Journal of Open Humanities Data

# "A Database of Intertexts in Valerius Flaccus' *Argonautica* 1: A Benchmarking Resource for the Evaluation of Computational Intertextual Search of Latin Corpora"

COLLECTION:
REPRESENTING THE ANCIENT WORLD THROUGH DATA

DATA PAPER

JOSEPH P. DEXTER
PRAMIT CHAUDHURI
PATRICK J. BURNS
ELIZABETH D. ADAMS
THOMAS J. BOLT
ADRIANA CÁSAREZ
JEFFREY H. FLYNT
KYLE LI
JAMES F. PATTERSON
ARIANE SCHWARTZ
SCOTT SHUMWAY

*Author affiliations can be found in the back matter of this article

ubiquity press

## ABSTRACT

Characterization of intertextual references among authors is fundamental for the study of Latin literature. In this paper, we describe a large-scale intertextuality dataset compiled from three modern commentaries on Valerius Flaccus' epic poem *Argonautica*. The dataset includes 945 references to earlier and contemporary Roman authors, as well as associated metadata required for use of multiple intertext search tools. To illustrate the dataset's reuse potential, we perform a new benchmark analysis of Fīlum, a sequence alignment tool for intertextuality detection.

CORRESPONDING AUTHORS:
**Joseph P. Dexter**
Harvard Data Science Initiative and Human Evolutionary Biology, Harvard University, Cambridge (MA), USA
jdexter@fas.harvard.edu
**Pramit Chaudhuri**
Classics, University of Texas at Austin, Austin (TX), USA
pramit.chaudhuri@austin.utexas.edu

KEYWORDS:
Latin; literature; computer science; linguistics

TO CITE THIS ARTICLE:
Dexter, J. P., Chaudhuri, P., Burns, P. J., Adams, E. D., Bolt, T. J., Cásarez, A., Flynt, J. H., Li, K., Patterson, J. F., Schwartz, A., & Shumway, S. (2024). "A Database of Intertexts in Valerius Flaccus' *Argonautica* 1: A Benchmarking Resource for the Evaluation of Computational Intertextual Search of Latin Corpora". *Journal of Open Humanities Data*, 10: 14, pp. 1–7. DOI: https://doi.org/10.5334/johd.153

# (1) OVERVIEW

## REPOSITORY LOCATION

JOHD Dataverse: https://doi.org/10.7910/DVN/S6RD4M

## CONTEXT

The epic poems of Vergil, Ovid, and other major Roman authors contain an extraordinary density of references to earlier Greek and Latin literature. These references encompass a broad array of intertextual relationships, ranging from overt quotation of memorable passages to subtle allusions requiring deep readerly expertise to appreciate. Identifying and interpreting these intertextual parallels constitutes a major activity of Latin literary criticism and is vital to understanding the poems' compositional artistry and cultural significance (Thomas, 1986; Hinds, 1998). As many parallels involve repetition or adaptation of short phrases, the study of Latin intertextuality is well-suited to computational approaches. Several foundational tools for corpus and intertextual search, including Diogenes, Tesserae, and TRACER, are now standard resources in the field (Heslin, 2019; Coffee *et al.*, 2012; Moritz *et al.*, 2016), and computational intertextual criticism of Latin literature continues to be an active topic of research (Bernstein *et al.*, 2015; Burns 2017; Dexter *et al.*, 2017; Forstall & Scheirer, 2019; Manjavacas *et al.*, 2019). As part of our ongoing work on the quantitative study of Latin intertextuality, we have created a benchmark dataset of intertextual parallels in Latin epic, which can be used for thorough and consistent evaluation of different search methods. The dataset was originally released with the following paper about language models and Latin intertextuality:

Burns, P. J., Brofos, J. A., Li, K., Chaudhuri, P., & Dexter, J. P. (2021). Profiling of Intertextuality in Latin Literature Using Word Embeddings. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4900–4907). DOI: https://doi.org/10.18653/v1/2021.naacl-main.389.

