From ancient wisdom to computational humanities

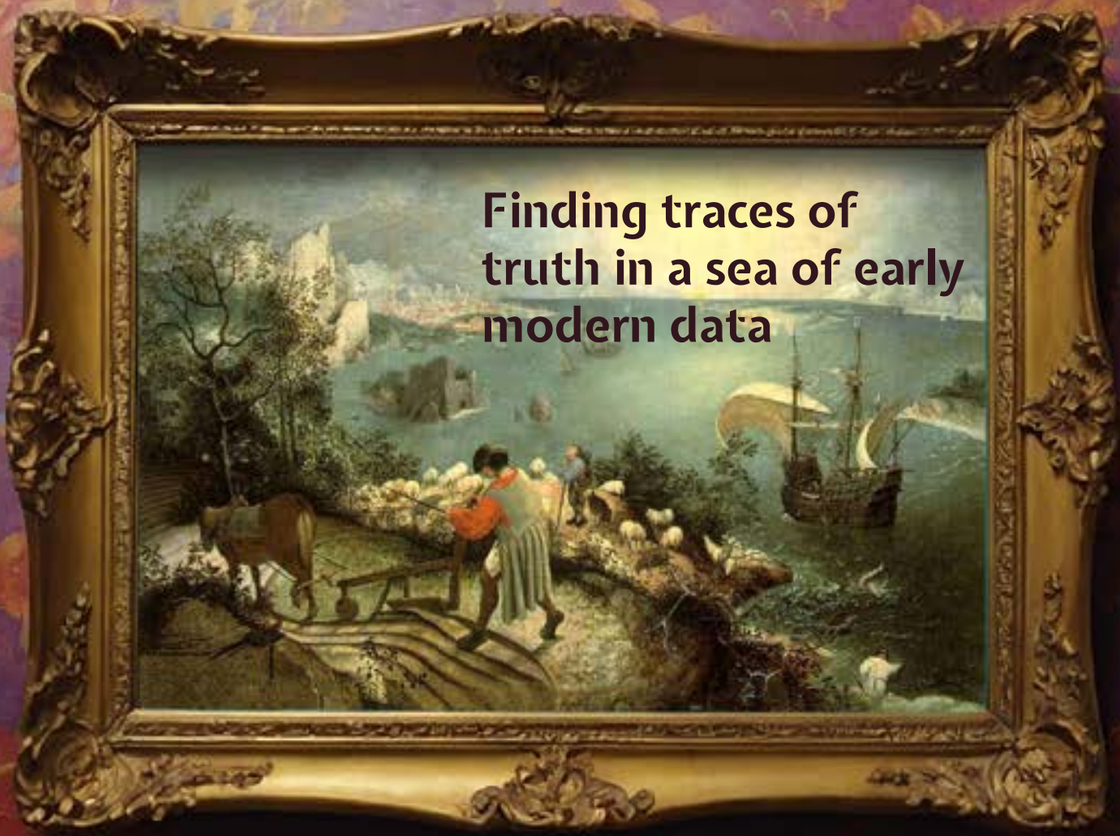

Figure 1: Pieter Bruegel the Elder, Landscape with the Fall of Icarus (c. 1560). A visual metaphor for intertextuality: the myth of Icarus from Ovid's Metamorphoses is subtly recontextualised within an everyday pastoral and maritime scene.

Professor Cornelis J. Schilt and the VERITRACE team
Vrije Universiteit Brussel

The European Research Council (ERC) Starting Grant project VERITRACE, or *Traces de la Verité: The Reappropriation of Ancient Wisdom in Early Modern Natural Philosophy*, is devoted to understanding how early modern scientists such as Nicolaus Copernicus, Johannes Kepler, Francis Bacon and Isaac Newton, considered knowledge to be not something new, but something old. Texts such as the *Corpus Hermeticum*, the *Chaldean Oracles*, the *Orphic Hymns* and the *Sibylline Oracles*, all understood to have been composed in antiquity, were interpreted by many as containing more than just ancient wisdom: they encompassed knowledge about the cosmos lost to mankind, waiting to be discovered. Many of these writings were only translated into Latin during the Renaissance, and their availability in print fuelled speculations about a perennial tradition of wisdom and knowledge stretching back millennia. All over Europe, editions and translations into the vernacular appeared, which graced the shelves not only of scholars of antiquity but also of those who devised new experiments and new theories about the world in and around us. Yet we hardly know what these luminaries took from these writings or how these sources influenced their knowledge-making because of diffuse early modern citation practices and how alien these writings are to us today. What we need, in essence, is an early modern plagiarism detector.

