DEVELOPING VERITRACE'S COMPUTATIONAL APPROACH TO TRACING THE INFLUENCE OF ANCIENT WISDOM IN EARLY MODERN NATURAL PHILOSOPHY

Jeffrey C. WOLF*

Abstract. This article introduces the computational approach of the VERITRACE project, which aims to trace the influence of ancient wisdom writings on early modern natural philosophy using advanced digital tools. It outlines the project's scope, data sources, and methodology, combining distant and close reading of a large multilingual, diachronic corpus. The article describes the initial phases of data acquisition, highlighting challenges encountered and solutions developed. Key digital capabilities being implemented are discussed, including keyword search, text matching, sentiment analysis, and latent thematic analyses like Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). Early results demonstrating the potential of these methods are shared through a case study examining the influence of the first English translation of *The Divine Pymander* in 1650. The article concludes with lessons learned so far and suggestions for shared digital humanities infrastructure to support future large-scale textual analysis projects.

Keywords: digital humanities, distant reading, text matching, keyword search, early modern philosophy, ancient wisdom

Introduction

Between 2023 and 2028, the European Research Council is funding the project *Traces de la Verité: The reappropriation of ancient wisdom in early modern natural philosophy*, aka VERITRACE (ERC-StG Project VERITRACE, 101076836).[1] Led by Professor Cornelis J. Schilt at the Vrije Universiteit Brussel (VUB), VERITRACE's primary goal, as has been described in more detail elsewhere in this special issue, is to trace the influence of the most prominent ancient wisdom writings throughout the early modern period, paying particular attention to how these writings functioned in natural philosophical discourse.[2]

VERITRACE accomplishes this by applying sophisticated digital analysis techniques on a large corpus of early modern texts, tracing referenced and unreferenced uses of the *Corpus Hermeticum* (including the *Asclepius*), the *Chaldean* and *Sibylline Oracles*, and the *Orphic Hymns*. Moreover, it analyses how these texts were

* Vrije Universiteit Brussel, e-mail: jeffrey.charles.wolf@vub.be

being used (employing Latent Semantic Analysis, among other tools), and with what sentiment they were discussed (using Sentiment Analysis) by their proponents and antagonists, and how these debates were influenced by key episodes in the transmission history of these texts. VERITRACE aims to provide a comprehensive analysis of the influence of ancient wisdom writings on early modern natural philosophy, and it does so by making use of methodologies hitherto not employed at this scale in the early modern history of science.

VERITRACE draws on the most ubiquitous intellectual materials found in this period: printed books. Although early modern debates followed other modes of discourse, in particular oral discussion and the circulation of manuscripts and letters, these often pertained to small circles and select readers. Books, on the other hand, were everywhere, connecting authors and readers all over Europe and beyond. Indeed, even if we only focus on works in Latin, English, German, French, Dutch, and Italian, the number of books that have come down from the early modern period is staggering, presenting lots of challenges–and opportunities.

Some of these challenges uniquely characterise VERITRACE as a digital humanities project. These features include:

1. **Multilingualism**: Our source material comes from at least six different languages, both modern and classical ones. Many digital humanities projects only contend with one or two languages, especially English. Since many natural language processing (NLP) techniques were initially developed on easily available textual corpora in contemporary English, the multilinguistic nature of VERITRACE raises its own set of challenges. In fact, even when available tools are comprehensive in terms of modern languages, they sometimes exclude, or have limited support for, classical languages like Latin.[3]
2. **Longue durée**: VERITRACE spans a period of almost two hundred years (1540-1728), so this adds diachronic complexity to our investigation of the corpus. An interpretation that applies to a smaller slice of the data cannot be assumed to apply to the whole, given change in historical context and linguistic meaning over time.
3. **Big Data**: With hundreds of thousands of texts as our source material, simple search processes and data management will not be adequate for the project. It will be resource intensive and sometimes require different tools and solutions, given the size of the data collections with which we are working.
4. **Complex Integration**: Because our data comes from different sources held in separate institutions collected over long periods of time, there are inherent challenges to integrating and harmonising the data. We pay particular attention to–and carefully document–any cleaning and transformations we apply to the data, so that we have a reasonable basis for subsequent analysis.

VERITRACE must also grapple with the familiar challenges of any distant reading project: issues with the accuracy of the underlying digital texts (OCR quality), the parameter-dependent nature of various NLP techniques, and so forth.

Distant Reading

To meaningfully trace the influence of the various ancient wisdom writings and their Renaissance popularisers on a case-by-case basis, we would need to work with a very large team of researchers, or drastically reduce the number of passages traced and books inspected, and presumably both. But this is where digital techniques come in, most notably from the perspective of distant reading, which have been developed specifically to query large corpora. These techniques, adopting methods from natural language processing, allow for the analysis of large text corpora, identifying patterns and uncovering both prominent and neglected works, the latter termed 'the great unread' by Margaret Cohen.[4] Early modern writers would rarely include references to their source material, which provides one of the key challenges for the project.

Significant advancements in digitising early modern books have expanded the application of distant reading techniques. Improved OCR technology now yields meaningful results even with suboptimal text recognition.[5] Online repositories, like those of the Bibliothèque nationale de France, provide standardised data for content extraction, facilitating large-scale analysis.[6]

The Distant Reading Corpus (DRC) we have chosen consists of several hundred thousand works from important European library collections, written in Latin, French, German, Dutch, English, and Italian, including:

- *Early English Books Online* (EEBO) (ProQuest), which in its EEBO-TCP format developed by the Text Creation Partnership[7] contains about 58,000 English and Latin texts published between 1540 and 1700 (hereafter **'EEBO'** unless we refer specifically to our custom version of EEBO, which we call 'VEEBO', described in more detail below)
- *Gallica* (Bibliothèque nationale de France) contains almost 125,000 books published between 1540 and 1728 in a variety of languages including French, Italian, Dutch, and Latin (hereafter **'Gallica'**)
- The *Digitale Sammlungen* of the Bavarian State Library, which contain more than 340,000 books published between 1540 and 1728, including in Latin, German, French, Greek, Italian, and Dutch, among others (hereafter **'BSB'**)

The rationale behind choosing these primary data sources is that we want to be able to make historical claims about the *Prisca sapientia* tradition, e.g. how prevalent it was and the level of interest in it over time. Did curiosity in the *Corpus Hermeticum*, for instance, decline after the first quarter of the seventeenth-century, or not? By interrogating a truly representative sample of books, we can make reasonable claims about levels of interest and prevalence. This is not the place to dive into the nitty-gritty of sample size and representativeness, but a rigorously statistical frame of mind underpins our approach, and we believe the sources we have chosen can be the basis for constructing a representative sample size of books published in Europe between 1540 and 1728.

Close Reading Too

VERITRACE combines proven techniques for distant reading on a much larger corpus with the close reading of a carefully selected corpus of Renaissance and early modern texts.[8]

The Close Reading Corpus (CRC) consists of all the relevant editions of the *Corpus Hermeticum* (including the *Asclepius*), the *Chaldean Oracles*, the *Sybilline Oracles* and the *Orphic Hymns* that were published during the Renaissance and early modern period, and the works that drew heavily on these ancient wisdom writings and promoted the idea of a *prisca sapientia*, such as Ficino's *Theologia platonica* and Steuco's *De perenni philosophia*, for a starting total of circa 80 works. Obviously, when it comes to the readership and geographical dissemination of these works, a census would be extremely welcome. Previous censuses of works such as Copernicus's *De Revolutionibus*, Vesalius's *De Fabrica*, and the first edition of Newton's *Principia* all revealed a wealth of information about how and by whom these works were read, how they changed ownership over time, and indeed seriously challenged myths surrounding their print run and popularity, as with the *Principia*.[9] As desirable as such a census may be, it would not be feasible within the scope of this project. However, any relevant annotated or otherwise interesting versions of works within the CRC the project team comes across will be included as case-studies.

The CRC will be dynamic in two ways. When it comes to influential editions, the research within this project will undoubtedly uncover seemingly innocuous works that turn out to have been of great influence. These works will subsequently be included in the CRC. Secondly, as the CRC will be mapped against the much larger Distant Reading Corpus that contains materials from 1540 to 1728, the actual contents of the CRC will vary depending on the dates of the works it is compared with.[10]

What this overall methodology allows for is a combination of highly granular and detailed close reading of a select corpus with the possibility to zoom out and inspect that corpus embedded in a much larger environment. In historical scholarship thus far, the influence of ancient wisdom texts on natural philosophy has been confined to a handful of case studies that involve the big names, primarily Copernicus, Kepler, Bacon, More, and Newton. But early modern natural philosophy was always so much more than the pursuit of a select few individuals, and even though many never put their ideas, theories and experiments to paper, thousands of others did. Likewise, the ancient wisdom writings were read and studied by many, with new editions of these texts in Greek, Latin, and the vernacular appearing regularly throughout the early modern period.

Critical Reflections

VERITRACE has now been active for over a year, so it is a good time to reflect on what we have done, what we still wish to do, and to offer some reflections about what we have learned so far.

It is also worth underscoring two general points: first, digital tools in this new age of machine learning and A.I., are changing so rapidly that anything written in the pages of this print journal will already be out-of-date by the time of publication. Thus,

the emphasis in what follows is less on the specific tools and more on what we hope to accomplish using them. And, since the VERITRACE project is still early in development, it makes more sense to discuss a preliminary case study or proof of concept, and our overall approach, rather than a detailed discussion and rationale for the set of tools we end up using for the final research platform. It is best seen as a snapshot in time of work in progress. Indeed, the accompanying figures we provide (see below) are not to be seen as final or even preliminary results but as views of our 'working data' as they were at that time, crude as they appear. Nonetheless, it is hoped that, despite this project-centred approach and the early stage of the work, our critical reflection about the challenges faced will be of interest to a wider audience.[11]

Second, it can never be emphasised enough that digital tools and techniques in the humanities are best seen as complementary to the skills of traditional, critical scholarship. At their best, these new techniques allow us to discover little-noticed data or uncover new ways of seeing our sources. They provide intellectual threads, which can then be unwound using the long-established tools of the humanities scholar–critical, nuanced thinking, serious attention paid to individual texts and the context in which they were written, and an awareness of the insights provided by multiple fields of inquiry. These tools always belong in the service of traditional scholarship.

I. VERITRACE Data

There is no VERITRACE project, or any digital humanities project, without the proper data. Thus, a large part of the initial months of the VERITRACE project were spent on a seemingly mundane but critical task: the acquisition of data. This is the traditional first phase of any data-heavy project, and without it, nothing further is possible. Therefore, what I will do now is describe data acquisition for the project so far, with all its twists and turns, as well as some of the challenges we have faced, and how we have decided to address them. It is probably typical of many digital humanities projects that have had to adjust as they encounter expected–and unexpected–issues.

To start, one abstract way of thinking about a data-heavy, digital humanities project is to view it as a data processing pipeline (see **Figure 1**). A simple model of this pipeline can be separated into 5 sequential stages: data acquisition, data cleaning, data transformation, data analysis, and finally, data storage.[12] In other words, one must acquire the data first and then clean and transform it, before being able to analyse it meaningfully and store it appropriately. The distinction between stages is not always so clear-cut, nor does a digital humanities project simply move from one stage to the next. Nonetheless, a data processing pipeline is a helpful way of envisioning the overall process, and VERITRACE began with data acquisition.

