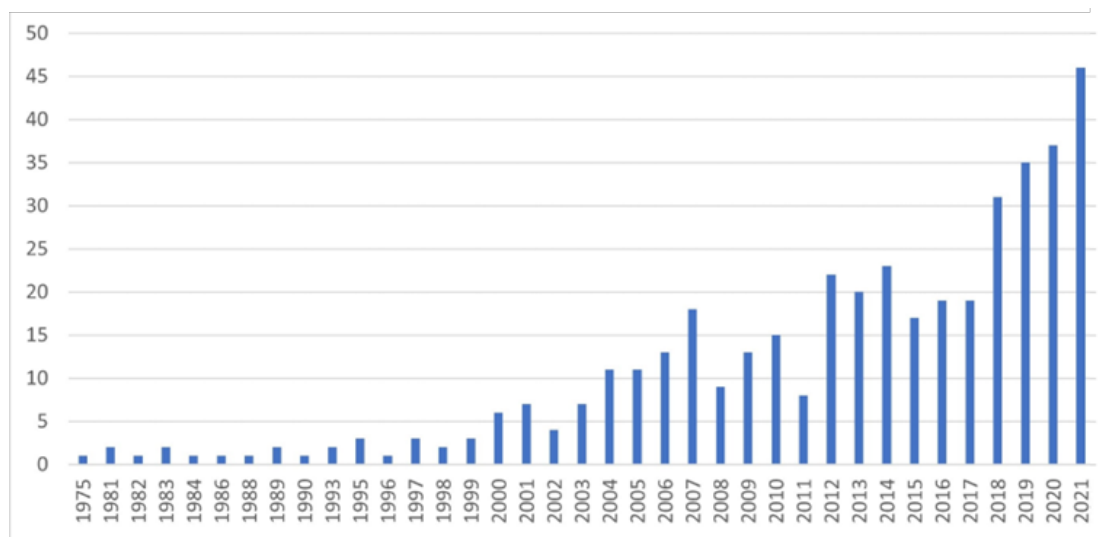


Figure 1. Count of books per year.

3.2 Index parsing

Indices are relatively structured data, and this facilitates reading them algorithmically. Nevertheless, there are many idiosyncrasies of different indices requiring both computerized and human processing to make use of the indices in a shared database. 9

a. Subject indices

Published subject indices are typically constructed of alphabetized subject headings and sub-headings, with the latter indented, as demonstrated in Figure 2. Therefore, in order to include all the information in the index, it was necessary to correctly identify subject headings and sub-headings. This was done algorithmically in Python 3.10, with human input and supervision. The stages were as follows: 10

- 1. Algorithmic conversion of PDF to text format (.txt).
- 2. Human identification of subject index format: whether sub-headings exist; whether each sub-heading is on a separate line and marked by an indent; if not, what character is used to mark sub-headings.
- 3. If the sub-headings are marked with a character, this is used to algorithmically differentiate them from main headings. Otherwise, since the order of the main headings is always alphabetized, a human marks on which row the first letter of the subject headings begins and this marking is used to identify main headings (beginning with this letter) and sub-headings (not beginning with this letter). Also, lines beginning with propositions ("and", "by", etc.) are identified as sub-headings, and additional minor corrections are applied (e.g., identifying sub-headings when they start with the same letter as the heading).

At the end of the process, the subject headings and sub-headings were joined by a comma, and page ranges were expanded. 11

Baptism	100
Christian	90-91
Jewish	93-95
With fire	12-14

Figure 2. An example of a main subject heading and sub-headings, with each sub-heading on a separate line.

Table 1 provides an example of data entry for the output appearing in Figure 2. 12

Subject	Book #	Page #
Baptism	1	100
Baptism, Christian	1	90
Baptism, Christian	1	91
Baptism, Jewish	1	93
Baptism, Jewish	1	94
Baptism, Jewish	1	95
Baptism, with fire	1	12
Baptism, with fire	1	13
Baptism, with fire	1	14

Table 1. Example of subject entry and corresponding page numbers from a book subject index

b. Source indices

Every reference in the source index was parsed and divided into the following columns: 1. author and title identifying number, 2. reference, and 3. Page number, as seen in Figure 3. 13

New Testament

1 Cor.

3:16-17	152-155
5:11-13	151
10:2-3	90

Book of Jubilees

1:23	51
6:6-9	233-4

Figure 3. Standardized Tiresias references vs. the original source references

Table 2 provides an example of data entry for the output appearing in Figure 3. 14

Ancient Work #	Reference	Book #	Page #
8056	3.16	1	152
8056	3.16	1	153
8056	3.16	1	154
8056	3.16	1	155
8056	3.17	1	152
8056	3.17	1	153
8056	3.17	1	154
8056	3.17	1	155
8056	5.11	1	151
8056	5.12	1	151
8056	5.13	1	151
8056	10.2	1	90
8056	10.3	1	90
2309	1.23	1	51

Table 2. Example of source entry and corresponding page numbers from a book source index

Clearly identifying the various ancient authors and works was challenging as publications use highly varied conventions to refer to ancient authors, works, and references [Romanello 2018]. The source indices contained more than 12 thousand unique ancient works from different traditions and languages. For example, The New Testament book, First Corinthians, can be written in various abbreviated forms such as “1 Cor.,” “First Corinthians,” “I Cor.,” etc. Many ancient works have several naming conventions for the title and for the author. Furthermore, the chapter and verse number can also be written according to different conventions, using Roman or Arabic numerals, and different types of dividers. To meet this challenge, we constructed an author-title database of modern naming conventions and abbreviations of ancient authors and works, which the program consults to identify the references included in the indices. This database utilized and collated existing free databases such as the TLG canon of Greek Authors and Works, abbreviation tables from various Greek and Latin dictionaries available online, the Classical Works Knowledge Base, SBL handbook, and other lists. Many additional authors and titles for Greek, Latin, Coptic, Hebrew and Syriac works were added incrementally to the database as more indices were processed and additional authors and titles were identified manually. The script checked each line in the source index against this database to identify authors and titles. When a match was made of both author and title, a work number was assigned. Fuzzy matches were not used because different works can have very similar titles (e.g., 1 Corinthians and 2 Corinthians). Lines which were not identified as authors or titles were assumed to be of references and page numbers. These were distinguished by manually inputting the differentiating character (usually a whitespace) between them. This script generally produces good results, but with some errors as a result of unrecognized authors or titles, or incorrect differentiation of reference and page numbers, usually in cases of unusual references. Therefore, an expert examined the outputs (in csv format) to delete or correct wrong rows.