With this in mind, VERITRACE is developing a set of bespoke tools and techniques for computational analysis. By mapping a close reading corpus (CRC) of editions, translations, and early theoretical adaptations of the four corpora mentioned against a distant reading corpus (DRC) of nearly half a million books, our project traces direct and indirect quotations, paraphrases and individual translations, as well as specific vocabularies. Our DRC covers nearly two centuries, from 1540, when Agostino Steuco published his *De Perenni Philosophia* (On the Eternal Philosophy), to 1728, when Isaac Newton's posthumously published studies of ancient history appeared, which showcase his dedication to the ancient wisdom tradition. They also cover six major languages—Latin, Italian, French, German, Dutch and English—drawing on major digitised collections that, through their extent and reach, are representative of everything that was published during the period, thus allowing VERITRACE to draw statistically meaningful inferences.

The linguistic background of VERITRACE's digital methods

Defining queries, evaluating the results, and drawing conclusions from them make necessary a theoretical reflection about the linguistic features of the multilevel corpora. This study comprises three research *foci*: the notion of a diachronic corpus, intertextuality and distinctive vocabulary of *prisca sapientia*.

The project's timeframe (1540–1728) delimits a corpus similar to those used in historical corpus linguistics, which has recourse to computational methods to map change in language usage. However, our corpus is not defined by linguistic change but by the pivotal moments in reflection about ancient wisdom. Corpus linguistic methods can be inspiring for our work in the detection of patterns in the DRC, such as the gradual regression of topics related to *prisca sapientia* or the appearance of linguistic markers of doubt about the authenticity of these sources.

Intertextuality, a partly literary, partly linguistic notion, denotes the way in which texts derive meaning through their relationship with other texts, including allusions, references, quotations and shared cultural context. Intertextuality is, in fact, at the heart of our project because early modern authors built on ancient wisdom via a network of implicit and explicit references. Belief or unbelief in an original God-given wisdom is manifest via quotations from ancient wisdom texts, commentaries on them and even via commentaries on commentaries. A shared cultural background accounts for the frequency of unreferenced quotations (for instance, the syntagm 'the voice of fire' evokes the *Chaldean Oracles* in seventeenth-century philosophical debates without an explicit reference). Detecting intertexts from *prisca* sources in early modern printed books, namely word-for-word quotations, close paraphrases or marginal references, often in an abridged form, is crucial to understanding cross-national and cross-cultural influence.

One of the project's tasks is to study *prisca* vocabulary or terminology using qualitative and quantitative methods. This task comprises:

- a preliminary study of the four corpora to define typical lexical categories in them
- technical processing of some of the digital editions of these corpora
- testing different named entity recognition methods for the retrieval of meaningful lexical elements
- comparing possible methods, such as dictionary, ontology, thesaurus, etc., based on terminological features.

While natural language processing (NLP) offers several useful methods for our CRC, giving the planned *prisca* vocabulary a final form is not a self-evident choice. We have encountered several methodological difficulties when working with ancient wisdom vocabulary. Many terms found in these texts resist definition and sources, such as the *Chaldean Oracles*, because of their fragmentary character, can also resist meaningful quantification. To resolve these difficulties, we continue to test NLP methods on longer transcriptions, but we also concentrate on smaller lexical datasets, such as *hapax legomena* and formulaic expressions in the *Chaldean Oracles*, personified notions in the *Orphic Hymns*, etc. A long-term goal is to create a dataset of relevant *prisca* terms and map their use in early modern natural philosophical discourse.

Developing state-of-the-art tools

VERITRACE, now in its second year, has begun developing concrete cutting-edge tools to support scholarship. We have created an early version of the VERITRACE research tool that allows users to search our extensive collection of historical texts (upwards of 400 000 documents) in a web browser.

While keyword searching is valuable, it is just the beginning of what our platform offers. Researchers will also be able to use our **text matching tool**—think of it as a historical version of plagiarism detection software. This feature identifies and highlights the most similar passages between a source text and multiple target texts, showing how ideas were borrowed, adapted or responded to throughout the early modern period.

What makes this particularly powerful is that text matching works across languages. The system can identify similar semantic content regardless of whether it appears in English, Latin, French or any of our other project languages. This adds an exciting computational dimension to the work of intellectual historians.

Figure 2: An engraving of Hermes Trismegistus, from Pierre Mussard's 1680 Frankfurt am Main edition of Historia Deorum Fatidicorum.