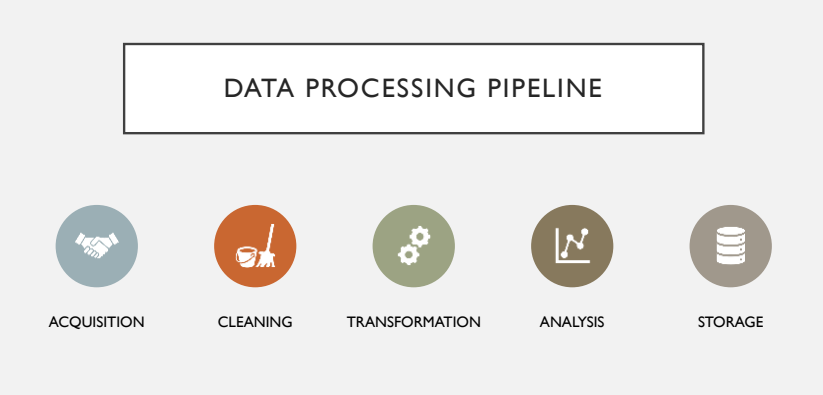


Figure 1. Schematic of a data processing pipeline

For each data source, there are two types of data about the original printed book we will need to obtain and later analyse: first, the metadata about the printed texts (e.g. author, title, publication date, and so forth) and, second, the digital versions, transcribed or produced through optical character recognition (OCR), of those same texts, often in the form of an XML, HTML, or HOCR file.

I refer to the data sources of the VERITRACE project collectively as the **VERITRACE data lake**. A data lake is "a centralized repository that ingests and stores large volumes of data in its original form…a data lake can accommodate all types of data from any source, from structured (database tables, Excel sheets) to semi-structured (XML files, webpages) to unstructured (images, audio files, tweets), all without sacrificing fidelity."[13]

The VERITRACE data lake contains the semi-structured digital text files from our three data sources along with the structured metadata records for each of the original printed texts.

Data Acquisition

Data acquisition began in earnest in October 2023. It sounds straightforward, and sometimes data acquisition is simply a matter of downloading a few files. But often, as in our case, there are numerous complications.

We started under the assumption that we needed to create a consolidated search tool that combined application programming interface (API) & local search results in a transparent and reliable way. There is nothing technically difficult about this; digital humanists have long been familiar with using APIs to connect to external, online data sources, and local search is something we all do when we search our own hard drives. In fact, this kind of basic search functionality should already be available using the various web interfaces for Gallica, the BSB, and even EEBO. Could we not simply use the web interface tools already provided online to power our research? To

some extent, yes. For instance, the Gallica search interface permits the export of search query results (as a CSV file)[14] but the BSB search interface, at time of writing, does not.[15]

So, our first steps appeared obvious: we would start the project by connecting to the Gallica Search API(s), the BSB Search API(s), and create our own search tool for the EEBO texts. Using these, we could then pass a single set of search parameters to all 3 data sources and combine the results in a reliable and meaningful way. Furthermore, it would keep our data storage requirements light because we would not need to download the metadata or digital texts from Gallica or the BSB; instead, we would simply be interacting with their data, on a 'need to know' basis, e.g. when we sent a search query and needed to obtain the results.

In terms of collecting data, each data source presented its own set of challenges. Gallica presented the conventional issue of connecting to an online data source via an external API. Although the API documentation is not always straightforward and hides some of its complexity, this was the first source to which we attempted to connect and were appreciative that the search interface offered the ability to download search results, as we could immediately get an overview of the full scope of the data we wanted to work with. While every API is different, and it takes a certain of amount of 'learning the API language', connecting to an external API is a familiar task and no further discussion is necessary here. Suffice it to say that we wrote a script in Python to parse the XML response from the API endpoints that we needed to access and used the *Pandas* python library to manage the results.[16]

The Bavarian State Library (BSB) proposed a different challenge: although the online interface is very user-friendly, it was hard to get an overview of the entire database data, without an ability to download search results. More worrisome for us, no search API for the BSB digital collections currently exists. What to do?

In consultation with one of their developers and the BSB's Munich Digitisation Centre itself, we were given permission to piggyback on their frontend web application to access their backend database, without an official API. As they informed us, they had an internal HTTP API which is used for the interaction between their search backend and frontend web application. And we could use it with under certain conditions.[17] They informed us that they cannot offer any formal support, that we should not make excessive requests at any one time, and that they could not guarantee its stability. But, in theory, we would have 'informal' access to an internal search API and could obtain the results we needed. With an ideal solution out of reach, we were happy to be practical. This, then, is how we connected to our second (online) data source, by writing a Python script that used an informal and internal API 'surface' to access the search results we needed.

We assumed that our third data source, EEBO, would be the most straightforward. We saved it for last because we knew no API connection would be necessary; the files can be downloaded directly and constitute a static data source.[18] The EEBO files contain, in theory, a digital copy–and a TEI-encoded one at that–of at least one edition of every book published in the English language up to 1700.[19] So the quality of the digital texts would be high, and there would be no API 'learning

curve' to access them. We would therefore simply download these files onto a local computer and run our own full-text searches against them.

What also helped was that we had come across other digital humanities projects that used EEBO as their primary data source. The project leaders of *EarlyPrint*, for instance, had set up a website with a search interface and Python tutorials and have gone to great efforts over several years to improve both the metadata and encoding quality of much of the EEBO collection.[20] They were equally generous with their time, at a very early stage in our own project, explaining to us how they approached the data and how we could programmatically access what they had done. We later discovered other ambitious projects, like *Project Quintessence*, that showcased alternative ways of interacting with the EEBO collection.[21]

Nonetheless, our confidence about the EEBO data source turned out to be naïve: one could download the data onto a local machine in a straightforward manner, but at more than 60,000 files, some as large as 100MB, the processing power and time it would take to manage them would be more challenging than using a public search API.

Also, we wanted to improve upon the EEBO data source by incorporating as many of the *EarlyPrint* transcriptions (that is, their digital text versions of the EEBO texts) as we could, in the belief that they were the most accurate versions available. But *EarlyPrint* intentionally did not transcribe many non-English-language works that are part of EEBO–works VERITRACE is still interested in–so we decided to incorporate additional EEBO texts that were not part of *EarlyPrint*, creating a unique collection of texts. Our custom collection of texts also has its own set of metadata records that match each text, and we refer to our specific version of EEBO as **VEEBO**.

By January 2024, we were able to create a prototype of our consolidated search tool, allowing the user–albeit inefficiently and slowly–to send a single search query to all 3 data sources and to receive a single set of search results:

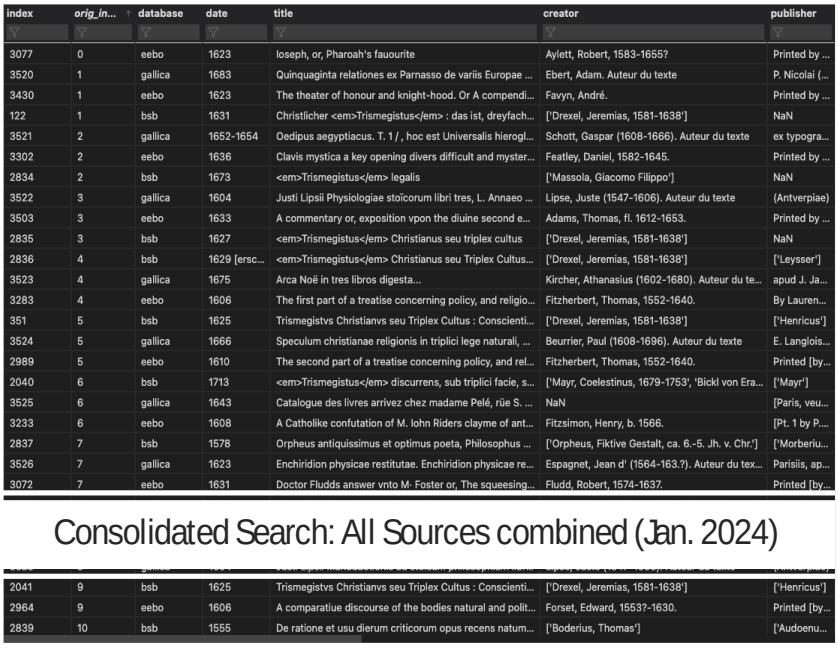


Figure 2. Results from the initial consolidated search tool (January 2024, for internal use)

Creating such a tool was illuminating in more ways than one, including by showing us its own inadequacy. We quickly realised that a not insignificant portion of metadata records was likely unreliable or, at best, not provided in a form conducive to subsequent analysis.

To take a simple but critical example: although each data source contains a 'Date' field that generally records the date of publication for the printed book represented by that record, dates are provided in a variety of incompatible and non-standard formats, e.g. in the form of Roman numerals instead of Arabic, or in a date range rather than a single 4-digit year. While a date range may sometimes be justified–for multivolume works, for instance, which present their own challenges–we need a clear, transparent set of rules that guide our data use such that the date values are in a consistent, standard format, reconcilable with the values from each of our data sources. Without this, our subsequent analyses will not be accurate. In fact, a closer look at our data showed that, for just one of our data sources (VEEBO), there are 527 unique value formats ('date signatures') in the Date field, instead of the single four-digit date format that we generally need.

This is not surprising, and we always expected that data cleaning would be necessary, indeed as the next step in the data processing pipeline. But to do that, we could not simply use the data provided to us via the search APIs for our library catalogue data. Instead, we needed more control over the underlying data, so we could

clean and transform it, and this meant that, we would need to download and manage the metadata and digital texts for *all* our sources, not just VEEBO.

Downloading the Data

Our API experience was still valuable because connecting to the Gallica API, for instance, would allow us to both download metadata records as well as any digital texts associated with those records. It would take some time, but we set up nightly 'data feeds' so that–over the course of a matter of weeks–using modified API scripts we had previously created, we were able to download all the text files and associated metadata for our Gallica data source (see **Figure 3**). For the BSB, we were able to download the metadata for the texts that constituted the printed works for our time period.

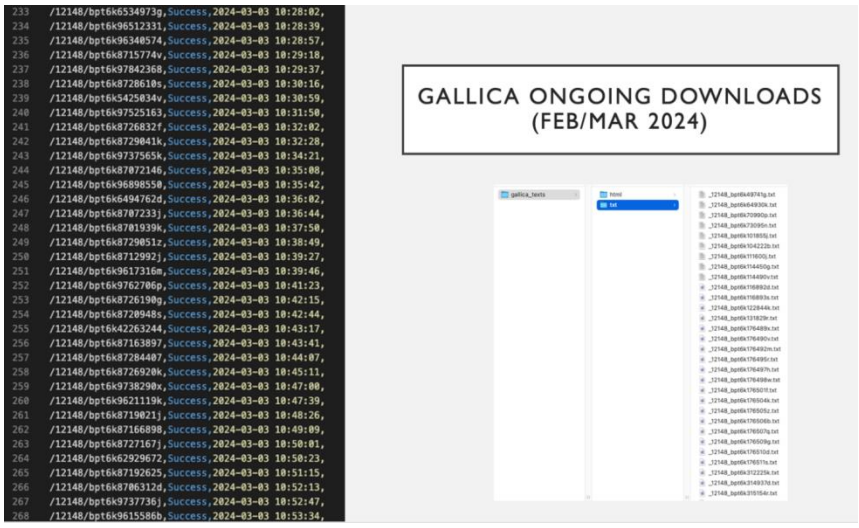


Figure 3. Nightly 'data feed' to download digital texts from Gallica (Bibliothèque nationale de France)

Once this was done, the final step in data acquisition was to find a way to obtain the digital texts from our largest data source, the BSB. However, as they explained to us, "there is no way to directly download the full text of a document in one go [from the BSB website]. The only way to access it is via the OCR API…"[22] meaning that we would have to go page image-by-page image through each document, downloading any associated text. While this could be done automatically, it would take a long time for the 340,000+ texts we were interested in – about 15 months, in our estimation.