3.3 Database construction

The main database, in a flat csv file, was constructed by combining the subject and source indices along the page column. This database currently contains 16,420,264 rows. For example, in Figure 3 the source index of a certain book tells us that the 1 Corinthians 10:2 is discussed on p. 90, while the subject index tells us that “baptism, Christian” is discussed on pages 90-91 (Table 1). Juxtaposing this data provides a certain probability that First Corinthians 10:2 deals with, is connected with, or relates to, Christian baptism. 1 Corinthians 10:2 is therefore tagged with the subject tag “baptism, Christian”. The resulting rows in the database are shown in Table 3.

Row #	Ancient Work #	Reference	Book #	Page #	subject
1	8056	10.2	1	90	baptism, Christian
2	8056	10.3	1	90	baptism, Christian

Table 3. Database rows

However, this process can produce false positives for several reasons: Book pages usually discuss several subjects, and source references on these pages may be connected to a specific subject and not to the others. Moreover, references may be supplied as examples for some side point in the discussion, having no link at all to the main subject of the page. Another problem is that references are not always consistent, with different editions dividing the source text in different ways. Therefore, the juxtaposition of indices alone will provide a large proportion of irrelevant, or only slightly relevant, results.

17

3.4 Validation

As a remedy to the false positives problem, we attempt to validate results by the following method. First, we retrieved all source references appearing in more than one book. Then, we extracted all the subject tags for these source references, split them into word tokens, removed non-significant words, and morphologically stemmed the words to capture variants such as plural/singular, or different verb forms. For example, Table 4 displays the unvalidated data, and Table 5 shows the same data after stemming.

18

Row #	Ancient Work #	Reference	Book #	Page #	Subject
1	8056	10.2	7	90	baptism, Christian
2	8056	10.2	7	150	baptismal, spirit
3	8056	10.2	8	70	Spiritual life
4	1100	1.2	9	10	baptism, Christian
5	1100	1.2	9	10	Spiritual life

Table 4. Unvalidated database rows

Row #	Ancient Work #	Reference	Book #	Page #	subject	tokens
1	8056	10.2	7	90	baptism, Christian	baptis
2	8056	10.2	7	90	baptism, Christian	christi
3	8056	10.2	7	150	baptismal, spirit	baptis
4	8056	10.2	7	150	baptismal, spirit	spirit
5	8056	10.2	8	70	Spiritual life	spirit
6	8056	10.2	8	70	Spiritual life	life
7	1100	1.2	9	10	baptism, Christian	baptis
8	1100	1.2	9	10	baptism, Christian	christi
9	1100	1.2	9	10	Spiritual life	spirit
10	1100	1.2	9	10	Spiritual life	life

Table 5. The rows of table 4, tokenized and stemmed

If the same token was found in the subject tags derived from more than one book for the same source reference, this token was considered validated for this reference and the row was entered into the validated table, under the assumption that a match of subjects from two books (even if only in one word) would usually mean that the pairing is significant, and not a mistake or coincidence. For example, as seen in table 4, if the source reference “1 Cor. 10:2” was

19

tagged with the subject tags “baptismal, spirit” in one book and “spiritual life” from another, then “1 Cor. 10.2” will be entered into the table of validated references with the token “spirit”, and the following rows in the validated table could be added as seen in Table 6.

Row #	Ancient Work #	Reference	Book #	Page #	subject	token
1	8056	10.2	7	150	baptismal, spirit	spirit
2	8056	10.2	8	70	Spiritual life	spirit

Table 6. Validated database rows

The two rows in the validated table were derived from rows 4 and 5 in the unvalidated table, which have the same source reference, an identical token (“spirit”) and derive from different books (7 and 8). These are the only rows meeting these criteria.

Currently, the validated table includes about 3.7 million rows. We intend to systematically assess the precision and recall performance of this validation method in future research. In a preliminary examination of 1075 randomly selected validated source-subject pairs, we looked at the tagged source text to assess the suitability of the subject tag. The results are summarized as follows:

1. 355 (33%) of the tagged source texts included the tagged word (e.g., a text tagged by “baptism”, where the word “baptism” appears in the source text);
2. 280 (26%) of the source texts included synonymous or closely related words (e.g., “immersion” or “initiation” in the same case);
3. 340 (31%) of the source texts included discussions or reference to subjects relevant to the subject tag, (for example, a scene from the Hebrew Bible which was later explained by Christian writers as relating to baptism or a text discussing pagan rituals similar to baptism).
4. 100 (9%) of the tags were a mistake or unclear – 28 (2.6%) an unclear reference, 10 (0.9%) other mistakes, 59 (5.4%) references which happened to appear on the same page as the tagged subject but were not connected to it.

Though this data is preliminary and more rigorous methods and larger sample sizes are needed, it indicates that less than 10% of the results were irrelevant (article 4 above), while a third of the results would have been identified also through full-texts searches (article 1); more than 50% of the results are relevant to varying degrees and would not have been located through full-text searches (articles 2 and 3).

Though only ~23% of the results are validated, this percentage will presumably rise with the expansion of the database. Beyond database size, two other reasons for the moderate validation rate are: a. subject tags which are semantically similar but not identical were not identified by the validation script; b. references to the same text which used different conventions and are not covered by the author-title database described above (Section 3.2b) and were therefore not identified as identical. The validation rate can be improved by solving these problems i.e., by semantic pairing, which will allow matching non-identical but synonymous subject tags, and by expanding the author-title database.