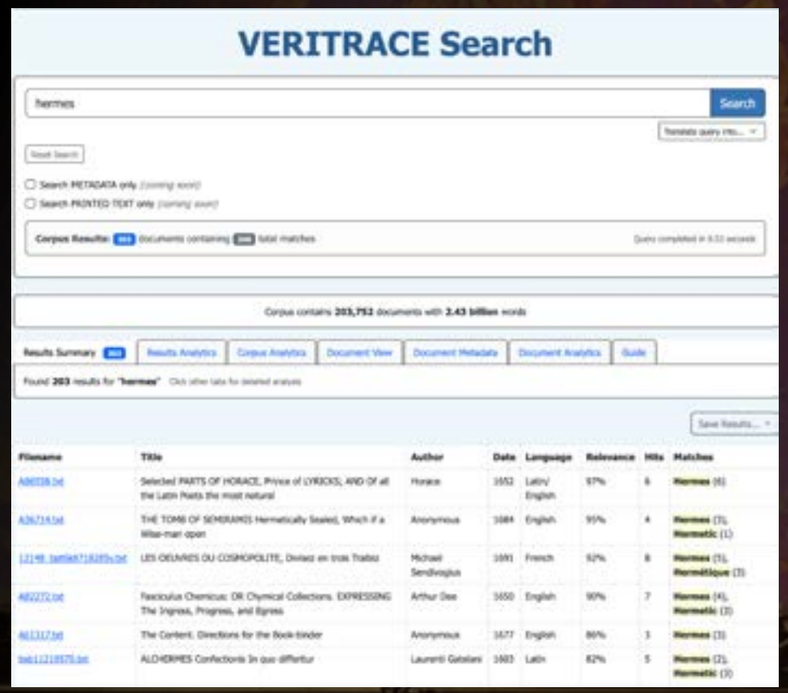

Figure 3: A preview of VERITRACE's research tool, still under development.

Both keyword searches and text matching results can be filtered by **sentiment**—the emotional tone of the passages. The system automatically analyses whether content expresses negative, neutral or positive sentiments, adding another layer of insight into how ideas were received and transmitted. Researchers can also browse the collection by topic, track how keywords evolved over time, or dive deep into individual texts. As users provide feedback, we will continue enhancing the platform with additional features. The goal is to provide scholars with an entire suite of tools they can use to explore, investigate and study the ancient wisdom tradition and early modern science more generally.

The path to our current progress was not always smooth. The project's first year was dedicated to obtaining the raw materials—both the bibliographic metadata (information about each text, like author, title, date, etc.) and the digital texts themselves. While some resources were readily available online, most of our data required more complex acquisition methods.

After gathering the raw data, we faced the substantial challenge of **cleaning** it—formatting and standardising information so computers could effectively process it. For example, just one of our data sources contained 527 different date formats—including Roman numerals, date ranges and notations with symbols like question marks. For consistent analysis across our collection, each date needed to be converted to a simple four-digit format (e.g. 1632).

To help tackle this challenge, we have begun testing large language models (similar to ChatGPT) to examine bibliographic records. These AI tools provide a preliminary analysis of record accuracy and suggest standardisation approaches that preserve the original meaning. This innovative application of AI promises to reduce the time and effort required for data preparation significantly, and we are currently refining how to best integrate these tools into our workflow. Our data cleaning efforts remain ongoing and will accelerate throughout the year.

All these efforts continue in VERITRACE's second year, and we are excited to share with the wider scholarly community how our tools will advance historical research and help answer the core research questions of our ERC-funded project.

Suggested literature

Allen, G. (2000) *Intertextuality*, London: Routledge.

Cohen, M. (2009) 'Narratology in the archive of literature', *Representations*, 108, pp. 51–75.

Colleoni, E., Rozza, A. and Arvidson, A. (2014) 'Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data', *Journal of Communication*, 64, pp. 317–332.

Dobson, J.E. (2019) *Critical Digital Humanities: The Search for a Methodology*. Urbana, IL: University of Illinois Press.

Dobson, J.E. (2021) 'Interpretable outputs: Criteria for machine learning in the humanities', *Digital Humanities Quarterly*, 15(2). Available at: http://www.digitalhumanities.org/dhq/vol/15/2/000555/000555.html (Accessed: 17 January 2024).

Foltz, P.W. (2011) 'Discourse coherence and LSA', in Landauer, T.K., McNamara, D.S., Dennis, S. and Kintsch, W. (eds.) *Handbook of Latent Semantic Analysis*. New York and London: Routledge, pp. 169–184.

Forstall, C.W., Scheirer, W.J. (eds.) (2019) *Quantitative Intertextuality*, Cham: Springer.

Garin, E. (1994) *Il ritorno dei filosofi antichi*. Reprint. Naples: Istituto Italiano per gli Studi Filosofici. Original work published 1984.

Hill, M.J. and Hengchen, S. (2019) 'Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study', *Digital Scholarship in the Humanities*, 34, pp. 825–843.

Imai, K. (2018) *Quantitative Social Science: An Introduction*. Princeton and Oxford: Princeton University Press.

Karsdorp, F., Kestemont, M. and Riddell, A. (2021) *Humanities Data Analysis: Case Studies with Python*. Princeton and Oxford: Princeton University Press.