Finding a way to avoid this would save us a lot of time. After some discussion with the Bavarian State Library, who have been incredibly helpful at all stages of our project, we came to an agreement with them that, in exchange for their providing us

with a hard drive containing the 340,000+ digital texts, we would only use the data for the purpose of our project, without redistributing it to other parties, and would share our results with them, as well as providing feedback about our use of their tools.[23] Once we obtained the hard drive from the BSB–which arrived in late May 2024–we could declare Phase 1–data acquisition–complete.

It is worth highlighting a data quality issue that we faced right away but which we are setting aside for later in the project: inaccurate digital texts due to poor OCR quality. We are not simply searching the metadata from our 3 data sources but also the digital texts themselves. Where does the digital text come from? In most cases, from the results of automated, or semi-automated, OCR performed on the individual book images that the holding library conducted at some earlier stage. Ultimately, then, the quality of our searches depends on the underlying OCR quality of the digital text.

This is not yet cause for despair, however: perfect OCR capture is neither realistic nor required. Importantly, statistically significant, and meaningful results are possible even with suboptimal OCR, as previously mentioned.[24] We decided that, though the OCR quality challenge is an important one, it was better left to a later phase in the project.

II. VERITRACE Case Study

What do we wish to do with all this data? It is true that we need to clean and, in many cases, transform the data for subsequent use. These are critical steps, and we are doing both. But that is a discussion for another place. Instead, I would like to take a step back to think critically about what we want to accomplish in the project. As outlined in the introduction to this article, scholars should be able to use the tools of the VERITRACE project to trace the influence of ancient wisdom writings on the development of early modern natural philosophy. This could mean searching for essential terms from *prisca* writings in a larger group of natural philosophical writings and authors. To do this, we need to provide scholars a way of searching efficiently through the distant reading corpus (keyword search). This is a traditional approach–user-driven search–but can be quite effective, especially with a very large corpus. Beyond user-driven search, a scholar might also wish to compare entire texts, passage by passage, between the *prisca* tradition and the early modern natural philosophical tradition. Which early modern philosophers were using *prisca* language, often unacknowledged, directly from the ancient wisdom corpora themselves? Can we trace these instances of textual re-use? To do this, we need to build a lexical matching tool–a kind of early modern plagiarism detector–that can find and highlight the re-use of essential vocabulary and lexical passages from one set of texts to another. But that is not all: lexical matching only works when there is shared vocabulary, when a source text is in the same language as a target text. But our corpus is multilingual, so we also need semantic analytic tools (e.g. LSA, LDA–see below) that work, no matter the language used. They need to operate at the more abstract level of semantic meaning – not lexical similarity.

To make this all more concrete, I will now present a proof of concept in the form of a case study, which will also illustrate the potential of some of these capabilities:

Research Question: *what was the influence (however vaguely defined) of the first English translation of the* Divine Pymander *(1650) upon the subsequent generation of thinkers, who published texts in English between 1650 and 1680?*

We will approach this in terms of the influence of a specific *source text* upon a much larger *target corpus*. The source text is the first translation into English of a work from the *Corpus Hermeticum*; namely, John Everard's *The divine Pymander of Hermes Mercurius Trismegistus,* published in 1650 (see **Figure 4**), and the target corpus (a subset of our larger Distant Reading Corpus) consists of all the English-language texts contained in VEEBO published between 1650 and 1680: 18,633 individual texts in total.[25]

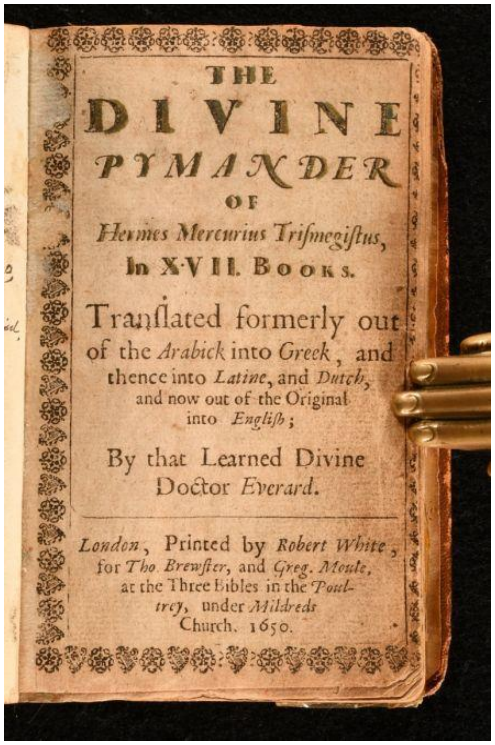

THE
DIVINE
PYMANDER
OF
Hermes Mercurius Triſmegiſtus,
In XVII. Books.
Tranſlated formerly out of the *Arabick* into *Greek*, and thence into *Latine*, and *Dutch*, and now out of the Original into *Engliſh*;
By that Learned Divine Doctor *Everard*.
London, Printed by *Robert White*, for *Tho. Brewſter*, and *Greg. Moule*, at the Three Bibles in the *Poultrey*, under *Mildreds* Church. 1650.

Figure 4. The title page from the 1650 English edition of *The Divine Pymander of Hermes Mercurius Trismegistus*

Keyword Search

A natural place to begin is keyword search, which is a familiar, long-established approach, with no less intellectual power for that. The name 'pymander' is an unusual but important one for our source text–in the very title itself–so let us begin by conducting a keyword search for 'pymander' in the target corpus. Perhaps later writers are referencing it directly? Here are the top results from this basic keyword search:

There are a few important observations worth making (see **Figure 5**). First, the 'score' column contains score values, which should be interpreted as equivalent to 'relevance scores' (relative to each other). The values provide a ranking of the most relevant search results, given the search query.

Second, both the title and the content of each digital text are searched by default, which is why–in the first result under 'matched terms'–'pymander' is listed twice, once for appearing in the title and once for appearing in the text itself.

Keyword Search: 'pymander'

filename	author	title	date	snippet	score	matched_terms	keyword_hits	doc_length
A43420.xml	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	, his Divine PYMANDER, in seventeen books: Hermes Mercurius Trismegistus, HIS Divine Pymander, IN Seventeen Books. Together with his Second Book, Called Asclepius; Containing fifteen Chapters, with a Commentary. Transl●..ted formerly out of the Arabick into Greek, and thence into Latine, and Dutch, and now out of ...	22.900000	pymander, pymander	3	65352
A31383.xml	Caussin, Nicolas,	The holy court in five tomes: the first, treating of motives, which should excite men of qualitie to Christian perfection. The second, of the prelate,	1650	The holy court in five tomes: the first, treating of motives, which should excite men of qualitie to Christian perfection, The second, of the prelate,: THE HOLY COURT IN FIVE TOMES: The FIRST, Treating of MOTIVES, Which should excite men of Qualitie to CHRISTIAN PERFECTION. The SECOND, Of the PRELATE, SOULDIER, STATES-MAN, and LADIE. The THIRD, Of MAXIMS of CHRISTIANTIE AGAINST PROPHANESSE, Divided into Three Parts, vi...	12.800000	pymander	12	106374
A42844.xml	Gadbury, John,	The just and pious scorpionist: or The nativity of that thrice excellent man Sir Matthew Hales, late Lord Chief Justice of England. Who was born in th	1677	The just and pious scorpionist: or The nativity of that thrice excellent man Sir Matthew Hales, late Lord Chief Justice of England. Who was born in th: The Just and Pious SCORPIONIST: OR THE NATIVITY Of that thrice Excellent Man Sir Matthew Hales, Late Lord Chief Justice of England. Who was born in the Year of our Lord 1609. on Wednesday, Novemb. the first, 7h 8' manè, Under the Coelestial Scorpion: Astrologically...	12.600000	pymander	1	4842
A43285.xml	Helmont, Jean Baptiste van, (1577-1644.)	Van Helmont's works: containing his most excellent philosophy, physick, chirurgery, anatomy. Wherein the philosophy of the schools is examin	1664	Van Helmont's works: containing his most excellent philosophy, physick, chirurgery, anatomy. Wherein the philosophy of the schools is examin: Van Helmont's WORKS: Containing his most Excellent PHILOSOPHY, CHIRVRGERY, PHYSICK, ANATOMY. WHEREIN The Philosophy of the Schools is Examined, their Errors Refuted, and the whole Body of Physick REFORMED and RECTIFIED. Being a New rise and progresse of PHILOSOPHY and ...	4.600000	pymander	2	106374
A39847.xml	Fludd, Robert,	Mosaicall philosophy: grounded upon the essentiall truth or eternal sapience. Written first in Latin, and afterwards thus rendred into English. By Rob	1659	Mosaicall philosophy: grounded upon the essentiall truth or eternal sapience. Written first in Latin, and afterwards thus rendred into English. By Rob: MOSAICALL PHILOSOPHY: Grounded upon the ESSENTIALL TRUTH OR ETERNAL SAPIENCE. Written first in Latin, and afterwards thus rendred into English. By ROBERT FLUDD, Esq & Doctor of Physick. The Lord giveth Wisdom, and out of his Mouth commeth Knowledg and Understanding, ...	4.600000	pymander	2	106374
A89818.xml	Naudé, Gabriel,	The history of magick: by way of apology, for all the wise men who have unjustly been reputed magicians, from the Creation, to the present age. / Wri	1657	The history of magick: by way of apology, for all the wise men who have unjustly been reputed magicians, from the Creation, to the present age. / Wri: THE HISTORY OF MAGICK By way of APOLOGY, For all the Wise Men who have unjustly been reputed Magicians, from the Creation, to the present Age. Written in French, by G. NAUDAEUS Late Library-Keeper to Cardinal Mazarin. Multos absolvemus, si caeperimus antè judicare ...	4.000000	pymander	1	61242
A37412.xml	Dee, John, (1527-1608.)	A true & faithful relation of what passed for many yeers between Dr. John Dee (a mathematician of great fame in Q. Eliz. and King James thei	1659	A true & faithful relation of what passed for many yeers between Dr. John Dee (a mathematician of great fame in Q. Eliz. and King James thei: The Order of the Inspirati MAHOMET receives his Law by Inspiration. APPOLONI. ª TYANEUS in Domitians tyme Edw Kelly Prophet or Seer to Dr. Dee. Roger Bacon an English man PARACELSUS Receits from the Inspiration of Spirits. Dr. Dee avoucheth his Stone is brought ...	2.600000	pymander	1	106374

Figure 5. A keyword search for 'pymander' within the target corpus

Third, it turns out that within our target corpus there is another edition of the *Divine Pymander*, published 7 years later (1657). It is almost identical to our source text, which is why it ranks first with the highest 'relevance score', when we search for 'pymander'. This is, in a way, a nice validation that our keyword search is accurate.

And it also means we must be a bit cautious in our interpretation of the influence of the 1650 edition; the 1657 edition is very similar to the earlier 1650 edition, so we should allow for the possibility that later writers are quoting from, or using, the 1657 edition instead of the source text. This is where the work of traditional scholarship would provide the final work (we will not adjudicate this now).