3.5 Online search interface

Finally, the database was made freely accessible to the scholarly community on the web with a search interface (<https://tiresias.haifa.ac.il/>). The main search currently offers three options:

1. for validated references only, which provides a list of all validated source references tagged by subjects containing the searched keyword, sorted by date of the sources.
2. for all references (i.e., unvalidated and validated combined), which provides a list of all references tagged by subjects containing the searched keyword, sorted by date of the sources.
3. for secondary literature, which provides a list of all subject tags containing the searched keyword and the

book pages from which these subject tags derived, without the source references. This provides a much shorter list focused on the subjects rather than the sources, and is sorted by alphabetical order of the subjects.

Search results can be filtered by one or two subject keywords, as well as by ancient author, work, or according to the century the ancient work was written. Filters are essential as many subject searches return thousands of results. Figure 4 shows that each result includes an ancient work reference (author, title, and internal reference); the approximate date of the ancient work; the subject tags for this specific reference in the database; the books and page numbers from which the subject tags derived; and the texts of the ancient work reference, in the original language and in English translation, when available.

25

7. **Josephus Flavius, *Jewish Antiquities*, 18.21** (1.0th cent. CE - 1.0th cent. CE)

Tagged with subjects: • De vita contemplativa, women in • Essenes, oaths • Josephus Essenes, and women • Josephus Essenes, oaths of commitment • Philo Essenes, and women • Therapeutae, women inclusion of • women • women and femininity, among the Therapeutae • women, and the Essenes • women, as community of wives

Found in books: Klawans (2019) 63, 64; Taylor (2012) 100, 101, 197, 198; Taylor and Hay (2020) 59, 102

18.21. καὶ οὐτε γαμετάς εισάγονται οὐτε δούλων ἐπιτηδεύουσιν κτῆσιν, τὸ μὲν εἰς ἀδίκιαν φέρειν ὑπεληφότες, τὸ δὲ στάσεως ἐνδιδόναι ποιήσιν, αὐτοὶ δ' ἐφ' ἑαυτῶν ζῶντες διακονίᾳ τῇ ἐπ' ἀλλήλοις ἐπιχρῶνται. / 18.21. and neither marry wives, nor are desirous to keep servants; as thinking the latter tempts men to be unjust, and the former gives the handle to domestic quarrels; but as they live by themselves, they minister one to another.

Figure 4. Example result of search for “women” and “oaths” in the unvalidated option.

oath, royal	Stavrianopoulou (2013) 122, 126
oath, sacrifices at oath-taking	Ekroth (2013) 42, 44, 48, 120, 158, 223, 252, 253, 259, 271
oath, seriousness of	Feldman (2006) 665
oath, solemn, in holy military service	Griffiths (1975) 15, 255

Figure 5. Example results of search for “oath” in the secondary literature option.

In order to facilitate study of the texts, more information was included:

26

1. Links were created for each subject tag to the relevant pages of the book in Google Books. This allows the user to go directly to the secondary literature which supplied the subject tag. While this is not its main function, Tiresias becomes a gateway not only to ancient sources but also to the scholarly literature (albeit only in books, not articles). Though Google Books allows only partial access to the books, this is currently the only way known to us for general users to directly access non-open access books online through a link to a specific page.
2. The ancient text in its original language (Greek, Latin, Hebrew, Aramaic, Arabic, Syriac, Coptic) and/or its translation to English is also presented, in cases in which the text is readily available in machine readable format – about 50% of the references. The user can thus search for a subject and immediately read the relevant texts, without the need to search for the text in other sites or offline. Full texts for Hebrew Bible and rabbinic texts is kindly supplied by Sefaria; for Greek and Latin texts, by Perseus Scaife; for the Quran, by Tanzil.net; for Syriac, from the Digital Syriac Corpus; for Coptic, from the Coptic Scriptorium. These projects provide large amounts of texts with TEI markup, which is essential for extracting the texts according to reference. Full texts for epigraphy and papyri is not yet supported but is planned in the future. English translations were provided for Rabbinic texts by Sefaria; for Greek and Latin texts by Perseus Scaife and Bill Thayer's website Lacus Curtius; for early Christian and Patristic texts, from publicly available 19th century translations.
3. A search form for searching by reference – the user enters an ancient source reference and receives all the subject tags and book pages concerning this source reference, as well as the full text itself when available. This feature is useful for studying specific ancient texts rather than a subject.

3.6 Visualizations

We used the database to allow users to create network and heatmap visualizations on subjects or works of their choosing. The visualizations help users understand how the subject tags in the database are connected to each other, and thus identify links of which they were unaware. Each type of visualization has its advantages: networks facilitate understanding of the degree and nature of connections between various subjects and their relative importance, while heatmaps can be used to understand changes in these connections over time. For ease of comprehension and to prevent visual overload, both types of visualizations were limited to validated subjects only. Furthermore, subject tags were split into one-word terms, so that the visualizations show relationships between one-word terms rather than whole subject tags.

27

Network visualizations

To perform network visualization, we first manipulated the files to create a network graph based on shared subjects and references. The validated table was grouped by references and the number of shared references between each subject was found. For example, the rows in Table 7 would produce the network graph in Table 8:

28

Row #	Ancient Work #	Reference	Book #	Page #	subject	token
1	1001	5.6	125	13	Evil spirit	spirit
2	9023	10.10	140	75	Holy spirit	spirit
3	1001	5.6	120	12	Water, from wells	water
4	9023	10.10	140	75	Holy water	water
5	1001	5.6	130	12	Sacrifice, of bread	bread

Table 7. Rows from validated table

source	target	Weight (=number of shared references)
water	spirit	2
water	bread	1
bread	spirit	1