# (2) METHODS

## STEPS

The dataset consists of a catalog of textual similarities (intertexts) between the first book of Valerius Flaccus' *Argonautica* (VF 1) and other works in the Latin epic tradition. The catalog collates all references to four major Latin epics (Vergil's *Aeneid*, Ovid's *Metamorphoses*, Lucan's *Bellum Civile*, and Statius' *Thebaid*) recorded in three modern commentaries on VF 1 (Spaltenstein, 2002; Kleywegt, 2005; Zissos, 2008; Figure 1).

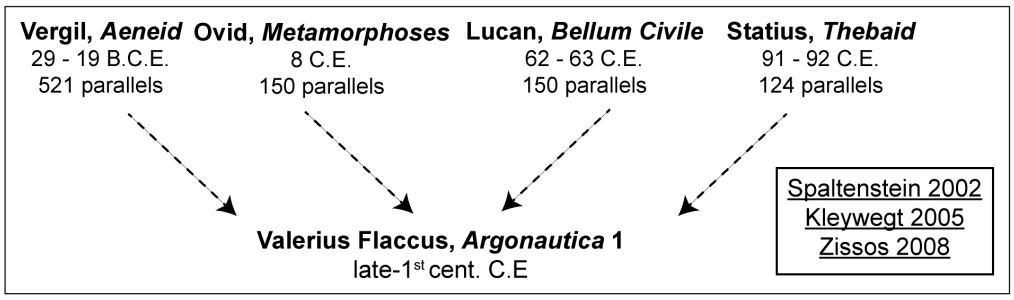


**Figure 1** Overview of intertextuality dataset compiled from three commentaries on VF 1.

Since the leading Latin intertextual search tool, developed by the Tesserae Project, analyzes two-word phrases, our dataset follows the same practice for ease of comparison (Coffee *et al.*, 2012). The Tesserae team also corroborated a common intuition among philologists that Latin intertexts frequently occur in the form of two-word phrases. On occasion, however, commentators refer to a single-word intertext or an intertext composed of several words. In these cases, our research team used expert judgment to choose a relevant two-word phrase. These phrases are composed of the key intertextual term plus a natural complement in close proximity (e.g., an adjective-noun, verb-object, or preposition-noun unit).

Our dataset uses the term "query phrase" to indicate a two-word phrase of interest in VF 1 derived from the commentaries and "result phrase" to indicate the two-word comparison phrase likewise derived from the commentaries. The terminology of "query" and "result" is intended to emphasize that the phrases are being deployed in a search and retrieval process that looks outward from VF 1 to a set of comparison texts. This unidirectional process is a technical simplification of human reading, of course, which takes account of multiple texts simultaneously in determining which phrases in any text might be of interest. In other words, a commentator's choice of lemma can never arise from consideration of a single text in isolation but always emerges from a personal history of reading situated within a broader literary critical and cultural context.

In addition to the list of intertexts, the commentary sources, and the citation information, the dataset also includes three parameters describing the relationship between the phrase in VF 1 and the parallel phrases noted by commentators, which are labeled "Order Free," "Interval," and "Edit Distance". "Order Free" is a binary parameter indicating whether the words comprising the result phrase (i.e., the intertext or parallel phrase) are either adjacent and in the same order as the source phrase (in which case the cell is marked "False") or non-adjacent or in reverse order ("True"). "Interval" indicates the number of words between the words comprising the result phrase for parallels that are labeled "Order Free;" when the result phrase consists of adjacent words, the interval is zero. For fixed-order searches, the "Interval" column is left blank. "Edit Distance" indicates the number of character substitutions, additions, or deletions required to turn the query phrase into the result phrase. The first two parameters are necessary input data for the method of semantic intertextual search described in our original paper (Burns *et al*., 2021). The same parameters plus the edit distance parameter are required for another tool, Fīlum, which we developed to identify phonetically similar phrases (Chaudhuri *et al*., 2015; Chaudhuri & Dexter, 2017; see Reuse Potential for further discussion).