Kintsch, W., McNamara, D.S., Dennis, S. and Landauer, T.K. (2011) 'LSA and meaning: in theory and application', in Landauer, T.K., McNamara, D.S., Dennis, S. and Kintsch, W. (eds.) *Handbook of Latent Semantic Analysis*. New York and London: Routledge, pp. 467–479.

Kurhekar, P., Nigam, S. and Pillai, S. (2021) 'Automated text and tabular data extraction from scanned document Images', in Sharma, N., Chakrabarti, A., Balas, V.E., and Bruckstein, A.M. (eds.) *Data Management, Analytics and Innovation: Proceedings of ICDMAI 2021, Volume 1*. Singapore: Springer Nature Singapore, pp. 169–182.

Lashari, I.A. and Wiil, U.K. (2016) 'Monitoring public opinion by measuring the sentiment of retweets on Twitter', in Bernadas, C. and Minchella D. (eds.) *Proceedings of the 3rd European Conference on Social Media*. Reading: Academic Conferences and Publishing, pp. 153–161.

Leinkauf, T. (2017) "Prisca scientia' versus 'prisca sapientia'. Zwei Modelle des Umgangs mit der Tradition am Beispiel des Rückgriffs auf die Vorsokratik im Kontext der frühneuzeitlichen Debatte und der Ausbildung des Kontinuitätsmodell der 'prisca sapientia' bzw. 'philosophia perennis", *Mediterranea. International Journal on the Transfer of Knowledge*, 2, pp. 121–143.

Rani, S. and Kumar, P. (2019) 'A sentiment analysis system for social media using machine learning techniques: Social enablement', *Digital Scholarship in the Humanities*, 34, pp. 569–581.

Ratna, A.A.P., Purnamasari, P.D., Adhi, B.A., Ekadiyanto, F.A., Salman, M., Mardiyah, M. and Winata, D.J. (2017) 'Cross-language plagiarism detection system using latent semantic analysis and learning vector quantization', *Algorithms*, 10, 69. Available at: https://www.mdpi.com/1999-4893/10/2/69/htm (Accessed: 17 January 2024).

Reid, D. (2019) 'Distant reading, "The great unread", and 19th-century British conceptualizations of the civilizing mission: A case study', *Journal of Interdisciplinary History of Ideas*, 15. Available at: http://journals.openedition.org/jihi/435 (Accessed: 17 January 2024).

Schmidt-Biggemann, W. (2004) *Philosophia Perennis: Historical Outlines of Western Spirituality in Ancient, Medieval and Early Modern Thought*. Dordrecht: Springer.

Schmitt, C.B. (1966) 'Perennial philosophy: from Agostino Steuco to Leibniz', *Journal of the History of Ideas*, 27, pp. 505–532.

Soni, S., Klein, L. and Eisenstein J. (2021) 'Abolitionist networks: Modeling language change in nineteenth-century activist newspapers', *Journal of Cultural Analytics*, 1, pp. 143.

Walker, D.P. (1972) *The Ancient Theology: Studies in Christian Platonism from the Fifteenth to the Eighteenth Century*. Ithaca, NY: Cornell University Press.

Zhou, Z.-H. (2021) *Machine Learning*. Singapore: Springer Nature Singapore.

VERITRACE
Traces de la Verité

PROJECT NAME

VERITRACE, or Traces de la Verité: The Reappropriation of Ancient Wisdom in Early Modern Natural Philosophy

PROJECT SUMMARY

VERITRACE enhances our understanding of the role of ancient wisdom writings in the development of early modern natural philosophy by employing state-of-the-art digital techniques on a large corpus of early modern texts. The project charts the spread of this discourse from Renaissance Italy to early modern Europe and identifies watershed moments in the reception of these ancient wisdom writings.

PROJECT LEAD PROFILE

Professor Cornelis J. Schilt is Associate Professor in History and Philosophy of Knowledge at the Vrije Universiteit Brussels. He was educated in physics and astrophysics, as well as history and philosophy of science, at Utrecht University, and he holds a DPhil in History of Science from Oxford University. Previously, he was a Junior Research Fellow at Linacre College, Oxford, and Senior editor with the Newton Project. He has published extensively on Renaissance and early modern science and religion, with a particular focus on the life and writings of Isaac Newton, including Isaac Newton and the Study of Chronology: Prophecy, History, and Method (Amsterdam University Press, 2021). He is very interested in the application of computational methods to study past, present and future.

PROJECT CONTACTS

Professor Cornelis J. (Kees-Jan) Schilt
Vrije Universiteit Brussel
Department of History, Archaeology, Arts, Philosophy and Ethics (HARP)
Pleinlaan 2 B-1050 Brussels
+32 2 629 39 05
cschilt@vub.be
https://veritrace.eu

FUNDING

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101076836.