Table 8. network graph derived from table 6

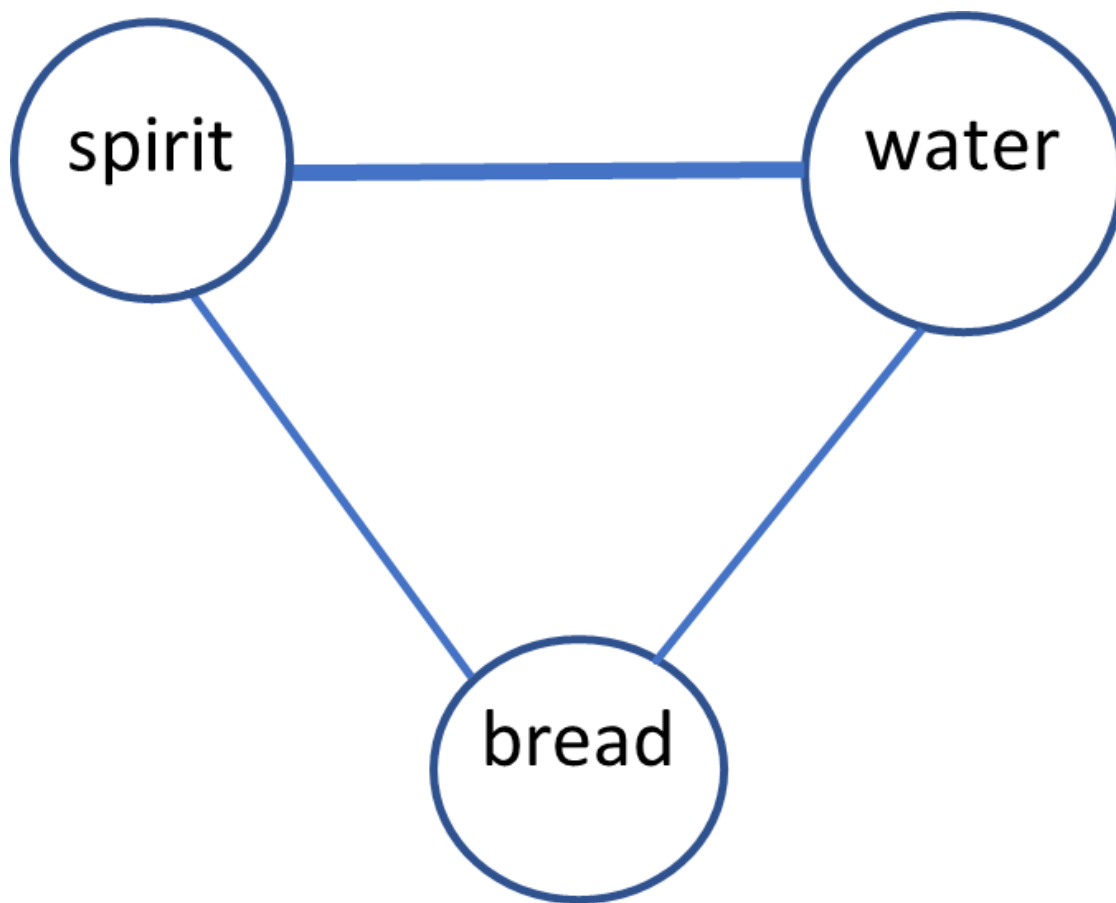


Figure 6. schematic visualization of network graph example from table 6

Using Gephi [Bastian et al. 2009] we visualized the network graph of subjects, as seen in Figure 7. In this visualization of the graph, words from subject tags which share the most references are adjacent on the network, while subjects not sharing references are far apart. The size of each node is determined according to the number of incoming and outgoing links (edges), so that subjects (nodes) with the most source references appear largest. Furthermore, communities of closely connected subjects are determined using the Louvain method for community detection and marked by color. This creates a mapping of the whole field based directly on the scholarly research, intuitively displaying the centrality or marginality of specific subjects, the relationship between subjects, and the different sub-disciplines of the study of ancient religion. The visualization is also instructive concerning subjects which are on the borders between sub-disciplines, that is, which are shared by one more of these sub-disciplines, as opposed to subjects which are in the core area of each. Figure 7 illustrates the part of the subject network dealing mainly with core subjects of two ancient religions - Jewish (in purple) and Christian (green), and some subjects connecting both areas.

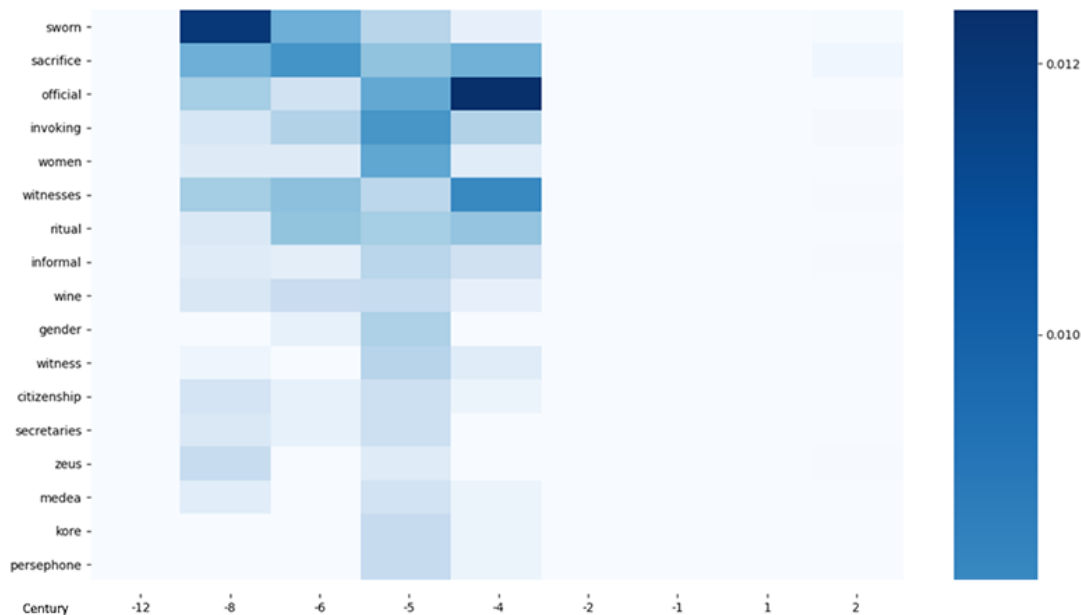


Figure 9. Linked subjects heatmap for “oath” from the 12th century B.C. to the 2nd century C.E

With the expansion of the database, heatmaps can serve to understand how the connection between subjects changed historically over time, providing insight on historical processes or on the scholarship on which the database is based. For example, Figure 6 indicates that most of the studies of oaths focused on Classical Greece rather than on later periods. It may also indicate that the sources speak of women swearing oaths only in sources from the 5th century BCE and not earlier, or that oaths by goddesses Persephone and Kore were less popular before this period. The heatmap cannot, of course, demonstrate actual historical trends, but it can provide pointers, questions and intriguing phenomena which can then be followed up by research.

