Category Type: **Short Paper**

Searching 'Ancient' Wisdom Using New Tools: The Challenges of Connecting and Integrating Multilinguistic, Interdisciplinary Data in the VERITRACE project

Author & Presenter: Dr. Jeffrey Wolf (https://orcid.org/0009-0007-9879-4476), Postdoctoral Fellow (Digital Humanities), VERITRACE (https://veritrace.eu), Vrije Universiteit Brussel

—

I will present early work-in-progress for the ERC Starting Grant project 'VERITRACE' or "*Traces de la verité*: The reappropriation of ancient wisdom in early modern natural philosophy" (2023-2028). Based at Vrije Universiteit Brussel and led by Professor Cornelis J. Schilt, VERITRACE anticipates significant breakthroughs in our understanding of the role of ancient wisdom in the development of early modern natural philosophy by using state-of-the-art digital techniques to analyse a large corpus of early modern texts. Because the VERITRACE corpus spans different data sets and institutions, it requires us to break down silos, span disciplines, and advance the integration of previously created data.

But first, the intellectual background: during the Italian Renaissance, the idea that true knowledge was not new but ancient was quite influential, fuelled by the rediscovery of texts like the *Corpus Hermeticum*, the *Chaldean Oracles*, the *Orphic Hymns*, and the *Sibylline Oracles*, all understood to have been composed in antiquity. These texts, and the associated idea of knowledge as old, were reappropriated into a perennial philosophy, a *prisca sapientia*, which was a tradition that valued these works for their insights into God, mankind, and the cosmos (Schmitt, 1966; Walker, 1972; Garin, 1984; Schmidt-Biggemann, 2004; Leinkauf, 2017). Key figures in early modern science, such as Nicolaus Copernicus, Johannes Kepler, Francis Bacon, Isaac Newton, and Gottfried Wilhelm Leibniz, ascribed to this tradition in one form or another. Yet no comprehensive account exists of exactly *what* these luminaries took from these ancient wisdom writings and *how* the concept of a perennial truth influenced their knowledge-making. VERITRACE focusses on the widespread dissemination and impact of the ancient wisdom tradition, employing specialized digital techniques adapted for early modern studies.

The VERITRACE project integrates close and distant reading of Renaissance and early modern texts. We do this by creating, and analysing, both a Close Reading Corpus (CRC) – which includes all the relevant editions of the four main corpora under scrutiny that were published during the Renaissance and early modern period but also works that drew heavily on these ancient wisdom writings – and a Distant Reading Corpus (DRC). The Distant Reading method, closely related to natural language processing, allows 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 (Cohen, 2009; Reid, 2019). The DRC consists of several hundred thousand works from important European digital text databases 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, contains about 58,000 English and Latin

texts published between 1540 and 1700 (For exemplary uses of this resource, see, e.g. Pizelo et al., 2023; https://earlyprint.org)
- 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
- The *Digitale Sammlungen* of the Bavarian State Library, which contain nearly 365,000 titles published between 1540 and 1728, in Latin, German, and French, among others

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 (Hill and Hengchen, 2019; Kurhekar et al., 2021, Sangiacomo et al., 2022). Online repositories, like those of the Bibliothèque nationale de France, provide standardised data for content extraction, facilitating large-scale analysis (Imai, 2018; Karsdorp et al., 2021).

It is clear, then, that the primary data sources for the VERITRACE project are large, diverse data sources, containing multiple languages, created in specific contexts, and stored in separate institutions. The initial challenge for VERITRACE then is not only how to access this data but to integrate and clean it in a transparent, defensible way.

This Short Paper presentation will discuss how we have been able to access and combine our primary data sources in a manageable way, as well as point out some of the many obvious – and more subtle – challenges that working with such data present.

As an example of innovate data reuse, I will show how we are repurposing the data from the Universal Short Title Catalogue (https://www.ustc.ac.uk) to enrich and cross-reference the disparate metadata we have obtained from the primary collections that underpin our DRC corpus. We are therefore reusing data from another project at a separate institution (University of St. Andrews in Scotland) to enhance our data pipeline, which will make our subsequent analyses more robust. By using the USTC, we will be spanning disciplines, combining the history of science and knowledge-making with book history and bibliographical studies.

To enable the entire VERITRACE team to make sense of the data, I will show how we use OpenRefine to provide a more accessible way of connecting our raw data to the enriched metadata from the Universal Short Title Catalogue.

Finally, I will present preliminary results that showcase our use of the latest data visualization techniques, in the form of Streamlit dashboards, to enable team members to explore live data in a communal way, without knowing any code.

REFERENCES

Banerjee, K. (2019). *The Data Wrangler's Handbook: Simple Tools for Powerful Results*. American Library Association.

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

Diaz-Ordoñez, M., Rodríguez Baena, D. S., & Yun-Casalilla, B. (2023). 'A new approach for the construction of historical databases—NoSQL Document-oriented databases: the example of *AtlantoCracies*'. *Digital Scholarship in the Humanities*, *38*(3), 1014-1032. doi:10.1093/llc/fqad033

Garin, E. (1994) *Il ritorno dei filosofi antichi*. Reprint. Napels: 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.

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.

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.

Osborne, J. W. (2013). *Best Practices in Data Cleaning: A Complete Guide to Everything you Need to do Before and After Collecting Your Data*. Thousand Oaks, CA: Sage.

Piotrowski, M. (2012). *Natural Language Processing for Historical Texts*. Morgan & Claypool Publishers.

Pizelo, S., Koehl, A., Nagda, C., & Stahmer, C. (2023). 'Project Quintessence: Examining Textual Dimensionality with a Dynamic Corpus Explorer'. *Digital Humanities Quarterly*, *17*(3).

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).

Sangiacomo, A., Hogenbirk, H., Tanasescu, R., Karaisl, Antonia, & White, N. (2022). 'Reading in the Mist: High-Quality Optical Character Recognition Based on Freely Available Early Modern Digitized Books'. *Digital Scholarship in the Humanities*, *37*(4), 1197-1209. doi:10.1093/llc/fqac014

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.

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