Fourth, our keyword search results do not simply provide a rank listing of documents with the highest number of 'keyword hits'. If it did, then the second result, Nicolas Caussin's *The Holy Court in Five Tomes*, would rank first, as 'pymander' shows up 12 times, whereas it shows up only 3 times in our most relevant result. This is in truth what we want to see because the search algorithm boosts the relevance score of any keyword that appears in the title of the target text, not just the text proper. This intuitively matches our sense of the importance of a keyword. Finally, the search algorithm also accounts for the length of the text in which the keyword is found, in relation to the number of 'hits' found in that document. Thus, a longer document with a greater number of hits might still rank below a shorter document with a smaller number of hits, under the assumption that the keywords are more central to the shorter text.[26]

Wildcard Search: 'trismeg*' (sample)

A50289.xml	Matthew, Edward	Karólou trismegístou 'epiphanía. The most glorious star, or, celestial constellation of the Pleiades, or Charles Waine, appearing, and shining most	1660	Karólou TRISMEGÍSTOU 'epiphanía: A Representation of his late Majesty King Charles the first His giving Thankes to Almighty God for the birth of his Sonne (now our most Gratious Soveraigne) King Charles the 2d. at St. Pauls London May the 29th: 163● ΚΑΠΟ'ΛΟΥ ⟨◊⟩ ⟨◊⟩ ⟨◊⟩ ⟨◊⟩ ⟨◊⟩ . THE MOS...	25.500000	trismegistou, trismegest
A81149.xml	Culpeper, Nicholas, 1616-1654.	Mr Culpeper's ghost, giving seasonable advice to the lovers of his writings. Before which is prefixed, Mris. Culpepers epistle in vindication of her h	1656	Mr Culpeper's ghost, giving seasonable advice to the lovers of his writings. Before which is prefixed, Mris. Culpepers epistle in vindication of her h: ...	24.800000	trismegistian-platonick-rosy-crucian, trismegistus
A31383.xml	Caussin, Nicolas,	The holy court in five tomes: the first, treating of motives, which should excite men of qualitie to Christian perfection. The second, of the prelate,	1650	The holy court in five tomes: the first, treating of motives, which should excite men of qualitie to Christian perfection. The second, of the prelate,: THE HOLY COURT IN FIVE TOMES: The FIRST, Treating of MOTIVES, Which should excite men of Qualitie to CHRISTIAN PERFECTION. The SECOND, Of the PRELATE, SOULDIER, STATES-MAN, and LADIE. The THIRD, Of MAXIMS of CHRISTIANITIE AGAINST PROPHANESSE, Divided into Three Parts, vi...	23.500000	trismegist, trismeg, trismegistus
A62913.xml	Tompson, Benjamin,	New-Englands tears for her present miseries: or, A late and true relation of the calamities of New-England since April last past. With an account of t	1676	New-Englands tears for her present miseries: or, A late and true relation of the calamities of New-England since April last past. With an account of t: New-Englands Tears FOR HER Present Miseries: OR, A Late and True RELATION of the CALAMITIES of NEW-ENGLAND Since APRIL last past. With an Account of the Battel between the English and Indians upon Seaconk Plain: And of the Indians Burning and Destroying of Marl...	16.700000	trismegestos
		Sculptura: or The history, and art of				

Figure 6. A keyword wildcard search for 'trismeg' within the target corpus (the * is the wildcard search operator)

We can do more than a simple keyword search. A helpful variation on this is a *wildcard search*, especially given the inconsistent use of spelling in early modern English, not to mention imperfect transcriptions of digital texts derived from automated OCR. For instance, we can search the target corpus for 'trismeg*' (part of the proper name 'Hermes Trismegistus' central to the *Divine Pymander* and the *Corpus Hermeticum* more generally) which finds any keyword that contains, as a minimum, the character sequence 'trismeg' and any additional characters:

Our wildcard search (see **Figure 6** above) picks up a variety of terms that contain 'trismeg', including 'trismegest', 'trismegestos', or even 'trismegistian-

platonick-rosycrusian', which would not necessarily be found using a basic keyword search for 'trismegistus'.

We can also run fuzzy searches across our target corpus, which are similar to but distinct from wildcard searches. In a *fuzzy search*, instead of picking up all additional characters to our base keyword, or partial keyword, like one does in a wildcard search, we find simple edit changes. For instance, if we run our original keyword 'pymander' as a fuzzy search, we get the following results:

Fuzzy Search: 'pymander~'

filename	author	title	date	snippet	score	matched_terms
A43420.xml	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	, his Divine PYMANDER, in seventeen books: Hermes Mercurius Trismegistus, HIS Divine Pymander, IN Seventeen Books. Together with his Second Book, Called Asclepius; Containing fifteen Chapters, with a Commentary. Transl●...ted formerly out of the Arabick into Greek, and thence into Latine, and Dutch, and now out of ...	43.100000	pymander, pomander, pimander, pymander
A86610.xml	Howard, Robert, Sir	Poems, viz. 1. A panegyrick to the king. 2. Songs and sonnets. 3. The blind lady, a comedy. 4. The fourth book of Virgil, 5. Statius his Achilleis, wi	1660	Poems, viz. 1. A panegyrick to the king. 2. Songs and sonnets. 3. The blind lady, a comedy. 4. The fourth book of Virgil, 5. Statius his Achilleis, wi: POEMS, viz. 1. A PANEGYRICK to the KING. 2. SONGS and SONNETS. 3. The BLIND LADY, a COMEDY. 4. The Fourth Book of VIRGIL, 5. STATIUS his ACHILLEIS, with ANNOTATIONS. 6. A PANEGYRICK to GENERALL MONCK. By the Honorable Sr ROBERT HOWARD. LONDON, Printed for Henry Herringman...	24.600000	pylander, pysander
A53472.xml	Orrery, Roger Boyle, Earl of,	Parthenissa, that most fam'd romance· The six volumes compleat. Composed by the right honourable the Earl of Orrery.	1676	Parthenissa, that most fam'd romance· The six volumes compleat. Composed by the right honourable the Earl of Orrery.: PARTHENISSA, That most Fam'd ROMANCE. The Six Volumes Compleat. COMPOSED By the Right Honourable The Earl of Orrery. LONDON, Printed by T. N. for Henry Herringman, at the Blue Anchor in the Lower-Walk of the New Exchange, MDCLXX-VI To my LADY NORTHUMBERLAND. MADAM, WE...	22.900000	symander
A47940.xml	L'Estrange, Roger, Sir	A whipp for the animadverter in return to his second libell. By R. L'S.	1662	A whipp for the animadverter in return to his second libell. By R. L'S.: A WHIPP For the Animadverter in Return to his Second LIBELL. By R. L'S. LONDON: Printed for Henry Brome, at the Gun in Ivy-lane. February the 12th, 1662. An Answer to a Libel, &c. I Had no sooner corrected One Libell, against the Bishop of Worcester, but out bolts...	17.700000	p●mander
A39847.xml	Fludd, Robert,	Mosaicall philosophy: grounded upon the essentiall truth or eternal sapience. Written first in Latin, and afterwards thus rendred into English. By Rob	1659	Mosaicall philosophy: grounded upon the essentiall truth or eternal sapience. Written first in Latin, and afterwards thus rendred into English. By Rob: MOSAICALL PHILOSOPHY: Grounded upon the ESSENTIALL TRUTH OR ETERNAL SAPIENCE. Written first in Latin, and afterwards thus rendred into English. By ROBERT FLUDD, Esq & Doctor of Physick. The Lord giveth Wisdom, and out of his Mouth commeth Knowledg and Understanding, ...	17.100000	pymander, pimander
A25743.xml	Aranda, Emanuel d'	The history of Algiers and it's slavery. With many remarkable particularities of Africk. Written by the Sieur Emanuel D'Aranda, sometime a slave there	1666	The history of Algiers and it's slavery. With many remarkable particularities of Africk. Written by the Sieur Emanuel D'Aranda, sometime a slave there: How the Christian Slaues are beaten at Algiers. THE HISTORY OF ALGIERS And it's SLAVERY. WITH Many Remarkable Particularities of AFRICK. Written by the Sieur EMANVEL D' ARANDA, Sometime a SLAVE there. English'd by JOHN DAVIES of Kidwelly. LONDON, Printed for John S...	16.900000	pysander
A43285.xml	Helmont, Jean Baptiste van, (1577-1644.)	Van Helmont's works: containing his most excellent philosophy, physick, chirurgery, anatomy. Wherein the philosophy of the schools is examin	1664	Van Helmont's works: containing his most excellent philosophy, physick, chirurgery, anatomy. Wherein the philosophy of the schools is examin: Van Helmont's WORKS: Containing his most Excellent PHILOSOPHY, CHIRVRGERY, PHYSICK, ANATOMY. WHEREIN The Philosophy of the Schools is Examined, their Errors Refuted, and the whole Body of Physick REFORMED and RECTIFIED. Being a New rise and progresse of PHILOSOPHY and	14.300000	pymander, pomander

Figure 7. A keyword fuzzy search for 'pymander' within the target corpus (the ~ is the wildcard search operator)

Here (**Figure 7**) we find the 1 'edit' changes to 'pymander', e.g. 'py**s**ander' or '**l**ymander'. Fuzzy searches have their uses, along with wildcard and basic keyword searches, providing another way to investigate and search through the target corpus.

Also, we can make the basic keyword queries more complex, specifying added parameters: perhaps only show texts between specific dates, published in London, that contain the search term as a fuzzy match. Wildcard searches, proximity searches, and exact phrases searches also belong in this suite of capabilities. These kind of keyword searches are a great way to explore the underlying data but are generally not sufficient to answer our research question(s).

In fact, there are some methodological pitfalls which are hard to avoid if one just uses keywords to search a corpus of texts. As Stephen Robertson, among others, has pointed out:

> Search struggles to deal with what lies outside a set of results. In returning only the terms one enters, a search filters out any alternative hypotheses…If we use the wrong search terms, we literally misread our sources…In other words, search radically decontextualizes the results it produces.[27]

Text Matching

Given the limitations of search, we need other ways to investigate our research question. Moving beyond simple keyword searches, we want the ability to match entire passages from our source text with similar ones from the target corpus. Similar lexical phrasing, regardless of the precise words used, should also be discoverable.

Text Matching, as we call this, can be seen as a more ambitious kind of search.[28] It does not take a single keyword as input but instead the entire source text itself–all its sentences collectively. Then we can find the most similar matching sentences or passages (groups of sentences) in the target corpus and rank them based on similarity.