4. Comparison with existing tools

In order to contextualize Tiresias as a tool in the digital humanities toolset, this section discusses several other tools including full text search tools, major subject indices of secondary literature, and other relevant online tools.

Several excellent online tools provide access to ancient primary sources in full-text format with search capabilities. Prime examples for these are the TLG for ancient and medieval Greek works (currently including more than 110 million words from over 10,000 works); the Brepols Library of Latin text (over 60 million words from 3100 works); *Ma'agarim*, for Hebrew works (over 9 million words from 4500 works as of 2008); CAL lexicon, for Aramaic (3 million words). Many other databases exist, but these examples stand out for offering parsing, lemmatization, searches by lemma, advanced searches with wildcards, filtering by author, period, genre, and other tools. Also notable is the Perseus site, which is at the forefront of facilitating access to Greek and Latin texts in the original and in translation, allowing the insertion of commentary, browsing and searching by reference, and linking to relevant artworks. A large percentage of published Greek and Latin papyri and epigraphical material are online via Trismegistos, PHI searchable Greek inscriptions, Papyri.info, and other sites, with lemmatization and extensive metadata for each of the documents. However, all of sites hardly provide information on the subject of the text, and therefore do not allow for searching by primary texts according to subject, or, in other words, searching through a controlled vocabulary as opposed to a free text search (unusual in this regard are the papyri search sites, where the metadata of each document occasionally includes a description of the text and keywords of its subject, usually in German. Research has shown that user retrieval of information is best accomplished through a combination of controlled vocabulary and free text searches, as these methods retrieve different types of data or have complementary functions [Croft 2002] [Gross et al. 2015].

As opposed to primary sources, much effort has been expended in the past decades on subject indexing and categorizing of secondary literature by controlled vocabularies. An important function of all the online databases for

articles and book chapters (e.g., JSTOR, Wiley Academic, Atla databases, *L'Annee Philologique*, RAMBI, Index Islamicus, etc.), as well as of library catalogues, is to help the user locate secondary literature on a specific subject. There has been extensive research on best practices of subject indexing in the context of such databases and catalogues (see [Yu and Young 2017]; [Landry et al. 2011]) as well as thought on alternative visualizations of indices and searches [Merčun 2016] [Włodarczyk 2013]. A central question that arises from this research is how to create comprehensive, efficient and understandable subject trees, while minimally imposing the viewpoint of its creator, in order to assist researchers rather than constrain them to a specific taxonomy. Researchers have recommended basing subject indices on topic maps, semantic webs or universal thesauri, which represent the language as a whole, rather than ad-hoc subject headings [Włodarczyk 2013] [Nilbe and Tarkpea 2014]. This question is even more significant in the case of a primary source index, as the source texts are the objects of study and not only, as in secondary literature, assisting instruments. In Tiresias, we currently use the subjects presented in the indices without any attempt at constructing a subject tree or topic map as an aid for subject retrieval. However, in the future we may attempt to parse the structure of the subjects and sub-headings of indices themselves as an innovative method to construct subject trees or ontologies of specific scholarly fields. The source indices alone can also be used to build innovative co-citation networks of source and research literature. See [Blidstein and Zhitomirsky-Geffet 2022] for an analysis of the different options based on the indices from the Tiresias project

A number of digital tools have been developed in the past decade for linking ancient sources to secondary literature or to other ancient sources. Biblindex is an index for locating quotations and allusion of biblical verses in early Christian literature, created at first manually and then using dedicated software. The Proteus project developed a quotation detection tool for locating quotations of classical Greek and Latin texts (in original and English translation) in the Internet Archive, which includes 16 million open access documents. The Tesseræ project [Scheirer et al. 2016] has developed tools for comparing identical passages in two ancient texts, and even for locating similar passages which are not verbatim quotes through topic modeling (however, the tool does not actually identify the topics discussed, only the similarity in vocabulary of two passages). The Cited Loci project locates all the articles on JSTOR discussing specific words or lines of an ancient text, and then allows reading the text in the central pane while scrolling through the relevant paragraphs of the articles discussing it in a side pane [Colavizza and Romanello 2019]. These projects demonstrate the current scholarly interest in utilizing computation to locate, compare and link ancient and modern texts, but also that these tools are based directly on the lexical level of the texts or on quotation, neglecting the level of topic or subject.

A small number of projects have attempted to harvest back-of-the-book index data for various aims. Piotrowski [Piotrowski 2010] has used place names listed in indices of a Swiss law book corpus in order to geo-tag the texts and has discussed some of the challenges and possibilities of utilizing indices in this way. Romanello, Berti, Babeu and Crane [Romanello et al. 2009] have discussed how to extract information from indices of critical editions to locate and identify fragments of ancient authors. Both publications provide important technical information on the tools needed to read and mine the information in indices. Michael Huggett and Edie Rasmussen [Huggett and Rasmussen 2013] constructed a small meta-index of subjects from the indices of a hundred open access books to allow users to approach the relevant pages of these books directly from the meta-index. They also studied the responses of users to their meta-index. This publication is especially useful regarding the challenges this project encountered in unifying different indices and their solutions, as well the features users most appreciated.