### SAMPLING STRATEGY

Select intertextual parallels previously recorded in the commentaries of Spaltenstein (2002), Kleywegt (2005), and Zissos (2008) were included in the dataset. No new parallels were added.

### QUALITY CONTROL

All entries in the dataset were reviewed by multiple members of the project team for completeness and accuracy.

## (3) DATASET DESCRIPTION

### OBJECT NAME

vf_intertext_dataset_2_0.csv

### FORMAT NAMES AND VERSIONS

CSV

### CREATION DATES

Start date: 2014-07-01

End date: 2023-08-31

### DATASET CREATORS

The dataset was created by Joseph P. Dexter, Pramit Chaudhuri, Patrick J. Burns, Elizabeth D. Adams, Thomas J. Bolt, Adriana Cásarez, Jeffrey H. Flynt, Kyle Li, James F. Patterson, Ariane Schwartz, and Scott Shumway.

### LANGUAGE

The language of entries in the dataset is Latin. The language of the metadata is English.

### LICENSE

CC0 1.0

### REPOSITORY NAME

JODH Dataverse

### PUBLICATION DATE

The dataset was published on 2023–09–04.

## (4) REUSE POTENTIAL

The dataset presented here provides a resource for researchers investigating intertextuality in historical language traditions, especially Latin. For philologists pursuing qualitative studies, the catalog assembles in a single source a systematic list of intertextual parallels between VF 1 and several major Latin epic poets who either influenced or were influenced by Valerius Flaccus. The availability of three modern, high-quality philological commentaries attests to the fact that this particular text is an especially rich model of intertextual relationships. The collation from multiple sources, furthermore, enables researchers to analyze the practices of modern philologists, in particular the patterns of reference and degree of overlap among the commentators. Such reuse could involve qualitative review of recorded intertexts, as well as quantitative and statistical analysis of the entire dataset. For instance, although the total number of parallels recorded by each commentator differs substantially (Kleywegt 757, Zissos 373, Spaltenstein 228), it is striking that the distribution over comparison texts is fairly consistent, with the *Aeneid* referred to most often (55–64%) and Ovid (7–17%), Lucan (17–18%), and Statius (11–13%) each referred to with approximately similar frequency.

Perhaps the greatest future potential, however, lies in the use of the dataset as a benchmark for testing intertextual search methods. This use case was already exemplified in the original paper (Burns *et al.*, 2021), which compared lemma matching and semantic scoring using word embeddings as two complementary approaches to Latin intertextual search. The addition of the edit distance parameter in the revised dataset now facilitates evaluation of search methods based on character similarity, such as Fīlum.

To illustrate this kind of reuse potential, we used the dataset to conduct a large-scale validation analysis of Fīlum. In prior work, we applied Fīlum to a variety of case studies involving intertextuality in classical and post-classical Latin literature (Chaudhuri *et al.*, 2015; Chaudhuri and Dexter, 2017), but we have not previously validated the performance of the tool at scale. We ran book-level Fīlum searches for all 945 intertexts in the dataset using the three parameters recorded therein ("Order Free," "Interval," and "Edit Distance"). To summarize these results, we calculated the precision@*k* and recall@*k* across a range of cutoffs (*k* = 1, 3, 5, 10, 25, 50, 75, 100, and 250), which are the metrics we used to validate the semantic search method (see Burns *et al.*, 2021 for details). As shown in Figure 2, Fīlum outperforms semantic search in both precision and recall; more than half of all intertexts can be recovered with no off-target results, and more than 90% can be recovered with at most 250 off-target results. One example of the type of intertext pair well-suited to discovery using phonetic search (but not semantic search) is *patuere doli* ("his deceptions were revealed," VF 1.64) and [*nec*] *latuere doli* ("deceptions did [not] escape the notice," *Aeneid* 1.130). Although the phrases are phonetically very similar – only a single character (p/l) distinguishes them – they are semantically distinct; the two verbs are almost antonymic. As a consequence, the similarity score for the pair calculated from word embeddings is not especially high, which contrasts with the very low edit distance and, therefore, the absence of off-target results in a Fīlum search.