Let us see this in action. First, we would like to identify the most similar matching sentences between source and target:

Most Similar Matching Sentences

	Source Sentence	Most Similar Corpus Sentence	Corpus Author	Corpus Title	Corpus Date	Similarity Score
55165	So doth God in Heaven sow Immortality in the Earth, Change in the whole Life and Motion.	So doth God in Heaven sow Immortality, in the Earth Change in the whole Life, and Motion.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	1.00
55164	Look upon the same Man, planting a vine, or an apple tree, or a fig tree, or some other tree.	Look upon the same Man, planting a Vine, or an Apple-Tree, or a Fig-Tree, or some other Tree.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	1.00
55159	For if he do not make, or do all things, he is either proud, or not able, or ignorant, or envious, which is impious to affirm.	For if he do not make, or do all things, he is either proud, or not able, or ignorant, or envious, which is impious to affirm.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	1.00
55154	But the vicissitude of Generation doth make them, as it were, to blossom out; and for this cause did make change to be, as one should say, The Purgation of Generation.	But the vicissitude of Generation doth make them, as it were to blossom out; and for this cause did make Change to be, as one should say, The Purgation of Generation.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	1.00
55152	For these are Passions that follow Generation, as Rust doth Copper, or as Excrements do the Body.	For these are Passions that follow Generation, as Rust doth Copper, or as Excrements do the Body.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	1.00
55092	But the things that pre-exist, and that are, being changed, are false.	But the things that pre-exist, and that are, being changed, are false.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	1.00
55100	For Death is destruction, but there is nothing in the whole World that is destroyed.	For Death is destruction, but there is nothing in the whole World that is destroyed.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	1.00
55104	The second is the World, made by him, after his own Image, and by him holden together, and nourished, and immortalized, and as from its own Father, ever living.	The second is the World, made by him, after his own Image, and by him holden together, and nourished, and immortalized; and as from its own Father, ever living.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	1.00
55080	It is Truth, and therefore is he only intrusted with the Workmanship of the World, ruling and making all things, whom I do both honour, and adore his Truth; and after the One, and First, I acknowledge him the Workman.	It is Truth, and therefore he is only intrusted with the Workmanship of the World, ruling and making all things, whom I do both honour, and adore his Truth; and after the One, and First, I acknowledge him the Workman.	[no entry]	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	1657	1.00

Figure 8. The most similar matching sentences between source text and target corpus

Above (**Figure 8**) we see something that is validating: sentences from the 1657 edition of the *Divine Pymander* are the top matching sentences to sentences in the

source text (the 1650 edition), which is precisely what we would expect, as they are identical to each other. It validates the logic of our text matching tool. But for illustrative purposes, it is not very interesting, so what happens if we exclude these from our results?

Not all the sentences that match are particularly meaningful (e.g. see the second result in **Figure 9** below). But we also find that many of the most similar sentences come from Thomas Traherne's *Christian ethicks* (1675), and it appears that he is copying directly from the source text (or the 1657 edition), though he sometimes alters the language in minor ways (see the two results I have highlighted).[29] So this is certainly more interesting and worthy of follow-up: is Traherne acknowledging the *Divine Pymander* as his source, or pretending he has written this himself? The VERITRACE Text Matching tool is an early modern plagiarism detector of sorts.

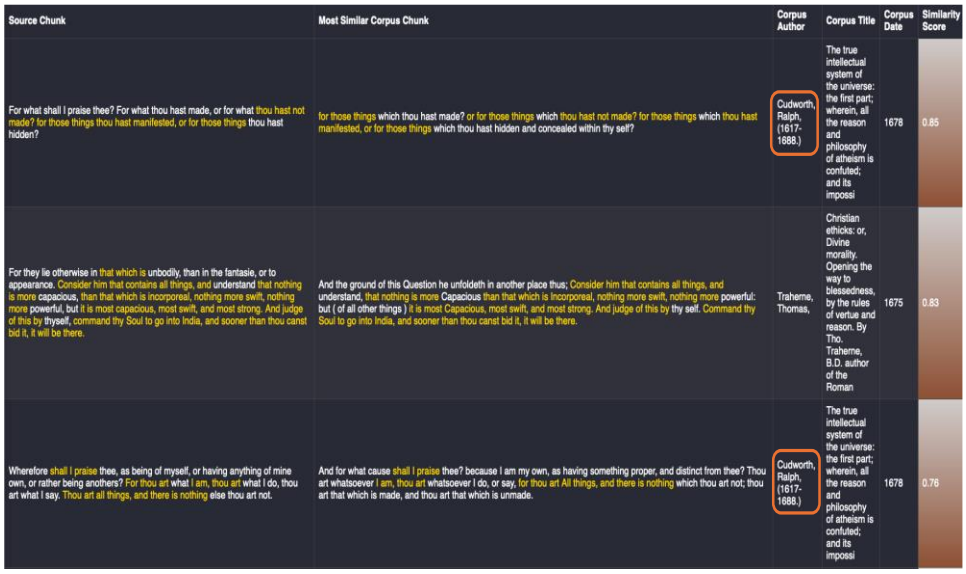
Most Similar Matching Sentences (excluding 1657 edition results)

47002	For it is the greatest Evil, not to know God.	For it is the greatest evil not to know GOD.	Traherne, Thomas,	Christian ethicks: or, Divine morality. Opening the way to blessedness, by the rules of vertue and reason. By Tho. Traherne, B.D. author of the Roman	1675	1.00
15539	How is that, quoth I?	How is that, quoth I?	Boethius,	Summum bonum, or An explication of the divine goodness, in the words of the most renowned Boetius. Translated by a lover of truth, and virtue.	1674	1.00
11848	And it is impossible it should be otherwise.	It is impossible it should be otherwise.	Farindon, Anthony,	LXXX sermons preached at the parish-church of St Mary Magdalene Milk-street, London: whereof nine of them not till now published. By the late eminent	1672	0.99
46994	After this manner, therefore, contemplate God to have all the whole world to himself, as it were, all thoughts, or intellections.	After this manner therefore contemplate GOD to have all the whole World in himself, as it were all Thoughts or Intellections.	Traherne, Thomas,	Christian ethicks: or, Divine morality. Opening the way to blessedness, by the rules of vertue and reason. By Tho. Traherne, B.D. author of the Roman	1675	0.99
46989	Bid it likewise pass over the Ocean, and suddenly it will be there; not as passing from place to place, but suddenly it will be there.	Bid it pass over the Ocean, and suddenly it will be there: not as passing from place to place, but suddenly it will be there.	Traherne, Thomas,	Christian ethicks: or, Divine morality. Opening the way to blessedness, by the rules of vertue and reason. By Tho. Traherne, B.D. author of the Roman	1675	0.98
47001	for thou canst understand none of those Fair and Good things, and be a lover of the body and Evil.	For thou canst understand none of those fair and good things, but must be a lover of the Body and Evil.	Traherne, Thomas,	Christian ethicks: or, Divine morality. Opening the way to blessedness, by the rules of vertue and reason. By Tho. Traherne, B.D. author of the Roman	1675	0.97
30217	But what is the Wisdom of God?	What is the Wisdom of God?	Vincent, Thomas,	An explicatory catechism: or, An explanation of the assemblies shorter catechism. Wherein those principles are enlarged upon especially, which obviate	1675	0.96
47000	But if thou shut up thy Soul in the Body, and abuse it, and say, I understand nothing, I can do nothing, I am afraid of the Sea, I cannot climb up to Heaven, I know not who I am, I cannot tell what I shall be: What hast thou to do with god?	But if thou shut up thy Soul in thy Body, and abuse it; and say, I understand nothing, I am afraid of the Sea, I cannot climb up into Heaven, I know not who I am, I cannot what I shall be; what hast thou to do with GOD?	Traherne, Thomas,	Christian ethicks: or, Divine morality. Opening the way to blessedness, by the rules of vertue and reason. By Tho. Traherne, B.D. author of the Roman	1675	0.95
47004	For there is nothing which is not the Image of God.	For it is nothing, which is not the Image of GOD.	Traherne, Thomas,	Christian ethicks: or, Divine morality. Opening the way to blessedness, by the rules of vertue and reason. By Tho. Traherne, B.D. author of the Roman	1675	0.93

Figure 9. The most similar matching sentences between source text and target corpus, excluding results from the 1657 edition of the *Divine Pymander*

Now, sentence matching is a start, but would it not be more intellectually meaningful if we could find entire passages–groups of sentences–that match between our source text and the target corpus? And indeed, we can. Once again, we first exclude the matching passages from the 1657 *Divine Pymander* edition, and then display the following results:

Matching Sentence Groups (excluding 1657 edition results)



Source Chunk	Most Similar Corpus Chunk	Corpus Author	Corpus Title	Corpus Date	Similarity Score
For what shall I praise thee? For what thou hast made, or for what thou hast not made? for those things thou hast manifested, or for those things thou hast hidden?	for those things which thou hast made? or for those things which thou hast not made? for those things which thou hast manifested, or for those things which thou hast hidden and concealed within thy self?	Cudworth, Ralph, (1617-1688.)	The true intellectual system of the universe: the first part; wherein, all the reason and philosophy of atheism is confuted; and its impossi	1678	0.85
For they lie otherwise in that which is unbodily, than in the fantasie, or to appearance. Consider him that contains all things, and understand that nothing is more capacious, than that which is incorporeal, nothing more swift, nothing more powerful, but it is most capacious, most swift, and most strong. And judge of this by thyself, command thy Soul to go into India, and sooner than thou canst bid it, it will be there.	And the ground of this Question he unfoldeth in another place thus; Consider him that contains all things, and understand, that nothing is more Capacious than that which is incorporeal, nothing more swift, nothing more powerful: but (of all other things) it is most Capacious, most swift, and most strong. And judge of this by thy self. Command thy Soul to go into India, and sooner than thou canst bid it, it will be there.	Traherne, Thomas,	Christian ethicks: or, Divine morality. Opening the way to blessedness, by the rules of vertue and reason. By Tho. Traherne, B.D. author of the Roman	1675	0.83
Wherefore shall I praise thee, as being of myself, or having anything of mine own, or rather being anothers? For thou art what I am, thou art what I do, thou art what I say. Thou art all things, and there is nothing else thou art not.	And for what cause shall I praise thee? because I am my own, as having something proper, and distinct from thee? Thou art whatsoever I am, thou art whatsoever I do, or say, for thou art All things, and there is nothing which thou art not; thou art that which is made, and thou art that which is unmade.	Cudworth, Ralph, (1617-1688.)	The true intellectual system of the universe: the first part; wherein, all the reason and philosophy of atheism is confuted; and its impossi	1678	0.76

Figure 10. The most similar matching passages (groups of sentences) between source text and target corpus, excluding results from the 1657 edition of the *Divine Pymander*

The most similar passage between the source text and the target corpus comes from Ralph Cudworth's *The True Intellectual System of the Universe* (1678)–see **Figure 10**. Scholars have known about the influence of the *Corpus Hermeticum* on the so-called Cambridge Platonists like Ralph Cudworth and Henry More for some time, and here is direct, lexical proof.[30]

Our Text Matching tool highlights all matching words from the passages in a yellow colour, so one can see how they overlap or differ. Notice that the passages in question are not exact matches; instead, they have minor differences in language and meaning, yet the corpus passages are clearly drawn from the original source ones. We are observing shades of influence. This is what we hoped to see, and there are a variety intellectual questions that could be pursued here, with just this small sample of results.

Sentiment Analysis

If we can search and match similar lexical phrases, we will also be able to conduct sentiment analysis on those phrases. While not a search feature per se, it is an important capability we wish to have (it is not part of our case study). **Sentiment Analysis** (**SA**) is used to analyse the sentiment that occurs in the discussion of a particular topic in a corpus of texts. It has been successfully used in historical projects, for example involving nineteenth-century English language newspapers and how news

items were reappropriated for particular audiences between the UK, Ireland, and the USA,[31] the role of women in the vanguard of the abolition movement,[32] and differences in which American newspapers covered the Civil War,[33] but also on modern Twitter discourse.[34] When we are reading a small number of texts, we can easily read each text individually and record with what sentiment a topic is being discussed: positive or negative, praising or condemning, with caution or with exuberance.

It becomes much more challenging when we try to analyse a larger corpus of several hundred thousand texts in multiple languages. Yet the sentiment found in the larger corpus would be much more representative than that drawn from just a sample of texts. As such, the ability to perform semi-automated SA on a larger corpus allows us to draw meaningful conclusions about the general thoughts and ideas surrounding a particular topic, text, or author. SA will allow us to measure the popularity and impact of the authors and works included in the CRC, differentiate between groups of readers in time and space, and provide answers to questions like, for example, whether Protestant authors were generally more susceptive to the 'truths' conveyed by the Sibylline Oracles than their Catholic equivalents. It will also allow us to measure the impact of events such as Casaubon's debunking of the *Corpus Hermeticum*, or Isaac Vossius's reinterpretation of the antiquity of the ancient wisdom texts.