5. Tiresias in Context

Tiresias has several salient advantages compared to existing tools:

1. Tiresias is not based on keyword searching of full text. Rather, it retrieves the texts based on the viewpoint of scholarship, filtered through book subject indices. It therefore provides a radically different perspective than that currently used in digital historical scholarship, and indeed in scholarship and text retrieval in general.
2. Tiresias is derived from a wide range of secondary sources, from many disciplines. It can therefore facilitate study of various subjects across periods, cultures and languages and help in reducing barriers between historical sub-disciplines. It can be especially helpful for comparative research between areas, cultures or

languages, or for scholars working on long-term trends, who are therefore less familiar with sources which are not in their area of specialization.

3. Tiresias' sources are multilingual. In other words, Tiresias's method provides English subject-based access to primary sources from different languages. This can help with the inherent difficulties of researching a subject in texts in various ancient languages, all of which require years of training for full proficiency.
4. Tiresias facilitates access to texts which have not yet been digitized. Despite ongoing digitization efforts, many ancient texts are still in print or in manuscript only, and/or were not translated into European languages. This is especially true in certain languages such as Aramaic, Coptic, Persian, Georgian, Armenian and Arabic, but even Greek and Latin works are far from fully digitized. These texts are therefore less accessible to the general researcher and even to the expert, who is frequently simply unaware of the contents of many texts (or even of their existence). The database helps researchers recognize the relevance of these texts for their subjects and seek them out.
5. As is usual in library catalogues, searches can be narrowed down by crossing two or more subjects; they can also be narrowed down by details such as author, date, religion, region or language.
6. While most library items or journal articles are tagged with 5–10 keywords, book indices typically include hundreds of subject headings, broken down into sub-headings. Therefore, Tiresias provides granularity far beyond what is now typically provided for secondary literature.

6. Usage scenarios

We shall present here three usage scenarios for Tiresias by researchers, students and teachers of ancient history. The use cases serve to exemplify usage for various skill levels and needs. Researchers use the full functionality of the database to surface unique patterns that give rise to research questions. Students may suffice with the search functionality to retrieve primary sources as required for term papers. Teachers can use Tiresias to show learners that novel technology can help uncover historical knowledge and they can give their class interactive assignments to promote curiosity and literacy.

Case no. 1: Research

A senior scholar on ancient religion is writing a book on oaths and swearing in antiquity. Searching for “oath” or “swear” and their derivatives in full text databases returns thousands of results. Therefore, to start their research, together with reading scholarship on the subject, they turn to Tiresias. Here they use several dimensions of the database to gain a general view of the subject in contemporary scholarship.

First, they search for “oath” among the subjects indexed by Tiresias. “Oath” appears in 621 different subjects. This list provides the scholar with a detailed overview of the different sub-headings in which oaths in antiquity have been discussed in the scholarship: for example, “language of oaths, and gender”, “homicide trials, oaths in” or “blood rituals, surrounding oaths”.

Second, they create a linked subjects heatmap for “oath” by entering the keyword in in the relevant search form and receives a heatmap graph showing the subjects sharing most references with “oath” in the database, sorted by century of the reference (see Figure 9). Here, they see that “sacrifice”, “women” and “witnesses” and “law court” are subjects strongly linked with oaths. Moreover, they see that almost all the references linked to oaths are from before the 4th century BCE. This can lead to the research question: are oaths in decline after this period? Or is there not enough scholarship on oaths after the 4th century BCE?

Pursuing these questions and a select set of sub-headings, they decide to limit their research to oaths by women. They search Tiresias for references tagged with both “women” and “oaths”. Currently there are 13 validated results for this search, and 220 unvalidated results (see above for an explanation of the validation process).

They start from investigating the validated results using the texts provided and can also go directly from the results to the modern publications in which they were discussed. After studying these, they turn to the unvalidated results, which were mentioned only in one book as connected to the subject. Some of these results do not appear relevant to the

subject and can be discarded, but some of them are relevant, and considering their large number this facilitates location of relatively less known texts on the subject.

Case no. 2: BA student

A BA student in religion, writing a term paper on conceptions of resurrection from the dead in early Christianity. The student reads scholarly literature on the subject, but as this is a large subject is finding it difficult to cope with the large and varied literature. Furthermore, the term paper instructions called for engagement with primary sources on the subject, and (beyond the New Testament) they are finding such sources difficult to locate. They turn to sites for searching the full text of Greek and Latin literature, and search for “resurrection”. This supplies them with some texts for analysis, but these are limited to cases in which the keyword explicitly appears in the text, and only in translations of Greek and Latin texts; furthermore, some of the cases turn out to be not actual discussions of resurrection but random mentions.

45

Turning to Tiresias, they search the same keyword among the validated results. They receive 76 results, with texts from the Hebrew Bible, second temple literature, the New Testament, Church fathers and rabbinic literature. They can read these texts in the original language (Greek, Latin, Hebrew and Aramaic) and in facing English translations. These results are not limited to cases in which the keyword appears in the text, but include texts discussing the subject without the keyword. Furthermore, all the results are relevant to the subject. Of these, they choose two ancient sources for further study, assisted by the detailed sub-heading tags. After choosing these, they click on to the modern publication in which these ancient sources were discussed, to find scholarship on the subject.

46

Case no. 3: Teaching

An instructor is building a new module on the history of emotions in antiquity. The study of the ancient sources is central in the module, as they are also interested in a broad, comparative view of several ancient cultures. When thinking of the general subjects and structure of the module, they turn to the heatmap and network visualizations to gain a general feeling of the subject. Here they see connections with Stoic and Platonic thinkers, but also the connections between emotion, the body and nature (Figure 10). Searching for “anger”, “love” and other emotions provides more information.