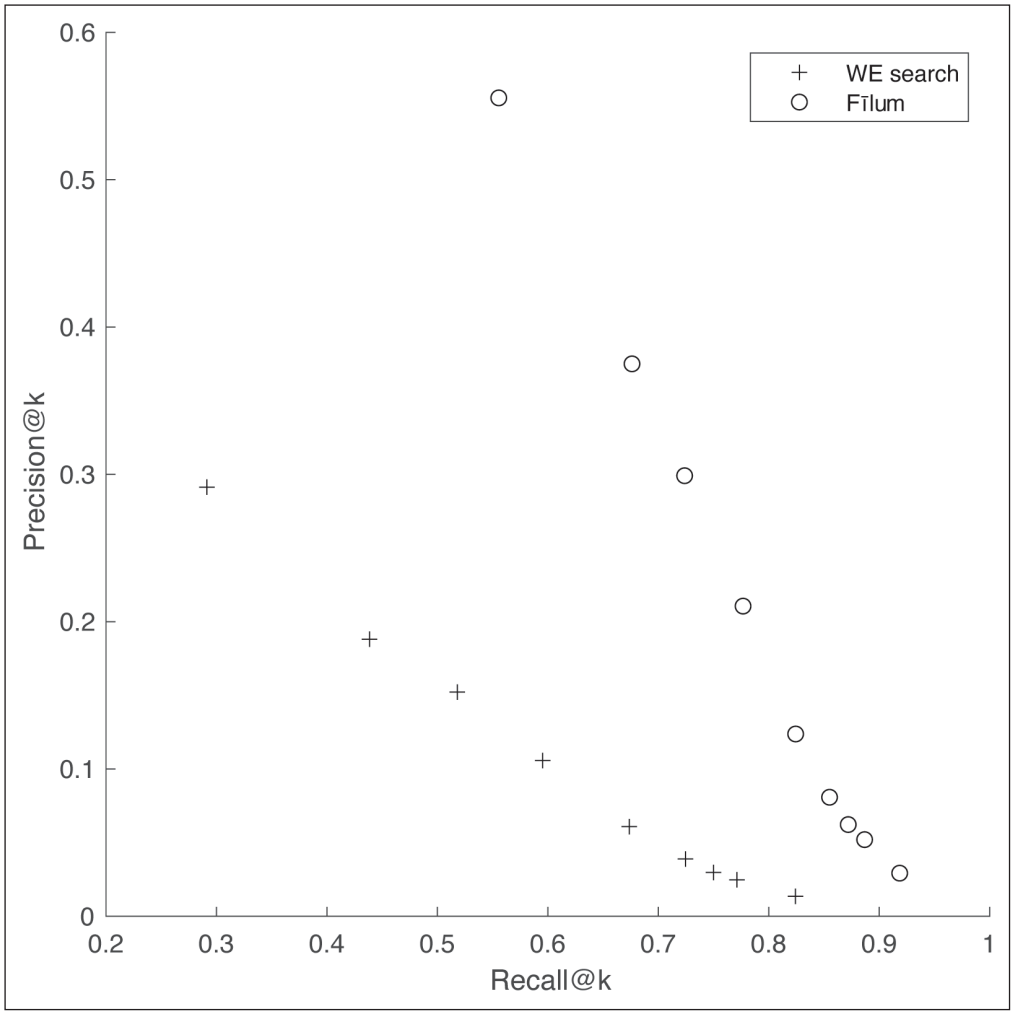


**Figure 2** Precision@*k* and recall@*k* for Fīlum and semantic search on the full VF 1 dataset. The semantic search data are reprinted from Burns *et al.*, 2021.

## ACKNOWLEDGEMENTS

We thank James A. Brofos, Jorge A. Bonilla Lopez, Tathagata Dasgupta, and Nilesh Tripuraneni for their contributions to our ongoing research on Latin intertextuality, and Neil Coffee and the Tesserae Project team for helpful discussions about intertextuality benchmarking.