Latent Thematic Analyses

Thus far the techniques have been directly anchored in the specific words of the texts – and rightly so, especially when it comes to search, text matching, and sentiment analysis. But there are other approaches that are better at finding latent or indirect, structures of meaning in our texts, providing a more nuanced understanding.

There are two related techniques in the Digital Humanities that have been used to uncover underlying structures of meaning in a collection of documents in the forms of themes, concepts, or topics. Both techniques–Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) –move beyond literal word matches to reveal what topics or themes pervade a textual corpus without needing predefined categories or relying on the exact words used. That is what makes them very complementary to our lexical Text Matching tool.

Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) is an NLP technique that allows for the comparison of texts within a large corpus. It compares the usage of words and passages, terminology, and phrasing, and enables us to detect both significant differences and similarities, even across language boundaries.[35] LSA makes use of open-source packages which are robust, user-friendly, and highly adaptable, and which can be integrated in any digital research environment.

One way to think of LSA is to imagine being faced with organising a huge collection of books. Then imagine skimming through the pages of all the books, looking for similar patterns in words and sentences. One could then group books together that seem to share these patterns, even if they do not explicitly discuss the same topics or use the same words for the same concept.

Think of LSA as a librarian who organizes books by the similarity in their writing style–prose, use of metaphors, complexity–regardless of the books' specific content. LSA creates sets of books based on the overall flavour of the language used, rather than specific, clear-cut topics. LSA reveals 'stylistic flavours' found in the text corpus – its *linguistic vibes*, so to speak.

Within this project, LSA will be performed on both the close and distant reading corpora. On the CRC, LSA will help analyse the influence of the various Renaissance authors upon each other, to detect the use and reuse of particular passages even though sometimes formulated rather differently. It has a major advantage over direct citation analysis–useful in itself, as has been shown recently[36]–in that LSA has no trouble working with vague referencing, multiple languages and so-called fuzzy data, including suboptimal OCR.[37] For the Distant Reading Corpus, LSA will allow us to identify the use of authors from the Close Reading Corpus in other Renaissance and early modern writings, be it verbatim or paraphrased. In both cases, LSA will strongly help to detect the influence of individual authors upon one another and upon larger bodies of text present in Renaissance and early modern society, influences hitherto unrecognised or underestimated.

Returning to our research question about the influence of the 1650 edition of the *Divine Pymander*, we can apply Latent Semantic Analysis to uncover the most similar documents–based on LSA, or 'linguistic vibes'–to our source text. First, we plot a visualisation of the application of LSA:

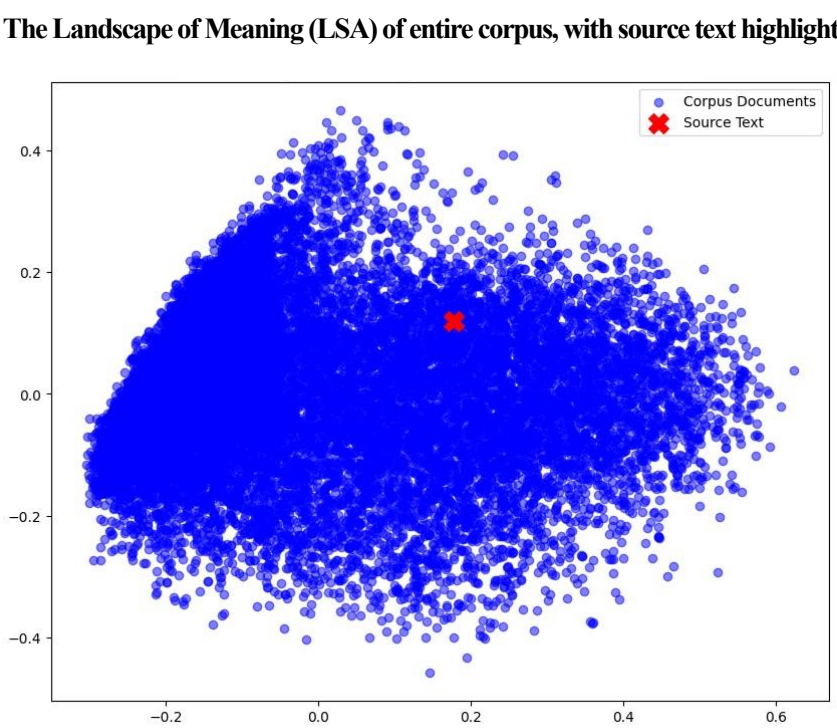


Figure 11. A visualisation of Latent Semantic Analysis (LSA) applied to our target corpus, with source text highlighted with a red X. Each dot represents a single document.

In this plot (**Figure 11**), each point (blue dot) represents a document from our target corpus–all 18,633 texts.[38] What is critical to understand is that the position of a point in the plot reflects the 'meaning' of the document represented by that dot, as understood by LSA. Documents that are similar to each other (in terms of their stylistic flavour, remember) are located close together. The red 'X' represents the location of our source text within the overall corpus. The closer other blue dots are to this 'X', the more similar those documents are to the source text. In summary, this plot is a way of visualising the 'landscape of meaning' within a collection of documents.

There is too much information in this visualisation for us to process. Would it not be helpful if we could zoom in and see which specific texts–say, the top 100–are most similar in meaning to the source text? We can:

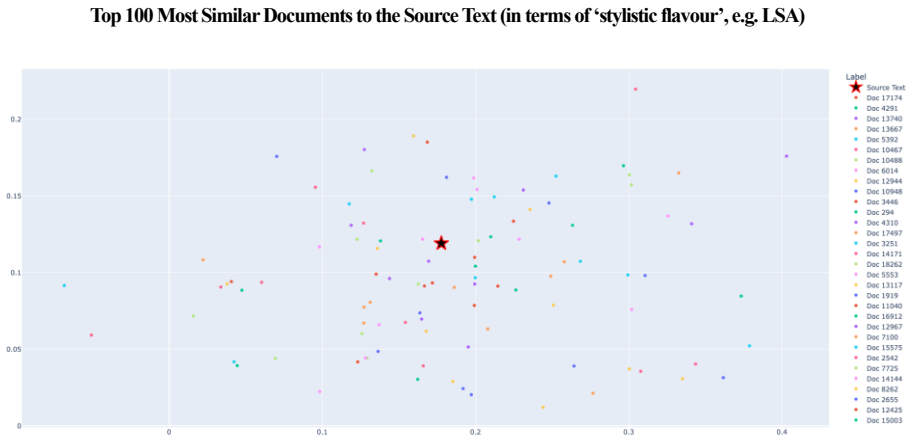


Figure 12. A 'zoomed in' view of Latent Semantic Analysis (LSA) applied to our target corpus, showing the top 100 most similar documents to the source text.

Above (**Figure 12**) we see a 'zoomed in' version of the 'landscape of meaning', centred around the source text, showing just the top 100 most similar documents. In our LSA tool, this plot is interactive (**Figure 13** below), such that hovering over each dot (document) shows its similarity score, along with title, author, and date.

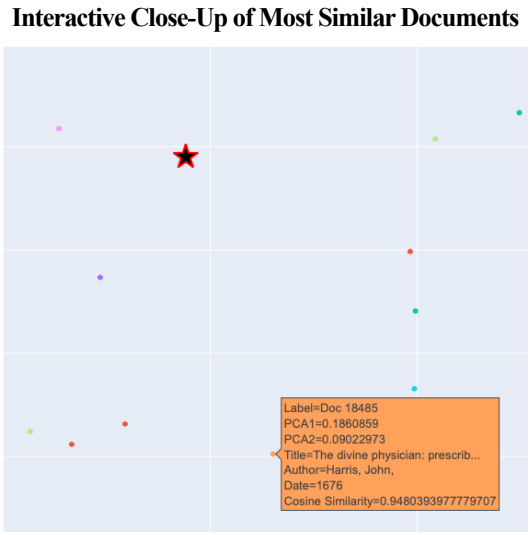


Figure 13. An interactive 'zoomed in' visualisation of Latent Semantic Analysis (LSA) applied to our target corpus, showing the top dozen or so most similar documents to the source text. Hovering over a dot shows additional information about that text.

We can see the most similar documents in list form as well (**Figure 14** below):

Most Similar Documents to the 1650 *Divine Pymander* using LSA ('linguistic vibes')

Label	Title	Author	Date	Similarity Score
Source Text	Source Text Title	Source Text Author	Source Text Date	1
Doc 17174	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius containing fifteen chapters,	[no entry]	1657	0.997
Doc 4291	Conjectura cabbalistica: or, a conjectural essay of interpreting the minde of Moses, according to a threefold cabbala: viz. literal, philosophical, m	More, Henry,	1653	0.989
Doc 13740	The felicity of a Christian life. By Hierome Savonarola.	Savonarola, Girolamo,	1651	0.981
Doc 13667	The beauty and order of the creation. Together with natural and allegorical meditations on the six dayes works of the creation. With the addition of t	Maynard, John	1668	0.98
Doc 5392	Psychosophia: or, Natural & divine contemplations of the passions & faculties of the soul of man. In three books. By Nicholas Mosley, Esq;	Mosley, Nicholas,	1653	0.977
Doc 10467	The grand prerogative of humane nature: namely, the souls naturall or native immortality, and freedome from corruption, shewed by many arguments, and	Holland, Guy, 1587?-1660.	1653	0.977
Doc 10488	Concerning the election of grace. Or Of Gods will towards man. Commonly called predestination.: That is, how the texts of Scripture are to	Böhme, Jakob, 1575-1624.	1655	0.976
Doc 6014	A work for none but angels & men that is, to be able to look into, and to know our selves. Or A book shewing what the soule is, subsisting and having	Jenner, Thomas,	1658	0.974
Doc 12944	Natural theology, or, The knowledge of God, from the works of creation; accommodated, and improved, to the service of Christianity. By Matthew Barker	Barker, Matthew,	1674	0.974
Doc 10948	Agnoia tou psychikou anthrōpou, or, The inability of the highest improved naturall man to attaine a sufficient and right knowledge of indwelling sinn	Hurst, Henry,	1659	0.973
Doc 3446	A posing question, put by the wise man, viz. Solomon, to the wisest men concerning making a judgment of the temporal conditions : wherein you have the	Baxter, Benjamin, Preacher of the Gospel	1662	0.972

Figure 14. A list of the top 11 most similar documents (using LSA) to the source text that are part of the target corpus

Again, we see–as expected–that the 1657 edition of the *Divine Pymander* is the most similar document with a similarity score just shy of 1 (1 means identical, 0 means completely dissimilar). The next 10 most similar documents are ranked after this, with Henry More's *Conjectura cabbalistica* being the most similar text (using LSA) after this. Each one of these results, and many more, could be investigated further, so that we can discover why LSA picks them out as being particularly similar in 'stylistic flavour' to the source text. Whether this measures true intellectual influence or not is a question that these tools cannot answer by themselves – that is a matter for traditional historical scholarship.

LDA (Topic Modelling)

Latent Dirichlet Allocation (LDA), aka topic modelling, is a similar but distinct analytic technique. Sticking with the library metaphor: LDA is like cataloguing each book in a library by identifying a set of topics it contains. LDA interprets clusters of words as indicators of topics – it is not just that certain words occur together, but that their co-occurrence represents a coherent topic. LDA helps one understand exactly what topics are inside each book and in what proportion, making it easier to find books on a specific subject.