47

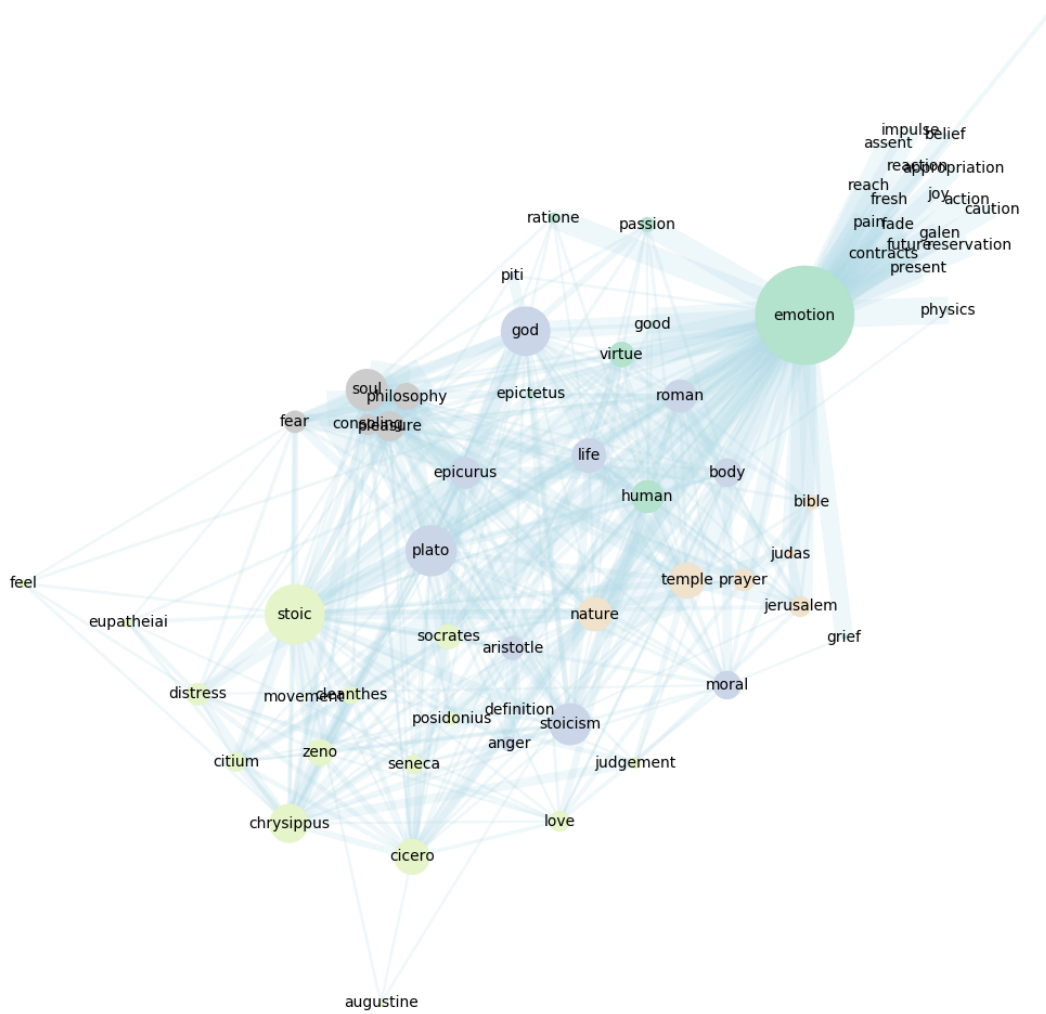


Figure 10. Network tool result for “emotion”

More concretely, searching the database for “emotion” leads to 54 validated results, from a wide range of periods and genres, as well as links to the modern publications. Some of these they find useful for background discussion, others for study together with their students, with texts and translations already available. Here too, searching for several other terms linked to emotions (some of them obvious such as “love” or “fear”, others found from the network, the heatmap or the subject subheadings of the validated results) leads to further texts on more specific subjects.

48

In class, the instructor asks their students to search for the term “anger” on Tiresias and to explore the results (for digital humanities usage in the classroom, see [Locke 2017], [Fyfe 2018]). This leads to a discussion of the validity of using english terms of emotion to describe ancient phenomena and to how using different types of indices can direct our focus and structure our worldview, as well as more specific questions concerning the validity of the the method used by Tiresias to locate subjects and texts.

49

In summary, the usage scenarios were developed based on our experience in academic research and teaching. Scholars are likely to become highly skilled in using the continually developing functionality of Tiresias and adopt it as one of their standard tools for aiding research based on primary sources. Users with simpler needs can also benefit substantially from access. For example, the BA student will realize that studying in the humanities can be supported uniquely by innovative digital technologies that were developed in academia, thus forming a mental link between humanities and innovation. Further, usage of Tiresias in addition to library databases will sharpen the understanding regarding the difference between primary and secondary sources. The teacher would be able to show younger learners that novel technology can help uncover historical knowledge and that the study of history may involve interactive assignments to promote curiosity and literacy, while problematizing the methods used to create such technologies and

50

their impact on our thought.

7. Generalizability

Book indices are a long-time subject of extensive research as part of the broader domain of information retrieval. Our contribution to this vein of research is both specific and generalizable. Specifically, we constructed a novel database based on book indices which aids in intellectually describing the content of ancient texts. This paper explains through description and usage scenarios how this improved access to ancient texts in Tiresias serves scholars and students of ancient history, helping them to save considerable research time, to raise new questions and to achieve a broader view of the field.

The generalizable aspect of our work is the notion to automatically construct new knowledge based on joining data from two independent indices pointing to the same book page, adding a contextual level to an otherwise technical reference. One may envision the applicability of the current approach to additional fields of research. For example, art books containing a picture index and subject index are good candidates for a similar approach, promoting a solution for another longstanding issue, the challenge of assigning meaning to visual materials. The 2016 scandal surrounding Facebook's removal of the Pulitzer prize winning, iconic photograph known as "napalm girl" is a vivid reminder for the critical importance of context and the poor ability of algorithms to contextualize, especially when visual materials are concerned. Another example could be literary, scientific or engineering books that contain a subject index and an author index. By crossing subject and author indices, one might draw inferences in two directions: 1) authors' areas of knowledge; 2) the prolific authors writing about a certain subject.

A practical contribution of this research concerns publishers who may find value in applying Tiresias logic to their collections of books as an added-value service to readers. Publishers could leverage their existing data to combine additional sources such as academic journals to enrich and further validate the data.