## FUNDING INFORMATION

This research was conducted under the auspices of the Quantitative Criticism Lab (www.qcrit.org) and was supported by a National Endowment for the Humanities Digital Humanities Start-Up Grant (grant no. HD-248410-16), a National Endowment for the Humanities Digital Humanities Advancement Grant (grant no. HAA-271822-20), an American Council of Learned Societies Digital Extension Grant, and a Neukom Institute for Computational Science CompX Faculty Grant. JPD was supported by a Neukom Fellowship and a Harvard Data Science Fellowship, and PC was supported by a New Directions Fellowship from the Andrew W. Mellon Foundation.

## COMPETING INTERESTS

PJB is guest editor of the special collection *Representing the Ancient World through Data* and a member of the editor board for JOHD; he did not take part in the editorial process pertaining to this manuscript. All other authors have no competing interests.

## AUTHOR CONTRIBUTIONS

Joseph P. Dexter: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft, Writing – Review & Editing.

Pramit Chaudhuri: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft, Writing – Review & Editing.

Patrick J. Burns: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Review & Editing.

Elizabeth D. Adams: Data Curation, Formal Analysis, Investigation, Validation, Writing – Review & Editing.

Thomas J. Bolt: Data Curation, Formal Analysis, Investigation, Validation, Writing – Review & Editing.

Adriana Cásarez: Data Curation, Formal Analysis, Investigation, Validation, Writing – Review & Editing.

Jeffrey H. Flynt: Conceptualization, Data Curation, Investigation, Methodology, Software, Validation, Visualization.

Kyle Li: Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation.

James F. Patterson: Data Curation, Formal Analysis, Investigation, Validation, Writing – Review & Editing.

Ariane Schwartz: Data Curation, Formal Analysis, Investigation, Validation, Writing – Review & Editing.

Scott Shumway: Conceptualization, Data Curation, Investigation, Methodology, Software, Validation, Visualization.

## AUTHOR AFFILIATIONS

**Joseph P. Dexter** orcid.org/0000-0001-8524-5792
Harvard Data Science Initiative and Human Evolutionary Biology, Harvard University, Cambridge (MA), USA

**Pramit Chaudhuri** orcid.org/0000-0003-2643-0829
Classics, University of Texas at Austin, Austin (TX), USA

**Patrick J. Burns** orcid.org/0000-0003-2158-866X
Institute for the Study of the Ancient World, New York University, New York (NY), USA

**Elizabeth D. Adams** orcid.org/0009-0008-0751-1352
Classics, University of Texas at Austin, Austin (TX), USA

**Thomas J. Bolt** orcid.org/0000-0002-3088-5562
Languages and Literary Studies, Lafayette College, Easton (PA), USA

**Adriana Cásarez**
University of Texas Libraries, University of Texas at Austin, Austin (TX), USA

**Jeffrey H. Flynt**
Independent Scholar, USA

**Kyle Li** orcid.org/0009-0009-1547-0826
Computer Science, Columbia University, New York (NY), USA

**James F. Patterson** orcid.org/0009-0006-3501-6266
Classics, Yale University, New Haven (CT), USA

**Ariane Schwartz** orcid.org/0009-0009-2249-9491
Independent Scholar, USA

**Scott Shumway** orcid.org/0009-0000-1777-2865
Independent Scholar, USA

## REFERENCES

**Bernstein, N., Gervais, K.,** & **Lin, W.** (2015). Comparative rates of text reuse in classical Latin hexameter poetry. *Digital Humanities Quarterly*, 9.3. https://www.digitalhumanities.org/dhq/vol/9/3/000237/000237.html

**Burns, P. J.** (2017). Measuring and Mapping Intergeneric Allusion in Latin Poetry using Tesserae. *Journal of Data Mining and Digital Humanities*, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages. DOI: https://doi.org/10.46298/jdmdh.3821

**Burns, P. J., Brofos, J. A., Li, K., Chaudhuri, P.,** & **Dexter, J. P.** (2021). Profiling of Intertextuality in Latin Literature Using Word Embeddings. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp.4900–4907). DOI: https://doi.org/10.18653/v1/2021.naacl-main.389