LDA will be employed to help identify information clusters (topics) where, for instance, authors might be following each other's patterns of evidencing. It can also be used to identify entire works, or parts of works, that cover similar topics, even

though the language used hardly overlaps, for instance when the texts are in different languages. It may also help us identify lesser-known works (the 'great unread' mentioned above) that share common topics with more studied works that are known to scholars.

Returning to our research question, let us apply LDA analysis to both our source text and the target corpus. First, we apply LDA to the entire target corpus to see what topics emerge. Here is the Topic Map for the entire corpus:

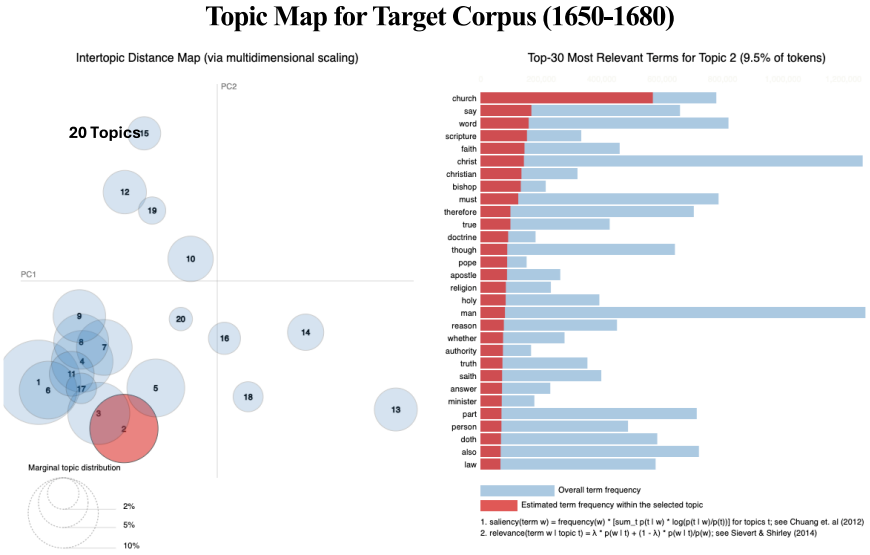


Figure 15. A Topic Map for the target corpus, showing 20 topics and their relation to each other, as well as the most relevant terms for each topic.

There are two panels here (**Figure 15** above): on the left-hand side, we have the Intertopic Distance Map, showing the 20 topics that constitute the corpus. Each topic is represented by a circle and the size of the circle represents the relative frequency of the topic within the corpus. Topics that are more similar to each other are closer together on the map. On the right-hand side, there is a bar chart that displays the top 30 most relevant terms for each topic (in the visualisation above, Topic 2 is shown).[39]

Next, to compare the topics in the corpus to those from our source, we need to apply LDA to the source text. This reveals what LDA considers to be the dominant topics contained in the 1650 *Divine Pymander*, as well as the key terms that constitute each specific topic found in the text (see **Figure 16** below).

Dominant Topics in Source Text:
Top Keywords per Topic
in Source Text

Topic 2 (38.90%):	Topic 16 (37.88%):	Topic 1 (18.32%):
• nature	• Christ	• man
• body	• spirit	• unto
• self	• thou	• say
• man	• man	• well
• reason	• light	• word
• part	• unto	• therefore
• though	• word	• saith
• must	• world	• others
• spirit	• power	• never
• world	• life	• self

Figure 16. A list of the dominant topics in the source text, with the top keywords per topic. These 3 topics account for more than 95% of the content in the source text.

It is important to underscore that the algorithm–LDA–is determining these topics from underlying word frequencies found in the corpus. That is all. The topics therefore may not, and often will not, "align with any human reader's experience of a text…what is being predicted is something no human reads or produces."[40] This is why it can be helpful, as we have done above, to see a list of the most relevant terms per topic, even if the topic itself does not align with a human-generated one.

From this foundation, we can now obtain a list (**Figure 17**) of the most similar documents to the source text, as interpreted through the lens of LDA, i.e. through comparing the topics contained in the texts:

Most Similar Documents by Topic (LDA)

Label	Title	Author	Date	Similarity Score
Doc 17174	Hermes Mercurius Trismegistus, his Divine pymander, in seventeen books. Together with his second book, called Asclepius; containing fifteen chapters,	[no entry]	1657	0.997
Doc 1175	The arrogancy of reason against divine revelations, repressed. Or, Proud ignorance the cause of infidelity, and of mens quarrelling with the	Baxter, Richard	1655	0.957
Doc 10504	Four tables of divine revelation signifying what God in himself is, without nature; and how considered in nature; according to the three pri	Böhme, Jakob	1654	0.947
Doc 5751	Magick e.t.c. astrology vindicated :$bfrom those false aspersions and calumnies, which the ignorance of some hath cast upon them. In which is containe	[no entry]	1651	0.947
Doc 1734	A discourse of the freedom of the will. By Peter Sterry, sometimes fellow of Emmanuel Colledge in Cambridge.	Sterry, Peter,	1675	0.94
Doc 3596	An account of the Oriental philosophy, shewing the wisdom of some renowned men of the East; and particularly, the profound wisdom of Hai Ebn Yokdan, b	Ibn Ṭufayl, Muḥammad ibn ʻAbd al-Malik,	1674	0.938
Doc 7322	The dominion of the seed of God throughout all generations: or, The heighth, and breadth, and length, and depth of the love of God, which passeth know	Bishop, George,	1667	0.928
Doc 18233	The displaying of supposed witchcraft. Wherein is affirmed that there are many sorts of deceivers and impostors, and divers persons under a	Webster, John, (1610-1682.)	1677	0.927
Doc 15590	Apokrypta apokalypta. = Velata quædam revelata: some certain, hidden, or vailed spiritual verities revealed. Upon occasion of various very prying, and	Fisher, Samuel	1661	0.919
Doc 18306	The new witnesses proved old hereticks: or Information to the ignorant; in which the doctrines of John Reeve and Lodowick Muggleton, which they stile,	Penn, William,	1672	0.915
Doc 6133	The way to the city of God described, or, A plain declaration how any man may within the day of visitation given him of God, pass out of the unrighteo	Keith, George,	1678	0.907

Figure 17. A list of the most similar documents to the source text (using LDA) found in the target corpus.

Once again, the 1657 edition comes first. But the list of most similar documents is different from those identified by LSA because we are not focused on 'linguistic vibes' as the similarity criterion, but the topics contained within the texts. Note that LDA can handle some mistakes in OCR (see the extraneous symbols in the fourth record, for instance). Regardless, we now have a different set of texts to investigate, and the historian can make a closer study to determine if these similarities are intellectually meaningful or not, and if so, how that changes our understanding of the influence of the first-ever English translation of the 1650 *Divine Pymander* upon a subsequent generation of printed works in English.

Note that we cannot definitively answer our research question using these techniques alone. In each case, they present intellectual threads to unwind, which we do using the traditional tools of research-driven scholarship.

A final point about capabilities: we are just scratching the surface of what one can do, even with these traditional tools. For Text Matching, we have only used one source text, but we can do so for our entire Close Reading Corpus or a subset of that, e.g. all editions of the *Corpus Hermeticum*. We could then look for matches and similarities to this larger source collection. We also will use our entire multilingual VERITRACE text collection of c.430,000 texts as the target corpus–not the 18,633 English texts we limited ourselves to for this specific research question. In short, the capabilities of VERITRACE will be expanded significantly, as we proceed.

III. Conclusions, or Reflections on the Project So Far

In terms of lessons learned and insights gained, the following jump out.

Standardised APIs

While working with access to multiple library catalogues, it was at times frustrating to have to learn new API terms to write scripts for each one. It is probably unrealistic, but one wishes for standard API language across library collections, no matter the catalogue queried. For instance, knowing how to query the Gallica search API did not help us query the BSB database. This is a larger problem that will likely not be remedied anytime soon.

Search Results and Bulk Downloads

The public search interfaces for the library catalogues could be better tailored to the needs of the digital humanities community. For instance, while Gallica provides the ability to download the results of a search query as a CSV file, BSB does not. This is the kind of feature that is helpful for digital humanities researchers, as it facilitates data exploration. This also applies to potential features that might support large distant reading projects, like VERITRACE. Why not include the possibility – after completing a registration form – for bulk downloads of digital texts? The Library of Congress has offered this function for some time now, and it would significantly support digital humanities and other research.[41] Of course, it requires a certain amount of infrastructure and staff, but if the feature were limited to a subset of (registered) users, many libraries could manage this.

High-Quality-OCR Digital Text Repository as Shared Resource

Finally, there is far too much repetition of work amongst different digital humanities projects. There seems to be no online database of digital texts that have been confirmed to be of high-quality (e.g. OCR quality of >80%). While there are many databases of digital texts, and many ways of accessing digital texts via API (e.g. Google Books, Internet Archive, or HathiTrust), what one really desires in a digital humanities project is a source of *confirmed high-quality* digital texts, so that downstream natural language processing can proceed confidently. I am not aware of any online database that collates these texts and provides information on their OCR quality. If such a database existed and offered API access to its contents, then it would become a shared resource of great value to the digital humanities community. After all, why should each digital humanities project have to repeat the work that others have done before? Again, this would be a great joint infrastructure project for DARIAH-EU.

Returning to VERITRACE itself, we can say that, already, after an initial year of work, we have many intellectual breadcrumbs to follow. We look forward to sharing our preliminary data and results with the academic community, and, as we continue to build our data pipeline and digital tools, we anticipate exciting scholarly research.

References

[1]The VERITRACE project website is https://veritrace.eu.

[2] Much of the text in this introductory section describing the VERITRACE project has been taken and adapted from the VERITRACE ERC-funding proposal. For a condensed version of this proposal, see Schilt, C.J., "*Traces de la Verité*: The Reappropriation of Ancient Wisdom in Early Modern Natural Philosophy, VERITRACE (ERC-2022-STG-101076836)", [Online] Available via https://veritrace.eu/wp-content/uploads/2023/04/Project-Traces-de-la-Verite-Condensed.pdf, cited 15.07.2024.

[3] For example, Latin is not available as a default trained pipeline package in the open-source natural language processing library *spaCy* (see https://spacy.io/usage/models), although Patrick J. Burns has created *LatinCy* to fill this gap (see https://spacy.io/universe/project/latincy). The *Natural Language Toolkit* (*NLTK*) likewise has limited support for Latin (e.g. for tokenisation), though the *Classical Language Toolkit* (*CLTK*) – which does support Latin – has been developed to supplement this (see http://cltk.org). Another example: *OpenSearch* has no built-in language analyser for Latin (see https://opensearch.org/docs/latest/analyzers/language-analyzers/).

[4] Cohen, M., "Narratology in the Archive of Literature", *Representations* 108 (2009); Reid, D., "Distant Reading, 'the Great Unread', and 19th-Century British Conceptualizations of the Civilizing Mission: A Case Study", *Journal of Interdisciplinary History of Ideas* 15 (2019).

[5] Hill, M.J., Hengchen, S., "Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study", *Digital Scholarship in the Humanities* 34/4 (2019); Kurhekar, P., Nigam, S., Pillai, S., "Automated Text and Tabular Data Extraction From Scanned Document Images", *Data Management, Analytics and Innovation, Proceedings of ICDMAI 2021*, ed. N. Sharma, A. Chakrabarti, V. E. Balas, A. M. Bruckstein, 1 (2021), 169-182.