8. Summary

In summary, Tiresias offers access to ancient primary texts through the lens of humanistic scholarship described in the subject indices of academic books. It is a valuable tool for researchers wishing to gain fast and broad retrieval of sources combined with modern interpretations of their meaning. The database provides opportunities for filtering and visualization which aid the research process. In future research, we aim to further augment the validity of the source-subject pairs as well as investigate aspects of historical knowledge extraction. We encourage our readers to visit the Tiresias web site (<https://tiresias.haifa.ac.il/>) and send us feedback. Finally, the method of cross-indexing may apply to other research fields and offers opportunities for developing databases on other fields of science.

Acknowledgments

We thank Mr. Jonathan Blam for his work on validation assessment (paragraph 21). This research was supported by the University of Haifa Data Science Research Center

Works Cited

- Bastian et al. 2009** Bastian, M., Heymann, S., and Jacomy, M. (2009) "Gephi: An Open Source Software for Exploring and Manipulating Networks", *Third International AAAI Conference on Weblogs and Social Media*, pp. 361–362.
- Blidstein and Zhitomirsky-Geffet 2022** Blidstein, M. and Zhitomirsky-Geffet, M. (2022) "Towards a New Generic Framework for Citation Network Generation and Analysis in the Humanities", *Scientometrics*, 127(7), pp. 4275–4297.
- Colavizza and Romanello 2019** Colavizza, G., and Romanello, M. (2019) "Citation Mining of Humanities Journals: The Progress to Date and the Challenges Ahead", *Journal of European Periodical Studies*, 4(1).
- Colavizza et al. 2017** Colavizza, G., Romanello, M., and Kaplan, F. (2017) "The References of References: A Method to enrich Humanities Library Catalogs with Citation Data", *International Journal on Digital Libraries*.
- Croft 2002** Croft W. B. (2002) "Combining Approaches to Information Retrieval", in Croft, W. B. (ed.) *Advances in Information Retrieval. The Information Retrieval Series 7*. Boston, MA: Springer.

- Dalton and Charnigo 2004** Dalton, M., and Charnigo, L. (2004) "Historians and Their Information Sources", *College & Research Libraries*, 65(5), 400–425.
- Fyfe 2018** Fyfe, P. (2018) "Reading, Making, and Metacognition: Teaching Digital Humanities for Transfer", *Digital Humanities Quarterly*, 12(2).
- Green 2000** Green, Rebecca. (2000) "Locating Sources in Humanities Scholarship: The Efficacy of Following Bibliographic References", *The Library Quarterly*, 70(2), pp. 201–29.
- Gross et al. 2015** Gross, Tina, Taylor, Arlene G., and Joudrey, Daniel N. (2015) "Still a Lot to Lose: The Role of Controlled Vocabulary in Keyword Searching", *Cataloging & Classification Quarterly*, 53(1), pp. 1–39.
- Hitchcock 2013** Hitchcock, T. (2013) "Confronting the Digital", *Cultural and Social History*, 10, pp. 9–23.
- Huggett and Rasmussen 2013** Huggett, Michael, and Rasmussen, Edie. (2013) "User Interface Evaluation of Meta-Indexes for Search" in *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '13)*, pp. 435–436.
- Huistra and Mellink 2016** Huistra, H., and Mellink, B. (2016) "Phrasing history: Selecting sources in digital repositories", *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 49, pp. 220–229.
- Landry et al. 2011** Landry, Patrice, Bultrini, Leda, O'Neill, Edward T., and Roe, Sandra K. (2011) *Subject Access: Preparing for the Future*. Walter de Gruyter.
- Locke 2017** Locke, Brandon T. (2017) "Digital Humanities Pedagogy as Essential Liberal Education: A Framework for Curriculum Development", *Digital Humanities Quarterly*, 11(3): pp. 116–123.
- Merčun 2016** Merčun, Tanja, Žumer, Maja, and Aalberg, Trond. (2016) "Presenting Bibliographic Families: Designing an FRBR-Based Prototype Using Information Visualization", *Journal of Documentation*, 72(3): pp. 490–526.
- Nilbe and Tarkpea 2014** Nilbe, Sirje, and Tarkpea, Tiit. (2014) "Using the Estonian Subject Thesaurus in the Digital Environment", *Cataloging & Classification Quarterly*, 52(1), pp. 32–41.
- Piotrowski 2010** Piotrowski, Michael. (2010) "Leveraging Back-of-the-Book Indices to Enable Spatial Browsing of a Historical Document Collection". In *Proceedings of the 6th Workshop on Geographic Information Retrieval (GIR '10)*, 17:1–17:2.
- Romanello 2018** Romanello, Matteo. (2018) "Large-Scale Extraction of Canonical References: Challenges and Prospects", January.
- Romanello et al. 2009** Romanello, Matteo, Berti, Monica, Babeu, Alison, and Crane, Gregory. (2009) "When Printed Hypertexts Go Digital: Information Extraction from the Parsing of Indices" in *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia (HT '09)*, pp. 357–358.
- Scheirer et al. 2016** Scheirer, W., Forstall, C., and Coffee, N. (2016) "The Sense of a Connection: Automatic Tracing of Intertextuality by Meaning", *Literary and Linguistic Computing*, 31(1), pp. 204–17.
- Sinn and Soares 2014** Sinn, D., and Soares, N. (2014) "Historians' use of digital archival collections: The web, historical scholarship, and archival research", *Journal of the Association for Information Science and Technology*, 65, pp. 1794–1809.
- Waskom 2021** Waskom, M. L. (2021) "Seaborn: Statistical Data Visualization", *Journal of Open Source Software*, 6(60), 3021. Available at <https://doi.org/10.21105/joss.03021>.
- Włodarczyk 2013** Włodarczyk, Bartłomiej. (2013) "Topic Map as a Method for the Development of Subject Headings Vocabulary: An Introduction to the Project of the National Library of Poland", *Cataloging & Classification Quarterly*, 51(7), pp. 816–29.
- Yu and Young 2017** Yu, Holly, and Young, Margo. (2017) "The Impact of Web Search Engines on Subject Searching in OPAC". *Information Technology and Libraries*, 23(4), pp. 168–80.



This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.