**Chaudhuri, P.,** & **Dexter, J. P.** (2017). Bioinformatics and Classical Literary Study. *Journal of Data Mining and Digital Humanities*, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages. DOI: https://doi.org/10.46298/jdmdh.1386

**Chaudhuri, P., Dexter, J. P.,** & **Bonilla Lopez, J. A.** (2015). Strings, triangles, and go-betweens: Intertextual approaches to Silius' Carthaginian debates. *Dictynna*, *12*. DOI: https://doi.org/10.4000/dictynna.1156

**Coffee, N., Koenig, J.-P., Poornima, S., Ossewaarde, R., Forstall, C.,** & **Jacobson, S.** (2012). Intertextuality in the digital age. *Transactions of the American Philological Association*, *142*, 383–422. DOI: https://doi.org/10.1353/apa.2012.0010

**Dexter, J. P., Katz, T., Tripuraneni, N., Dasgupta, T., Kannan, A., Brofos, J. A., Bonilla Lopez, J. A., Schroeder, L. A., Cásarez, A., Rabinovich, M., Haimson Lushkov, A.,** & **Chaudhuri, P.** (2017). Quantitative criticism of literary relationships. *Proceedings of the National Academy of Sciences USA*, *114*, E3195–E3204. DOI: https://doi.org/10.1073/pnas.1611910114

**Forstall, C. W.,** & **Scheirer, W. J.** (2019). *Quantitative Intertextuality: Analyzing the Markers of Information Reuse*. Cham: Springer International Publishing. DOI: https://doi.org/10.1007/978-3-030-23415-7

**Heslin, P.** (2019). *Diogenes*. Retrieved from https://d.iogen.es/d/index.html (last accessed 4 December 2023).

**Hinds, S. E.** (1998). *Allusion and Intertext: Dynamics of Appropriation in Roman Poetry*. Cambridge: Cambridge University Press.

**Kleywegt, A. J.** (2005). *Valerius Flaccus, Argonautica, Book I*. Mnemosyne, Bibliotheca Classica Batava. Supplementum. Leiden: Brill.

**Manjavacas, E., Long, B.,** & **Kestemont, M.** (2019). On the Feasibility of Automated Detection of Allusive Text Reuse. *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 104–114). DOI: https://doi.org/10.18653/v1/W19-2514

**Moritz, M., Wiederhold, A., Pavlek, B., Bizzoni, Y.,** & **Büchler, M.** (2016). Non-Literal Text Reuse in Historical Texts: An Approach to Identify Reuse Transformations and its Application to Bible Reuse. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1849–1859). DOI: https://doi.org/10.18653/v1/D16-1190

**Spaltenstein, F.** (2002). *Commentaire des "Argonautica" de Valérius Flaccus (livres 1 et 2)*. Bruxelles: Éd. Latomus.

**Thomas, R. F.** (1986). Virgil's Georgics and the Art of Reference. *Harvard Studies in Classical Philology*, *90*, 171–198. DOI: https://doi.org/10.2307/311468

**Zissos, A.** (2008). *Valerius Flaccus' Argonautica Book I, Edited with Introduction, Translation, and Commentary*. Oxford: Oxford University Press.

**TO CITE THIS ARTICLE:**
Dexter, J. P., Chaudhuri, P., Burns, P. J., Adams, E. D., Bolt, T. J., Cásarez, A., Flynt, J. H., Li, K., Patterson, J. F., Schwartz, A., & Shumway, S. (2024). "A Database of Intertexts in Valerius Flaccus' *Argonautica* 1: A Benchmarking Resource for the Evaluation of Computational Intertextual Search of Latin Corpora". *Journal of Open Humanities Data*, 10: 14, pp. 1–7. DOI: https://doi.org/10.5334/johd.153

**Submitted:** 02 September 2023
**Accepted:** 29 November 2023
**Published:** 29 January 2024

**COPYRIGHT:**
© 2024 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

*Journal of Open Humanities Data* is a peer-reviewed open access journal published by Ubiquity Press.