[6]Imai, K., *Quantitative Social Science: An Introduction* (Princeton and Oxford: Princeton University Press, 2018); Karsdorp, F., Kestemont, M., Riddell A, *Humanities Data Analysis: Case Studies With Python* (Princeton and Oxford: Princeton University Press, 2021).

[7] The Text Creation Partnership (TCP) is a "consortium of (mostly) university and college libraries that have joined together to create standardized, accurate, and faithful XML/SGML-encoded electronic text editions of early printed books…To date, the project has created more than 70,000 transcribed and encoded historical texts." See https://textcreationpartnership.org/about-the-tcp/, cited 15.07.2024.

[8] For distant reading over a large corpus of texts, see Moretti, F., *Graphs, Maps, Trees: Abstract Models for Literary History* (London and Brooklyn, NY: Verso, 2005); Moretti, F., *Distant Reading* (London and Brooklyn, NY: Verso, 2013); Beals, M.H., "Stuck in the Middle: Developing Research Workflows for a Multi-Scale Text Analysis", *Journal*

of Victorian Culture 22/2 (2017); Underwood, T., *Distant Horizons: Digital Evidence and Literary Change* (Chicago: University of Chicago Press, 2019).

[9]Kuhn, H.C., "Counting What May Count Regionally. The Presence of Prints of Works By Frane Petrić in German Libraries", *Synthesis Philosophica* 21 (2006); Palumbo, M., "Books on the Run: The Case of Francesco Patrizi", in *Fruits of Migration: Heterodox Italian Migrants and Central European Culture 1550-1620*, ed. C. Zwierlein, V. Lavenia (Leiden: Brill, 2018); Margocsy, D., Somos, M., Joffe, S.N., *The Fabrica of Andreas Vesalius: A Worldwide Descriptive Census, Ownership, and Annotations* (Leiden and Boston: Brill, 2018); Feingold, M., Svorenčík, A., "A Preliminary Census of Copies of the First Edition of Newton's *Principia* (1687)", *Annals of Science* 77 (2020): 253–348.

[10] The CRC has already grown from its initial core of c. 80 works to more than 130 editions today.

[11] This article offers some critical reflections of our work in progress but is not sufficiently theory-laden to put it in the realm of critical data studies, or similar approaches. For work of this type that is in some ways comparable, see Dobson, James E., *Critical Digital Humanities: The Search for a Methodology* (Urbana, IL: University of Illinois Press, 2019). Or, in the realm of science, see, e.g. Lowndes, J.S.S. et al. "Our path to better science in less time using open data science tools", *Nature Ecology & Evolution*, 1, 0160 (2017), https://doi.org/10.1038/s41559-017-0160.

[12] Vasiliev, Y., *Python for Data Science: A Hands-On Introduction* (San Francisco: No Starch Press, 2022), 9.

[13] See Microsoft Corporation, "What is a Data Lake?" [Online] Available via https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-a-data-lake, cited 15.07.2024.

[14] See Gallica, "Advanced Search", [Online] Available via https://gallica.bnf.fr/services/engine/search/advancedSearch/, cited 15.07.2024.

[15] See Münchener Digitalisierungszentrum Digitale Bibliothek, "Search the Digital Collections", [Online] Available via https://www.digitale-sammlungen.de/en, cited 15.07.2024.

[16] The Pandas Development Team, pandas-dev/pandas: Pandas, 20.01.2024, v.2.2.0, Zenodo, https://doi.org/10.5281/zenodo.10537285

[17] Sölch, D., Münchener Digitalisierungszentrum (2023). Email to the author, 09.11.2023.

[18] The raw EEBO texts are available for download for free. Information about how and where to download them can be found online here: Text Creation Partnership, "Can I download the raw files?", [Online] Available via https://textcreationpartnership.org/faq/#faq05, cited 15.07.2024.

[19] The Text Encoding Initiative (TEI) is an international consortium with the mission "to develop and maintain guidelines for the digital encoding of literary and linguistic texts." They publish the widely used 'Text Encoding Initiative Guidelines for Electronic Text Encoding and Interchange.' See Text Encoding Initiative, "Text Encoding Initiative: About", [Online] Available via https://tei-c.org/about/, cited 15.07.2024. EEBO is freely searchable, at least in part, as part the University of

Michigan's Digital Collections, available online at https://quod.lib.umich.edu/e/eebogroup/, cited 15.07.2024.

[20] Led by Joseph Loewenstein and Martin Mueller, *EarlyPrint* "is a collaborative effort – centered doubly at Northwestern University and Washington University in St. Louis – to transform the early English print record, from 1473 to the early 1700s, into a linguistically annotated and deeply searchable text archive." See *EarlyPrint*, "EarlyPrint: About", [Online] Available via https://earlyprint.org/about/, cited 15.07.2024.

[21]*Project Quintessence* can be accessed online here: *Project Quintessence*, [Online] Available via http://quintessence.ds.lib.ucdavis.edu, cited 15.07-2024. The EECO-TCP website has a longer list of related projects that use TCP texts here: Text Creation Partnership, "Projects and publications using TCP texts", [Online] Available via https://textcreationpartnership.org/using-tcp-content/projects-and-publications-using-tcp-texts/, cited 15.07.2024.

[22] Sölch, D., Münchener Digitalisierungszentrum (2024). Email to the author, 14.02.2024.

[23] VERITRACE would like to thank in particular Dr. Klaus Ceynowa, the Director General of the Bayerische Staatsbibliothek, for his support, as well as the entire staff of the Munich Digitisation Centre.

[24]Hill, M.J., Hengchen, S., (2019); Kurhekar, P., Nigam, S., Pillai, S., (2021); Sangiacomo, A., Hogenbirk, H., Tanasescu, R., Karaisl, A., White, N., "Reading in the Mist: High-Quality Optical Character Recognition Based on Freely Available Early Modern Digitized Books", *Digital Scholarship in the Humanities* 37/4 (2022).

[25] John Everard's 1650 work was entitled, in full, *The divine Pymander of Hermes Mercurius Trismegistus, in XVII. books. Translated formerly out of the Arabick into Greek, and thence into Latine, and Dutch, and now out of the original into English; by that learned divine Doctor Everard* (London: printed by Robert White, 1650). A digital transcription of this text can be found online at https://sacred-texts.com/eso/pym/index.htm, cited 15.07.2024.

[26] The search algorithm in this example relies on the Term Frequency – Inverse Document Frequency (TF-IDF) method. TF-IDF is a statistical method which "measures how important a term is within a document relative to a collection of documents (i.e., relative to a corpus)...importance of a term is high when it occurs a lot in a given document and rarely in others." See LearnDataSci, "TF-IDF – Term-Frequency-Inverse Document Frequency", [Online] Available via https://tinyurl.com/2pmdn9hp, cited 15.07.2024.

[27] Robertson, S., "The Differences between Digital Humanities and Digital History", *Debates in the Digital Humanities* (2016), [Online] Available via https://dhdebates.gc.cuny.edu/read/untitled/section/ed4a1145-7044-42e9-a898-5ff8691b6628, cited 15.07.2024.

[28] Several projects have conducted 'text re-use analysis' or 'text similarity analysis', as we do here. See, e.g. Rosson, D. E., Mäkelä, E., Vaara, V., Mahadevan, A., Ryan, Y. C., & Tolonen, M. (2023). "Reception Reader: Exploring Text Reuse in Early Modern British Publications". *Journal of open humanities data*, *9*,

https://openhumanitiesdata.metajnl.com/articles/10.5334/johd.101, and Ryan Y., Mahadevan A. and Tolonen M. (2023). "A Comparative text similarity analysis of the works of Bernard Mandeville". *Digital Enlightenment Studies*, 1, 28–58. DOI: 10.61147/des.6

[29] The connection between Traherne and the *Corpus Hermeticum* is not a novel observation, though the Text Matching tool is a new way to observe it. Already by the 1960s, historians were connecting the two. See, e.g., Marks, C., "Thomas Traherne and Hermes Trismegistus", *Renaissance News* 19/2 (1966): 118-131.

[30] A classic discussion can be found in Walker, D.P., *The Ancient Theology: Studies in Christian Platonism from the Fifteenth to the Eighteenth Century* (Ithaca, NY: Cornell University Press, 1972).

[31] Beals, M.H., (2017).

[32] Soni, S., Klein, L., Eisenstein, J., "Abolitionist Networks: Modeling Language Change in Nineteenth-Century Activist Newspapers", *Journal of Cultural Analytics* 1 (2021).

[33]Nelson, R.K., "Mining the Dispatch", (2010) [Online] Available via https://dsl.richmond.edu/dispatch/, cited 15.07.2024.

[34]Colleoni, E., Rozza, A., Arvidson, A., "Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data", *Journal of Communication* 64 (2014); Lashari, I.A., Wiil, U.K., "Monitoring Public Opinion By Measuring the Sentiment of Retweets on Twitter", *Proceedings of the 3rd European Conference on Social Media, EM Normandie, Caen, France, 12-13 July, 2016* (Reading, UK: Academic Conferences and Publishing International Ltd, 2016), 153-161.

[35] Foltz, P.W., "Discourse Coherence and LSA", in *Handbook of Latent Semantic Analysis*, ed. T.K. Landauer, D.S. McNamara, S. Dennis, W. Kintsch (New York and London: Routledge, 2011), 169–84; Kintsch, W., "LSA and Meaning: In Theory and Application", in *Handbook of Latent Semantic Analysis*, ed. T.K. Landauer, D.S. McNamara, S. Dennis, W. Kintsch (New York and London: Routledge, 2011), 467–79; Ratna, A.A.P., Purnamasari P.D., Adhi, B.A., Ekadiyanto, F.A., Salman, M., Mardiyah, M., Winata, D.J., "Cross-Language Plagiarism Detection System Using Latent Semantic Analysis and Learning Vector Quantization", *Algorithms* 10 (2017): 69; Dobson, J.E. *Critical Digital Humanities: The Search for a Methodology* (Urbana, IL: University of Illinois Press, 2019).

[36]Winnerling, T., "Moving Around in Narrowing Circles: How Four Scholars Got Forgotten in Eighteenth-Century Learned Journals", *Journal for the History of Knowledge* 2/1 (2021).

[37] Zhang, R., "Hierarchical and Pairwise Document Embedding for Plagiarism Detection", *Advanced Data Mining and Applications: 16th International Conference, ADMA 2020, Foshan, China, November 12-14, 2020, Proceedings* (2021): 148–56.

[38] The meanings of the X and Y axes are not straightforward. But for the technical-minded, this plot uses the Truncated SVD mathematical technique, a type of Singular Value Decomposition, to perform LSA. LSA itself performs dimensionality reduction,

taking the high-dimensional term-document matrix and reducing it to 8 dimensions. However, for the purpose of human visualisation, we reduce this further to just two dimensions (represented by the X and Y axes) using Principal Component Analysis (PCA). For additional details, see Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W., (ed.), *Handbook of Latent Semantic Analysis* (New York and London: Routledge, 2007).

[39] For a sample explanation of this type of visualisation (using different data), see https://www.kaggle.com/code/ykhorramz/lda-and-t-sne-interactive-visualization?scriptVersionId=1508230&cellId=22, cited 15.07.2024.

[40]Karsdorp, F., Testemont, M., Riddell, A., (2021). Also available online at "Mixed-Membership Model of Texts", [Online] Available via https://www.humanitiesdataanalysis.org/vector-space-model/notebook.html, cited 15.07.2024.

[41] Library of Congress, "Collection Items", [Online] Available via https://www.loc.gov/collections/selected-datasets/?fa=contributor:library+of+congress.+cataloging+distribution+service, cited 15.07.2024.